Energy Efficient Network Function Virtualisation in 5G Networks

Ahmed Noori Naima Al-Quzweeni

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Electronic and Electrical Engineering

September 2018

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 4 is based on the work from:

A. Al-Quzweeni, T. E. H. Elgorashi, L. Nonde, and J. M. H. Elmirghani, "Energy efficient network function virtualization in 5G networks," presented at the 17th International Conference on Transparent Optical Networks (ICTON), 2015.

This paper has been published jointly with my PhD supervisor prof. Jaafar Elmirghani, Dr Leonard Nonde, and my co-supervisor Dr. Taisir Elgorashi.

And:

A. Al-Quzweeni, A. Lawey, T. Elgorashi, and J. M. H. Elmirghani, "A framework for energy efficient NFV in 5G networks," in Transparent Optical Networks (ICTON), 2016 18th International Conference on, 2016, pp. 1-4.

This paper has been published jointly with my PhD supervisor prof. Jaafar Elmirghani, Dr Ahmed Lawey, and my co-supervisor Dr. Taisir Elgorashi.

Chapter 5 is based on the work from:

A. Al-Quzweeni, A. Lawey, T. Elgorashi, and J. M. H. Elmirghani, “Energy-efficient NFV in 5G network: the impact of backhaul traffic and VMs inter-traffic” to be submitted to IEEE Journal of Lightwave Technology.

This paper has been published jointly with my PhD supervisor prof. Jaafar Elmirghani and my co-supervisor Dr. Taisir Elgorashi.

Chapter 6 is based on the work from:

A. Al-Quzweeni, T. Elgorashi, and J. M. H. Elmirghani, "Energy-efficient content caching in 5G networks" to be submitted to IEEE Journal of Lightwave Technology.

This paper has been published jointly with my PhD supervisor prof. Jaafar Elmirghani and my co-supervisor Dr. Taisir Elgorashi.

Chapter 7 is based on the work from:

A. Al-Quzweeni, T. Elgorashi, and J. M. H. Elmirghani, "Energy-efficient integrated framework for NFV and content caching in 5G networks", to be submitted to IEEE Transactions on Network and Service Management.

This paper has been published jointly with my PhD supervisor prof. Jaafar Elmirghani and my co-supervisor Dr. Taisir Elgorashi.

Dedication

This thesis is dedicated with unrelieved grief to the memory of my beloved ***Mother***, ***Father***, *and* ***Brother***.

It is really hurts too much to think how I will never see you again.

You always on my mind, forever in my heart.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor; Professor Jaafar Elmirghani for his unceasing support, shrewd leadership, colossal knowledge, and patience throughout my entire PhD journey. He has always been a friend and brother who backed me up even in my personal life.

I would like to acknowledge my co-supervisor; Dr. Taisir Elgorashi for her support and useful discussion during thesis writing.

I wish to thank my friend Dr. Ahmed Lawey for all the constructive discussions and support.

I would like to acknowledge the Higher Committee for Education Development in Iraq (HCED) for funding my PhD.

I would like to express my deep gratitude to my friends Mohammed Safa, Haider Kais, and the beloved Muhannad Yousif for their support, encouragement, and for bringing joy to my life.

I wish to thank my beloved friend Dr. Katherine Tylor for her unceasing support to my family and I.

I wish to thank my friends and colleagues, past and present, in the Institute of Communications and Power Networks (ICapNet) in the School of Electronic and Electrical Engineering and particularly Dr Mohamed Musa for his support. I really enjoyed working with them.

A big thank you flows from the bottom of my heart to my family. My sister Thanaa, for being like a mother to me. My uncle Falah, who supported me and my late parents during many hard times. My cousin Muhammad for his unlimited support. My parents-in-law and brothers-in-law who always backed and supported me. My daughter Retaj, my little angel and the cool of my eyes, for her patience during my PhD. My beloved wife Zaineb, the faithful companion of the journey of my life, for her love, understanding, encouragement, belief in me and support.

Abstract

Once the dust settled around 4G, 5G mobile networks become the buzz word in the world of communication systems. The recent surge of bandwidth-greedy applications and the proliferation of smart phones and other wireless connected devices has led to an enormous increase in mobile traffic. Therefore, 5G networks have to deal with a huge number of connected devices of different types and applications, including devices running life-critical applications, and facilitate access to mobile resources easily. Therefore given the increase in traffic and number of connected devices, intelligent and energy efficient architectures are needed to adequately and sustainably meet these requirements. In this thesis network function virtualisation is investigated as a promising paradigm that can contribute to energy consumption reduction in 5G networks.

The work carried out in this thesis considers the energy efficiency mainly in terms of processing power consumption and network power consumption. Furthermore, it considers the energy consumption reduction that can be achieved by optimising the locations of virtual machines running the mobile 5G network functions. It also evaluates the consolidation and pooling of the mobile resources. A framework was introduced to virtualise the mobile core network functions and baseband processing functions. Mixed integer linear programming optimisation models and heuristics were developed minimise the total power consumption. The impact of virtualisation in the 5G front haul and back haul passive optical network was investigated by developing MILP models to optimise the location of virtual

machines. A further consideration is caching the contents close to the user and its impact on the total power consumption. The impact of a number of factor on the power consumption were investigated such as the total number of active users, the backhaul to the fronthaul traffic ratio, reduction/expansion in the traffic due to baseband processing, and the communication between virtual machines. Finally, the integration of network function virtualisation and content caching were introduced and their impact on improving the energy efficiency was investigated.

Table of Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

AIMMS	Advanced Interactive Multidimensional Modelling System
ALU	Arithmetic Logic Unit
AN	Access Network
AON	All Optical Network
APON	Asynchronous Transfer Mode Passive Optical Network
ATM	Asynchronous Transfer Mode
AxC	Antenna Carrier
B&B	Branch and Bound
BBU	Baseband Unit
BPON	Broadband Passive Optical Network
CBC	Coin Branch and Cut
CCITT	Consultative Committee for International Telegraphy and Telephony
CDM	Code Division Multiplexing
CN	Core Network
CO	Central Office
CP	Cyclic Prefix

CPRI	Common Public Radio Interface
CWDM	Coarse Wavelength Division Multiplexing
D2D	Device to Device
DL	Downlink
DWDM	Dense Wavelength Division Multiplexing
EDFA	Erbium Doped Fibre Amplifier
EEFCZ	Energy Efficient Fixed Cache Size
EENFVnoITr	Energy Efficient Network Function Virtualisation without Inter-Traffic
EENFVwithITr	Energy Efficient Network Function Virtualisation with Inter-Traffic
EEVarCZ	Energy Efficient Variable Cache Size
EIST	European Telecommunications Standards Institute
EO	Electrical to Optical
EPC	Evolved Packet Core
EPON	Ethernet Passive Optical Network
FBMC	Filter Band Multi-Carrier
FSAN	Full Service Access Nework
FTTB	Fibre To The Building
FTTC	Fibre To The Cabnet

FTTH	Fibre To The Home
GEM	Gigabit Passive Optical Network Encapsulation Method
GPON	Gigabit Passive Optical Network
HDTV	High Definition TV
IoT	Internet of Things
IOV	Input Output Virtualisation
IP	Internet Protocol
IPTV	Internet Protocol Television
ISG	Industry Specification Group
ITU-T	International Telecommunication Union - Telecommunication
LTE	Long Term Evolution
MILP	Mixed Integer Linear Programming
MIMO	Multi-Input Multi-Output
MME	Mobility and Management Entity
mMIMO	Massive Multi-Input Multi-Output
mmWave	Millimetre Wave
MNO	Mobile Network Operators
NFV	Network Function Virtualisation
OBS	Optical Burst Switching

OC	Optical Carrier
OCS	Optical Circuit Switching
ODN	Optical Distribution Network
OE	Optical to Electrical
OFBMC	Orthogonal Frequency Band Multi-Carrier
OFDM	Orthogonal Frequency Division Multiplexing
OLT	Optical Line Terminal
ONF	Open Network Foundation
ONU	Optical Network Unit
OPL	Optimisation Programming Language
OPS	Optical Packet Switching
OS	Operating System
P2MP	Point to Multi-Point
P2P	Point to Point
PCRF	Policy and Charging Role Function
PGW	Packet Gateway
PON	Passive Optical Network
PRB	Physical Resources Block
QAM	Quadrature Amplitude Modulation

QoE	Quality of Experience
QPSK	Quadrature Phase-Shift Keying
RAN	Radio Access Network
RAU	Radio Aggregation Unit
RE	Resources Elements
RRH	Remote Radio Head
RRU	Remote Radio Unit
SC-FDMA	Single-Carrier Code Division Multiple Access
SDH	Synchronous Digital Hierarchy
SDN	Software Defined Network
SGW	Switching Gateway
SONET	Synchronous Optical Network
TDM	Time Division Multiplexing
UFMC	Universal Filter Multi-Carrier
UHDV	Ultra-High Definition Video
UL	Uplink
VM	Virtual Machine
VMM	Virtual Machine Monitoring
VNIC	Virtual Network Interface Card

VQA	Video Quality Assessment
WDM	Wavelength Division Multiplexing
XG-PON	10 Gigabit Passive Optical Network

Chapter 1

Introduction

According to Cisco Visual Networking Index report in June 2017, mobile data traffic will witness seven pleats between 2016 and 2021 and will grow at a Compound Annual Growth Rate (CAGR) of 46% reaching 48.3 exabytes per month by 2021 [1]. This phenomenon is driven by a number of factors such as the enormous amount of connected devices and the development of data-greedy applications [2]. With such a tremendous amount of data traffic, a revolutionary mobile network architecture is needed. This network will represent the next generation of mobile network (5G) and it will be a mix of a multiple access technologies supported by a significant amount of new spectrum to provide different services to a massive number of different kind of users (i.e. IoT) with a high data rate at any time with less than 1 ms latency [3]. 5G networks are expected to be operational by 2020 where a huge number of devices and application will be using it [4].

Users, application, and devices of different kinds and purposes need to send and access data in both distributed and centralised servers and databases using public and/or private networks and clouds. To facilitate these requirements, 5G mobile networks have to be characterised by some traits such as intelligence, flexible traffic management, adaptive bandwidth and at the forefront of these traits is the energy efficiency. Information and Communication Technology (ICT) including services and devices were responsible for about 8% of the total world energy consumption

[5] and contribute about 2% of the global carbon emissions [6]. It is estimated that, if the current trends continue, the ICT energy consumption will reach about 14% of total worldwide consumption by 2020 [5].

Nowadays fossils fuel is the main energy resource in the world where it represented about 90% of the total available energy resources [7]. However, such kind of resources are not sustainable resources and they are very harmful to the environment and are a great contributor to the CO_2 emissions [8]. Considering this fact together with the energy cost, it is necessary for researchers in both industrial and academic sectors to focus on an energy-efficient paradigm for 5G networks that contributes to the reduction of the ICT energy consumption.

In the recent years, there have been significant endeavours around the world in both academic and industrial sectors intended to improve the energy-efficiency in ICT networks. This trend gave birth to a number of collective efforts, consortiums, and projects such as the EARTH project [9] and GreenTouch [10]. There have also been various efforts from researchers on reducing the power consumption in 5G networks. For instance, the authors in [11] focused in their work on the power consumption of base stations. They have provided a time-triggered sleep model for the future base stations in order to reduce the power consumption. The authors in [12] have investigated the base stations computation power and compared it to the transmission power. They concluded that the base station computation power will play an important role in the 5G energy-efficiency. The authors of [13] have developed an analytical model to address the planning and the dimensioning of 5G Cloud RAN (C-RAN) and compared it to the traditional RAN. They have showed that C-RAN can improve the 5G energy-efficiency. The research carried out in [14]

focused on offloading the network traffic to the mobile edge to improve the energy-efficiency of 5G mobile networks. The authors have developed an offloading mechanism for mobile edge computing in 5G where both file transmission and task computing have been considered.

Virtualisation has been proposed as an enabler for optimum use of network resources, scalability, and agility. In [15] the authors have stated that NFV is the most important recent advance in mobile networks where among its key benefits is the agile provisioning of mobile functions on demand. The fact that it is now possible to separate the functions form their underlying hardware and transfer them into software-based mobile functions as well as provide them on demand presents opportunities for optimising the physical resources and improve the network energy efficiency.

In this thesis, network function virtualisation has been identified as a promising key technology paradigm that can contribute to the energy-efficiency improvement in 5G networks. Moreover, this thesis goes beyond the limits of investigating the deployment of only virtualisation in 5G, it also investigates the impact of integrating the virtualisation with content caching on the energy-efficiency of the future mobile networks. In addition, an optical-based novel architecture has been proposed and investigated in this thesis to carry out the infrastructural burden of the 5G network and support NFV and caching the contents. In literature, NFV was investigated either in mobile core networks [16-18] or in the radio access network [19-21] of the mobile network and mostly using pooling of resources such as the work done in [22, 23]. In contrast, virtualisation in this thesis is not limited to a certain part in the mobile network, but it has been applied in both mobile core network and radio

access network. Moreover, it is not confined to pooling the network resources, but it actually concerns mobile functions-hardware decoupling and converting them into software-based functions. In contrast to works mentioned above, this thesis investigates the energy-efficiency under the impact of integrating the virtualisation with another technology, namely caching the contents.

Mixed Integer Linear Programming models and real-time heuristics have been developed in this thesis with the goal of improving the energy-efficiency in 5G mobile networks and reducing the emission of CO_2 into the environment.

1.1 Research Objectives

The main hypothesis in this thesis is that resource sharing via virtualisation can reduce the energy consumption of 5G networks and by integrating virtualisation with caching the contents, another level of energy-efficiency could be achieved. Therefore, the aim of this thesis is to investigate the energy-efficiency of Network Function Virtualisation (NFV) and caching the contents separately and jointly in 5G networks. In order to meet the overall goal, the following objectives were set:

1. Propose an energy efficient framework for NFV in 5G networks with an optical-based architecture to reduce the overall power consumption in 5G networks. In this framework, the mobile network functions are decoupled from their underlying hardware and converted into software-based functions.
2. Investigate the energy-efficiency benefit of passive optical networks (PONs) as wired access networks in 5G network and explore their impact on the optimum virtual machine placement and the total power consumption.

3. Study the impact of baseband processing virtual machines optimised location on the backhaul traffic and the effect of the backhaul to fronthaul traffic ratio on the optimisation problem.
4. Investigate the impact of the traffic between virtual machines on their optimised locations and resources consolidation as well as the energy-efficiency and the total power consumption.
5. Study the relationship between the total number of active users in the network and the energy-efficiency under the proposed virtualisation framework and investigate the VM servers' utilisation as well as their optimised locations in response to the total number of users.
6. Investigate the capability of the proposed optical architecture for the deployment of other technologies such as content caching and study the impact of different caching techniques (fixed and variable cache size) on the energy-efficiency.
7. Propose an integrated architecture for virtualisation and caching the content based on optical architectures and investigate the outcome of integrating the virtualisation with caching the contents.

1.2 Original Contributions

The following are the main contribution of this thesis:

1. A novel network function virtualisation model for mobile functions was developed based on mixed integer linear programming in 5G networks with optical-based architecture. The proposed architecture consists mainly of two parts: IP over WDM network and passive optical network (PON) as a wired access network.

2. A new MILP model was developed to minimise the total power consumption by optimising the VMs locations and VM servers' utilisation. The MILP model results are investigated under the impact of core network VMs (CNVM) inter-traffic, the variation of total number of active users during different times of the day, and the ratio of backhaul to fronthaul traffic.
3. An Energy-Efficient NFV heuristic model (EENFVnoITr) was developed as a real-time implementation of the proposed MILP model when the CNVMs inter-traffic is not considered whilst an Energy-Efficient NFV with CNVMs inter-traffic (EENFVwithITr) heuristic was developed when the CNVMs inter-traffic is considered.
4. A new MILP model was developed to optimise the total power consumption of video on demand services by optimising the fixed and variable cache size locations at different nodes of the proposed optical-based architecture. In addition, the cache size is optimised in the case of variable cache size. The MILP model considers different number of active users during the day and it was validated by new heuristics: Energy Efficient Fixed Cache Size (EEFCZ) heuristic for the fixed cache size case and Energy-Efficient Variable Cache Size (EEVarCZ) heuristic for variable cache size case.
5. Proposed a novel integrated framework architecture for virtualisation and caching in 5G networks to improve the energy-efficiency. A MILP model was developed to minimise the total power consumption by optimising the location and utilisation of VMs as well as the size and location of the cache nodes. The developed MILP mode was validated by an Energy-Efficient Virtualisation and Caching (EEVIRandCa) heuristic.

1.3 Related Publications

This work resulted in the following conference and journal papers that have been published and submitted for publication:

1. A. Al-Quzweeni, T. E. H. El-Gorashi, L. Nonde, and J. M. H. Elmirghani, "Energy efficient network function virtualization in 5G networks," presented at the 17th International Conference on Transparent Optical Networks (ICTON), 2015.
2. A. Al-Quzweeni, A. Lawey, T. El-Gorashi, and J. M. H. Elmirghani, "A framework for energy efficient NFV in 5G networks," in Transparent Optical Networks (ICTON), 2016 18th International Conference on, 2016, pp. 1-4.
3. A. Al-Quzweeni, A. Lawey, T. El-Gorashi, and J. M. H. Elmirghani, "Energy-efficient NFV in 5G network: the impact of backhaul traffic and VMs inter-traffic" to be submitted to IEEE Journal of Lightwave Technology.
4. A. Al-Quzweeni, T. El-Gorashi, and J. M. H. Elmirghani, "Energy-efficient content caching in 5G networks" to be submitted to IEEE Journal of Lightwave Technology.
5. A. Al-Quzweeni, T. El-Gorashi, and J. M. H. Elmirghani, "Energy-efficient integrated framework for NFV and content caching in 5G networks", to be submitted to IEEE Transactions on Network and Service Management.

1.4 Thesis Organisation

Following the introduction in this chapter (Chapter 1), this thesis is organised as follows:

Chapter 2 reviews the next generation of mobile networks (5G), and explains the main requirements of 5G networks and the key technologies needed. Additionally, it reviews the history of optical networks and passive optical networks.

Chapter 3 is a detailed review of virtualisation and NFV. In addition, it discusses the principles of virtualisation and the types of virtualisation and paves the way to the principles and the definition of NFV. It also sheds light on various energy efficient approaches that have been introduced in 5G networks. Additionally, the basics of optimisation problems and their formulation is also explained in this chapter.

Chapter 4 introduces a novel framework and MILP model for energy efficient NFV in 5G networks. The model investigates the impact of deployment of PONs and the effect of different VM workloads.

Chapter 5 extends to the work done in Chapter 3 by investigating the impact of the backhaul to the fronthaul traffic ratio and the influence of the inter-traffic between VMs of the mobile core functions. In addition, it investigates the effect of variation of the total number of users during the day on the energy-efficiency and the optimisation problem. The MILP model is validated by two heuristic models: An Energy-Efficient NFV without Inter-traffic (EENFVnoITr) heuristic for no CNVMs inter-traffic and an Energy-Efficient NFV with CNVMs inter-traffic

(EENFVwithITr) heuristic for the case where CNVMs inter-traffic is considered. In addition, different values of CNVMs inter-traffic are considered.

Chapter 6 tackles the problem of high power consumption of video streaming services by proposing caching video contents in IP over WDM and PON (ONU and OLT) nodes. A full evaluation of the proposed MILP model is provided and discussed for both fixed and variable cache sizes. The MILP model results are validated by a new Energy Efficient Fixed Cache Size (EEFCZ) heuristic for the fixed cache size case and by a new Energy-Efficient Variable Cache Size (EEVarCZ) heuristic for the variable cache size case.

Chapter 7 introduces a novel integrated architecture of both virtualisation and content caching in 5G networks. In addition, it introduces a MILP model to investigate the impact of integrating virtualisation and content caching on the energy-efficiency of 5G networks by jointly optimising the cache size, VM server utilisation, and the location of both VMs and cache nodes in the proposed network. The MILP model results are validated by an Energy-Efficient Virtualisation and Caching (EEVIRandCa) heuristic.

Finally, thesis conclusions are drawn in Chapter 8 where the major contributions of this work are presented and future directions are discussed.

Chapter 2

5G the Next Generation of Mobile Network and Optical Networks

2.1 Introduction

5G will be a paradigm shift that includes enormous number of connected devices, massive bandwidth, dense base stations and huge number of antennas. To support this, a number of key technologies are proposed. Some of these key technologies are explained in this chapter. This chapter reviews the next generation of mobile networks (5G) and explains the main requirements of 5G networks and the key technologies needed. In addition, this chapter presents the evolution of optical networks including passive optical networks (PON) as the optical networks play a very crucial role in the development of the Internet.

2.2 Next Generation of Mobile Networks "5G"

According to Cisco forecast, there will be seven pleats of mobile data traffic between 2016 and 2021 [1]. This tremendous growth in mobile data traffic is driven by several determinants such as continuous growth in the number of wireless devices, development of data-hungry applications, social media, and video on demands.

The existing cellular systems are not sturdy enough to support the upsurge in mobile data traffic [24] and this is considered to be a crucial driver towards a new generation of mobile wireless networks, which is 5G. Therefore, 5G, "the buzz word on everyone's lips", will emerge to meet the requirements that are beyond the

capability of the existing cellular network systems. These requirements have occupied the researchers' interest in both the industrial and academic sectors, but the way in which the light is shed upon these requirements differs from one researcher to another. The authors of [25] and [26] presented some of the 5G requirements and the emerging technologies to ameliorate the 5G architecture. Although they focused in their papers on designing the architecture of 5G, they elaborated that the current technologies such as OFDMA could be used for at least half century and only the strategy of designing the 5G architecture needs to change drastically. In contrast, the authors of [27] explained that OFDM might be excluded from 5G for some of its features and alternatively other technologies might be deployed such as: Filter band Multi-Carrier (FBMC), Universal Filter Multi-Carrier (UFMC), and Orthogonal Frequency Band Multi-Carrier (OFBMC). Agiwal et al listed eight main requirements for 5G in [28]. Some of these requirements may however be considered as one concern, such as battery life and energy efficiency which were considered separately. They studied the 5G requirements with a Radio Access Network (RAN) focus. They have believed along with the authors of [29] and [30] that the requirements of 5G will be driven towards exploring the high frequencies in the band range 3 ~ 30 GHz. Accordingly, M. Alsharif and R. Nordin focused in [29] on three techniques which are mm-wave, massive MIMO, and small cells deployment. Although researchers differ in the way they present, categorise, and bracket the requirements of 5G networks, the following are considered as the main requirements of 5G networks:

- ***Latency***: 5G networks are characterised by 4 *any*(*s*); meaning that 5G will support a fully connected society where the information and data are shared

for *anything* and *anyone*, *anywhere* and *anytime* [31]. Therefore, there will be many life-critical devices and real-time application that require almost zero latency. Accordingly, 5G should consider extremely low latency on the line of 1 ms or less [31-34].

- ***Ubiquity***: 5G networks will provide a ubiquitous experience for end users [15, 33, 35-37] since everything will be connected to the Internet even cows may have sensors [38] and may be connected to the Internet [39].
- ***High data-rate and data volume***: 5G networks are expected to afford user high data-rate in the order of 100 times the current capacity [32, 40-42] which is around 10 Gbps [43-45] and data volume of 1000 times of today's data volume [46, 47] which is approximately 10 $Tb/s/km^2$ [32].
- ***Availability***: In 5G era there will be more than 50 billion connected devices [48] many of these devices will serve life-critical and real-time applications which require a very high network availability. Accordingly, 5G networks are expected to provide a high level of availability in the order of 99.99% [49-51].
- Other key features of 5G networks includes **100% *coverage*** [52-55], ***10 times lower power consumption*** than current systems [24, 56], up to **10 *years battery life*** of devices [57-62], and **10x – 100x *number of connected devices*** [63-66].

Figure 2.1summarise the main requirements and key features of 5G networks.

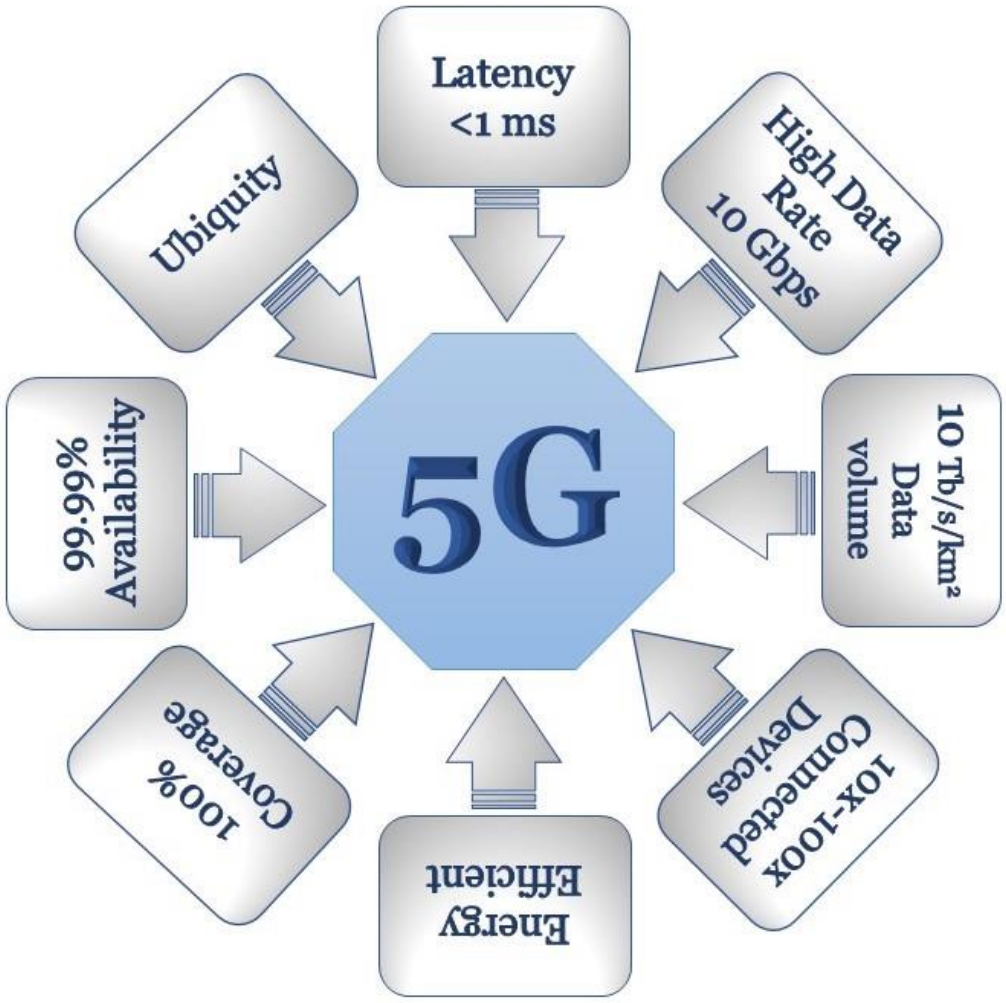


Figure 2.1 5G networks requirements

2.3 5G Key Technologies:

The revolutionary requirements of the envisaged 5G networks have inspired the researchers around the world to look for new methodologies, architectures, and technologies [24] as well as amalgamate different technologies to cope with the 5G requirements [67]. The following are the main key technologies that have occupied the researchers' attention:

- **mMIMO:** massive Multiple-Input Multiple-Output (mMIMO) (also known as Large-Scale Antenna System [68] , Large-Scale MIMO, hyper MIMO, full dimension MIMO, and Multi-User MIMO (MU-MIMO) [69]), refers to the simultaneous use of hundreds of antennas in a system [70]. In [71], the authors mentioned three main leverages of mMIMO on systems which are: 1) it reduces the air interface latency, 2) system robustness improvement, 3) reduces the complexity of media access layer. Adding to this, the energy efficiency of the wireless system could be improved by the deployment of mMIMO, since the transmit power of a single antenna could be decreased by a factor proportional to the number of antennas as elaborated in [72]. The authors of [67] mentioned that mMIMO increases the system capacity and improves the energy efficiency simultaneously. They also illustrated that the deployment of mMIMO will face some challenges that have to be addressed such as the need for fast processing algorithms.
- **mmWaves:** are waves that occupy the region of the electromagnetic spectrum between 1 and 10 millimetres which corresponds to the band of frequencies between 30 and 300 GHz [73, 74] as shown in Figure 2.2. Although mmWave communications have high propagation loss and blockage sensitivity, they will play a crucial role in 5G networks to support multi-gigabit service such as high definition TV (HDTV) and ultra-high definition videos (UHDV) [75]. In addition, the mmWave spectrum will be suitable for mobile broadband that will enable low-cost fibre replacement with mobile backhauls (wireless backhauls), highly dense small cells with low-interference, [76], indoor, and device to device (D2D) communications

[77]. The authors of [78] explained that by adequate implementation of mmWave communications all the limitations of mmWave communication could be turned into advantages for 5G networks. For instance, short-range communication is one of mmWave limitations but it meets the 5G networks requirements as the short-range communication allows frequency reuse without interference with other cells.

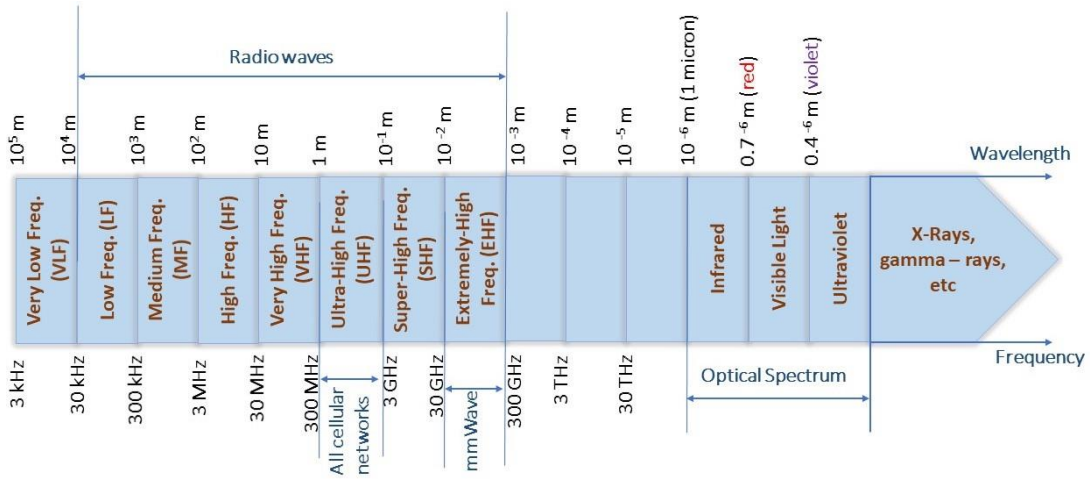


Figure 2.2 mmWave in the electromagnetic spectrum

- **Network Densification:** the dense deployment of many small cells is referred as "Network densification" [79] which is considered a promising technology to provide a high capacity to 5G networks [30] and deals with the explosively increasing number of devices [80] and the user traffic [81]. In term of energy-efficiency, network densification is considered an attractive technology as it reduces the distances between users and BTSs calling for a lower transmit power and a higher data rate [82].

- **Full-duplex communication:** The current wireless systems are generally half-duplex. They can transmit or receive, but not both simultaneously [75]. In contrast to half-duplex, the full-duplex technique enables the communication systems to simultaneously transmit and receive signals at the same frequency band. Full-duplex enables cellular networks to duplicate and improve the flexibility of RF spectrum use [83, 84]. Full-duplex is considered one of the promising technologies that can be employed in 5G networks for two reasons [85]; first, most conventional approaches to increase the spectrum efficiency have been now exhausted such as modulation techniques and MIMO. Second, the dense deployment of small cells in 5G makes the self-interface cancelation problem more manageable than macro-cell networks as the small cells have less transmit power and path loss due to their short coverage.
- **Heterogeneous Network (HetNet):** HetNet is a network which consists of a mixture of different cells with different sizes such as macro and small cells [86-88]. Different types of cells have different coverage as summarised in Table 2.1. Table 2.1 is constructed based on scattered data and parameters in [24, 86, 89, 90]. HetNet is one of the key technologies that will characterise the architecture of 5G networks. It has multi-tier architecture in which the base stations of each tier have different characteristics from the base stations of other tiers such as coverage area, access technology, and transmit power [33] as shown in Figure 2.3. According to [91], 5G networks will be based on dense HetNet architectures (densification) where anchor-booster architecture will be used in case the access to the ideal backhaul (low

latency, high data-rate) is not available. In anchor-booster architecture, the macrocell performs the mobility and control functions of the anchor base station whilst the small cells (micro, pico, femto) boost the data traffic. Although the authors in [67] mentioned that the dense HetNet suffers from inter-cell interference, they explained along with the authors of [92] that dense HetNet will enhance the network capacity and decrease the power consumption in 5G. They explained that the network operators need to deploy advanced power control and resource allocation to eliminate the inter-cell interference, whilst the authors of [93] suggested partial spectrum reuse as an efficient technique to overcome the interference in the HetNet. Heterogeneous networks enable the network operators to exploit the unlicensed spectrum such as Wi-Fi and provide a strong point of integration between the licensed and unlicensed spectrum as two tier heterogeneous network.

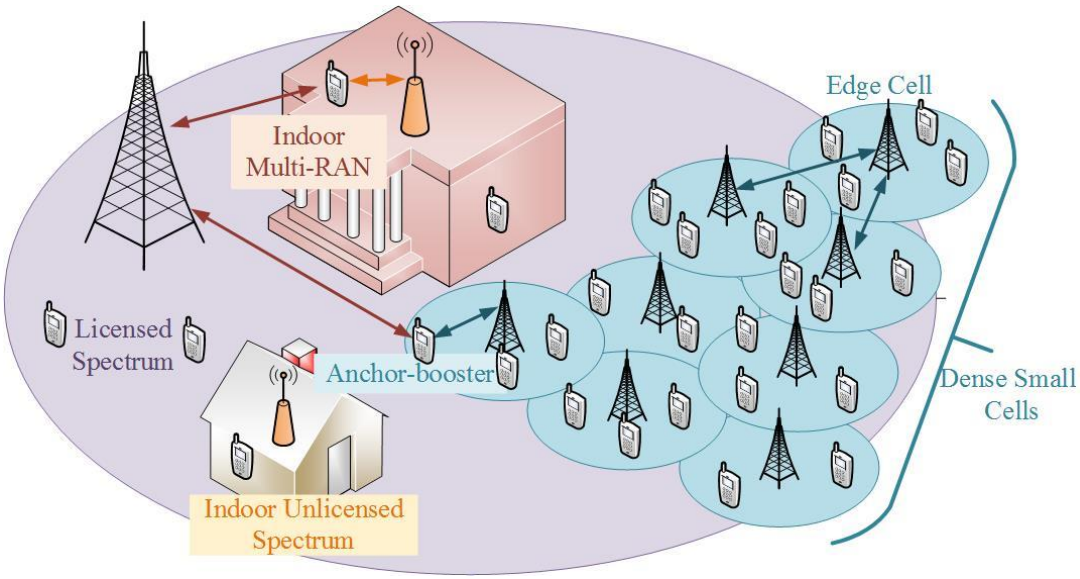


Figure 2.3 HetNet principle

Table 2.1 Cellular cell classification

Cell Type	Typical cell size	Users	Typical BS transmit power
Macrocell	30 - 35 (km)	Many	5 - 40 (W)
Microcell	200 (m) - 2 (km)	100	≤ 2 (W)
Picocell	4 - 200 (m)	20 - 40	0.25 - 2 (W)
Femtocell	10 - 20 (m)	few	≤ 100 (mW)

- **Software-Defined Networking (SDN):** SDN is a versatile technique that decouples the forwarding process (Data Plane) from the network control (Control Plane) in order to facilitate design, management, implementation, and operation of networks [94]. The Open Networking Foundation (ONF) definition of SDN is "*In SDN architecture, the control and data planes are decoupled, network intelligence and state are logically centralized, and the underlying network infrastructure is abstracted from applications* "[95]. Figure 2.4 illustrates the SDN architecture [96]. The decoupling of the forwarding process from the network control is realised by a programmable interface between the SDN controller and the forwarding devices [97] such as the OpenFlow interface [98]. Therefore, the separation of data and control planes along with the programmable interface could bring many merits to the network operators such as improved performance, competitive innovation in

network architecture, and a great level of control and operation [94, 99]. In addition, SDN can greatly support multi-RAN by providing an improved integration of the access technologies, as well as a smooth handover across the access technologies for 5G networks [100]. The authors of [101] explained that the main advantage of the SDN in 5G lies in its ability to provide and create new services and capabilities such as Network Function Virtualisation (NFV).

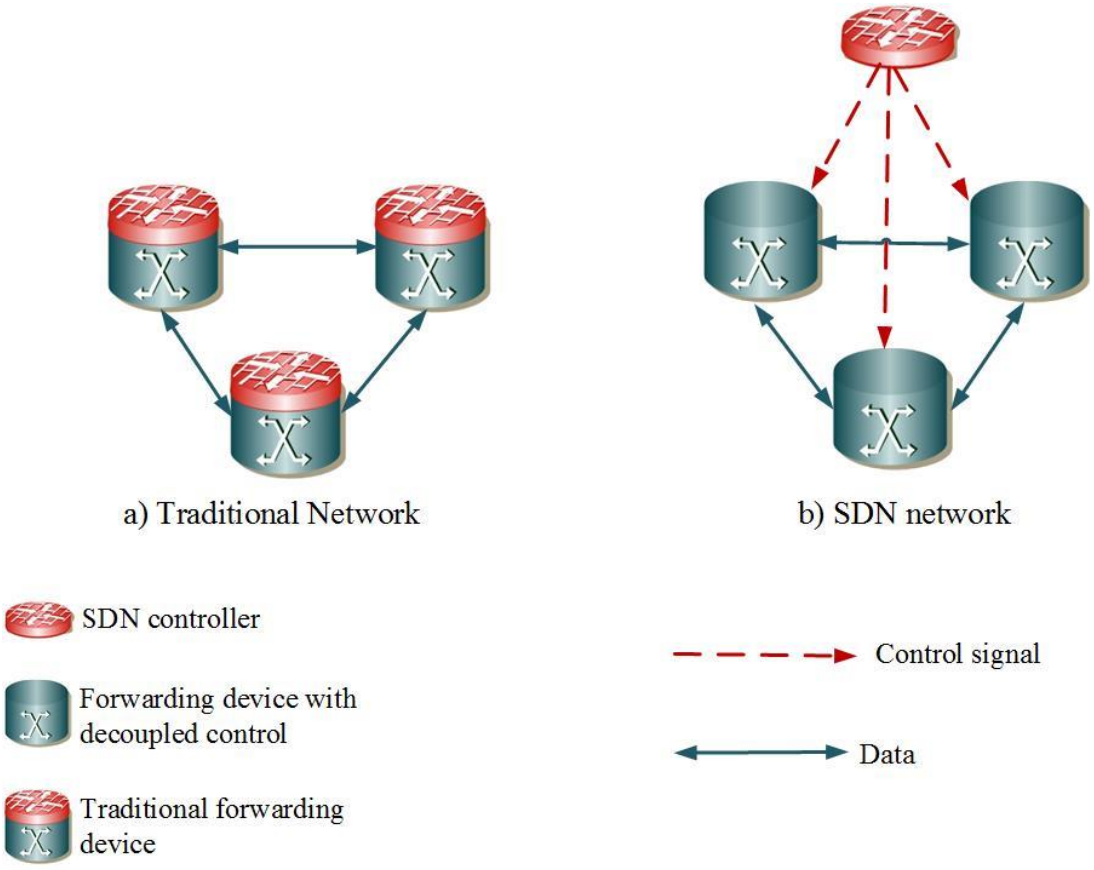


Figure 2.4 SDN architecture

- **Network Function Virtualisation (NFV):** it is an Information Technology (IT) virtualisation technology that aims to alter the way by which the network operators architect their networks through consolidating many

network functions types onto standard servers [102]. NFV is considered one of the key technologies of 5G networks [15]. The definition, advantages, and applications of NFV in 5G will be discussed in detail in the next chapter.

2.4 Optical Networks

2.4.1 Evolution of Optical networks

Optical networks have witnessed revolutionary developments since 1841 when Daniel Colladon did a crucial turn in optical history by illuminating water jets during his lecture on dynamics of fluid [103]. Colladon's experiment had widely opened the doors for the light-guiding competition which continued for years. In the summer of 1880 it was announced in NATURE that Alexander Graham Bell had made a discovery which would rival the telephone and phonograph [104]. Graham Bell and his laboratory assistant, Sumner Tainter, used a sunlight's beam to convey a telephone signal for a distance of 213 meters [105]. Such experiments and demonstrations paved the way to the development of optical fibre in 1960s [106, 107].

The deployment of optical fibres in communication networks provides significant merits for their high bandwidth capabilities, low attenuation and low loss compared with the copper wires. The earlier generation of optical system can be viewed as a point-to-point transmission system. In this generation, a single fibre carries a single wavelength and all network intelligence functions like routing and switching are done in the electrical domain [108]. This requires providing electrical to optical (EO) and optical to electrical (OE) converters at each node. This generation had introduced the synchronous optical network (SONET) standard in 1988 [109] by the

International Telegraph and Telephone Consultative Committee CCITT (now ITU-T) and its international version (SDH). SONET/SDH provided some key advantages over the older telecommunication systems such as simplified multiplexing and de-multiplexing techniques and multi-vendor equipment interconnection [110].

The potential bandwidth of a single mode fibre is around 50Tb/s [111]. This data rate is unlikely to be fully occupied by a single user. Partial bandwidth utilisation is costly and inefficient. Therefore, multiplexing techniques are used to allow optical carrier signals from different sources to travel on single optical fibres. Three main multiplexing techniques exist in optical networks: wavelength division multiplexing (WDM), optical time division multiplexing (TDM), and optical code division multiplexing (CDM) [112]. The most promising multiplexing technique is WDM, which had been introduced during the mid 1990s. WDM supports the routing function in multi wavelength-based networks by either electronic or all optical based switching mode. Optical networks with optical routing capabilities could be considered as the second generation of optical networks [108].

When the demands exceed the capacity of an existing optical fibre, multiple wavelength signals can share a single optical fibre using WDM techniques. Figure 2.5 depicts four channels point-to-point WDM system [111]. In this figure, four wavelengths sourced by four different sources are multiplexed in a single fibre link. The received signal is de-multiplexed at the receiver side and sent to different users.

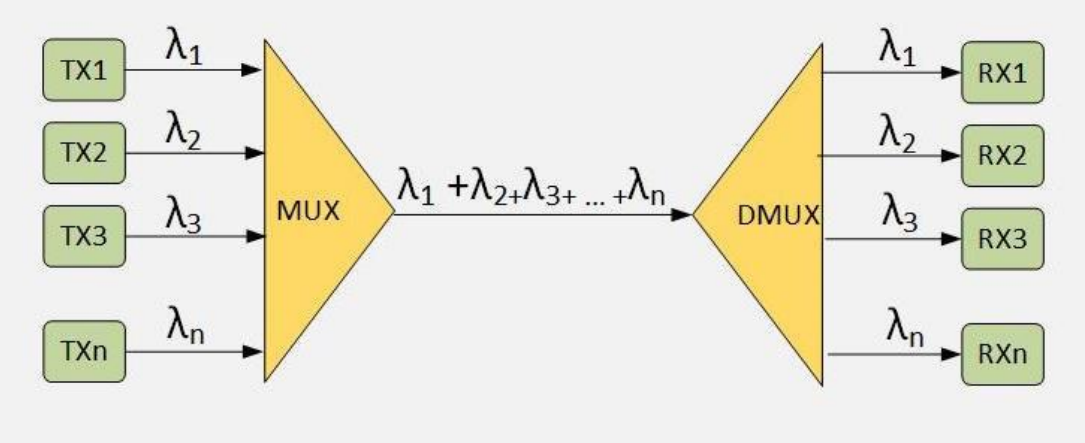


Figure 2.5 Typical four point-to-point WDM system

Many advantages can be provided by the use of WDM technology such as scalability, transparency, and fast dynamic provisioning of network connections [113]. In terms of network transparency, two kinds of WDM are identified: opaque and transparent [114]. The opaque network implies OE and EO conversion where the photonic signal is converted to electrical to be processed and converted back to the optical form. On the other side, transparent networks eliminate the need for OE and EO to process the data whilst the optical devices are used to rely switching and multiplexing functions. WDM can supports a combination of different higher layer technologies such as SONET/SDH, ATM and IP as well as different bit rates such as optical carriers OC-48 and OC-192 as illustrated in Figure 2.6.

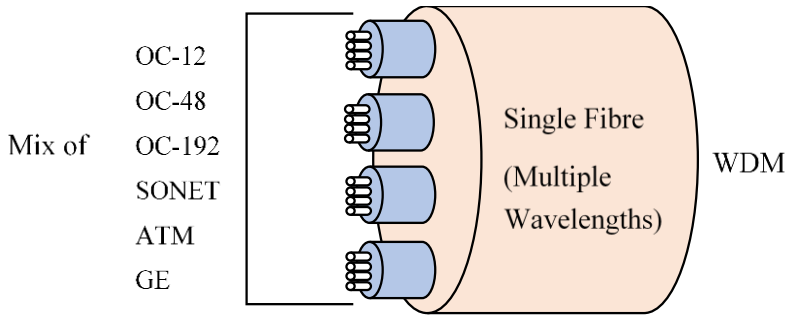


Figure 2.6 WDM interface

Generally, two WDM techniques are deployed in optical networks: Coarse Wavelength Division Multiplexing (CWDM), and Dense Wavelength Division Multiplexing DWDM. A small number of wavelengths (less than 10) can be multiplexed in a single fibre using CWDM techniques [114] whilst DWDM offers multiplexing of huge number of wavelengths (more than 300) in a single fibre [115].

The continuous efforts to avoid the electronics bottleneck and overcome the needs for OE-EO conversion had led to the introduction of all-optical networks (AONs). All-optical network refers to the transparent optical WDM network [116] which had opened the doors widely for the third generation of optical systems [117] characterised by optical switching techniques.

2.5 Optical switching techniques

Researchers' passion to offer all optical networks and replace the electronic switches and routers by all optical elements have led to the introduction of a number of optical switching techniques which are summarised as follows:

2.5.1 Optical circuit switching

Electrical network elements such as switches and routers are incapable of processing the huge data transmitted in the optical domain. Therefore, this challenge

has led to the introduction of Optical Circuit Switching (OCS) [118]. Although OCS eliminates the need for O-E-O conversion, it has some drawbacks such as the two-way lightpath reservation [116, 119].

2.5.2 Optical packet switching

With the increased use of the Internet and the need for large transmission capacity, service providers and network operators have to tackle the large consumption of energy and the lack of capacity of the electronic switching and routing networks [120, 121]. In contrast to OSC, Optical Packet Switching (OPS) was developed to exploit the transmission bandwidth efficiently and decrease the latency by providing a connectionless service.

There are many challenges that stand against the implementation of pure OPS such as the absence of optical RAM and the difficulty of implementing optical Arithmetic Logic Unit (ALU). Therefore, to implement OPS, the switching process must ensure that the processing of the payload and data switching are carried out within the optical domain without the need for O-E-O whilst the packet header is electronically processed.

2.5.3 Optical burst switching

Optical burst switching (OBS) is a hybrid of both OCS and OPS techniques. It is a promising technique that combines the advantages of both OCS and OPS and eliminates their drawbacks [122, 123].

The basic transmission unit in OBS is the “burst” [124]. Bursts are assembled by aggregating packets at the source node (ingress) and sent to the destination nodes (egress) where they are disassembled to their original packets [125, 126]. There are many proposed algorithms for burst aggregation based on the time interval of the

aggregation process (time-based aggregation) and the burst size limits (burst-length based aggregation). Algorithms based on both factors exist and are referred to as burst-length/time based aggregation [126, 127].

2.6 IP over WDM Network

Packet-based communication is rapidly growing today with the dramatic increase in daily use of Internet applications. Various types of packet-based networks such IP over WDM networks are deployed to meet the needs for high bandwidth communication channels and data rates. The packet header of the complex multi-layer IP network such as IP over ATM and IP over SDH occupies around 25% of the bandwidth [128]. The insufficient bandwidth utilisation of such networks resulted in the evolution of the IP over WDM network.

IP over WDM network is composed of two layers: IP layer and optical layer [129, 130]. The IP layer is responsible for services whilst the optical layer is in charge of the high bandwidth provisioning. The IP layer consists of a number of core routers that aggregate the traffic from end routers. These routers are connected to the optical layer. The optical layer is the WDM network which consists of a group of optical switches connected to each other through fibre links. For each fibre link a pair of multiplexer / de-multiplexer is deployed for wavelength multiplexing/de-multiplexing [129]. In addition, Erbium-Doped Fibre Amplifiers (EDFAs) are used to enable long distance transmission of optical signals. For each optical channel, a pair of transponders are provided for end to end lightpath data transmission [131].

Mainly, there are two approaches for traffic forwarding in IP over WDM networks: Bypass and Non-Bypass. In the bypass approach the lightpath avoids

passing through the intermediate core routers of the IP over WDM network, unlike the non-bypass approach [132]. The IP over WDM network architecture is shown in Figure 2.7.

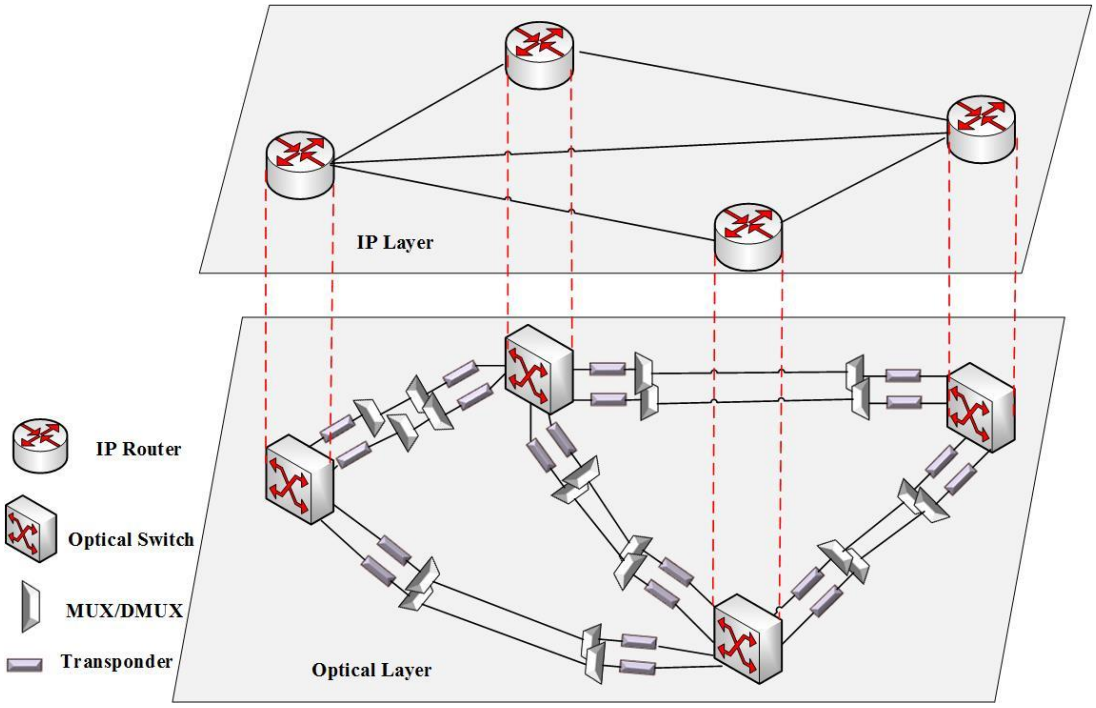


Figure 2.7 IP over WDM architecture

2.7 Passive Optical Networks

Subscriber access network, usually referred to as Access Network (AN), is the part of the telecommunication network infrastructure that connects the subscribers to the service provider central offices (CO). It is well known as “the Last Mile” or “The First Mile” as called by the Ethernet community [133]. Because of the growing demand for higher bandwidth, the access network has been considered as a bottleneck in the telecommunication network infrastructure that is known as “the last mile problem” [134]. Access network providers are making significant

investment in fibre-to-the-home technology (FTTH) and broadband wireless access to tackle the last mile problem. Among these investments is the Passive Optical Network (PON) that is well-known for its cost-efficient capabilities to carry gigabit data rates [134].

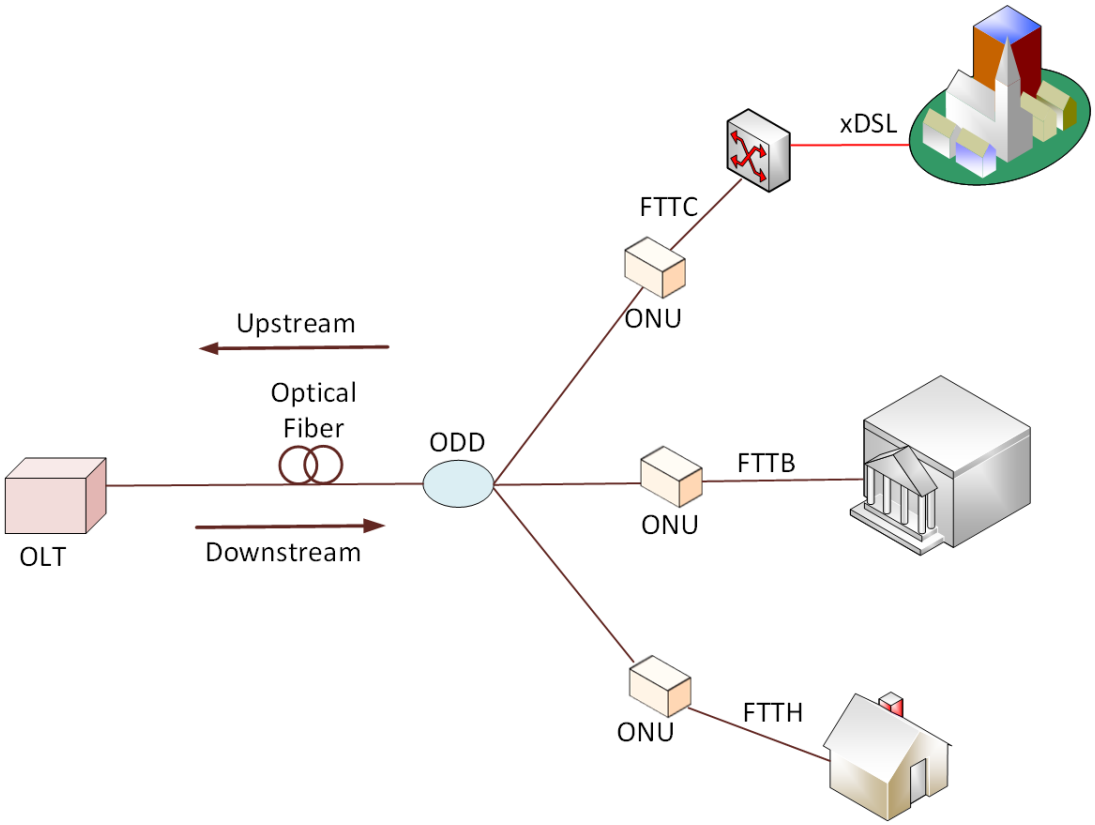


Figure 2.8 Basic PON architecture

PON networks are attractive to access network providers because of the non-active elements (passive) in the transmission line [135]. The general architecture of PON network is depicted in Figure 2.1 [136]. The PON network consists of three elements: the Optical Line Terminal (OLT), the Optical Network Unit (ONU) and the Optical Distribution Network (ODN) [137, 138]. OLTs are deployed at the CO whilst ONUs are deployed close to the subscribers. Depending on the multiplexing techniques, three types of PONs can be identified today [139]; these are:

2.7.1 Time Division Multiplexing PON (TDM-PON)

At the early stages of the work on optical access networks, the Full Service Access Network (FSAN) consortium recommended time division multiplexing passive optical network TDM-PON in 1990 [140]. This recommendation was adopted in 1996 by International Telecommunication Union Standardisation (ITU-T) as ITU-T G.983.1 and referred to as ATM PON (APON) [139] which is evolved to Broadband PON (BPON) [141]. APON and BPON provides data rates of 155 Mbps for upstream traffic and 622 Mbps for downstream traffic [142]. The TDM-PON is most popular nowadays among other types of PON because of its cost efficiency and feasibility [143]. In 2003, ITU-T ratified new standards for PON which were included in the G.984 series of ITU-T recommendation, called Gigabit PON (GPON) [144]. GPON supports a mix of ATM, TDM, and Ethernet services [144] using Gigabit PON Encapsulation Method (GEM). GPON provides 1.25 Gbps downstream data rate and 2.55 Gbps upstream data rate [145]. In 2004, IEEE 803.2ah group standardised an Ethernet-friendly technology PON called Ethernet PON (EPON) [146] which provides 1.25 Gbps data rate in both down and upstream [144]. EPON grabbed enormous attention in eastern countries such as Japan, Korea, China, and Thailand [139]. In 2007 IEEE 802.3av increased the downstream of EPON to 10 Gbps to introduce 10G-EPON which supports two downstream data rates of 1 Gbps and 10 Gbps [140] whilst in 2009 10G-EPON improved to support asymmetric 10 Gbps for both down and upstream bandwidth as well as 1 Gbps upstream [147]. ITU-T published in 2010 the first generation of 10 Gigabit-capable Passive Optical Network (XG-PON1) within the standard G987 to offer 10 Gbps downstream and 2.5 Gbps upstream data rates [148]. The standard G987 also refers

to a second phase in the development of XG-PON2 to offer symmetric 10 Gbps for both up and downstream bandwidths [149]. XG-PON1 and XG-PON are also referred to as the Next Generation of PON networks (NG-PON) whilst the letter X refers to the Latin number 10 [148].

2.7.2 Wavelength Division Multiplexing PON

Although WDM-PON architecture has been proposed in the mid of 1990s, it has not been commercialised yet [150] for many reason among them the high installation and maintenance cost [151]. In WDM-PON architecture peer-to-peer (P2P) connectivity is achieved via dedicated wavelength channels between the OLT and the individual ONUs [147]. As WDM-PON supports multiple wavelength channels over single fibre, it provides an excellent scalability. There are a number of variations of WDM-PON such as Dense WDM-PON (DWDM-PON) and TDM-WDM PON (TWDM-PON). WDM-PON is one of the PON architectures that have been suggested for the next generation PON (NG-PON) [152].

2.7.3 OFDM-PON

OFDM-based PON is considered as an effective paradigm in optical access network for its features that meet the NG-PON [153]. OFDM-PON is considered as a point to multi-point (P2MP) system with one wavelength for downstream and another for upstream [152]. It is one of the concepts that are suggested to realise the NG-PON paradigm [149].

2.8 Summary

This chapter provided a review of 5G networks. It has shed the light on the requirements of 5G networks and the key technologies that are suggested to cope with these requirements. In addition, this chapter has presented the optical network including IP over WDM networks and PON networks.

Chapter 3 : Background: Network Function Virtualisation, Energy Efficiency, and Optimisation Problem Formulation

3.1 Introduction

Network function virtualisation has been identified as a solution to the gradual ossification of the Internet architecture [154].NFV provides a high degree of flexibility, on demand resources allocation, and easy resource management by allowing multiple network function to coexist on a same hardware infrastructure. In addition, NFV improves the energy efficiency by consolidating more than one function on a single hardware [155]. For instance, different processors have different cycles and operations per second (CPS/OPS) with different power consumption. However, a mobile function that run on low (CPS/OPS) and high power consumption processor could be virtualised and run alongside with another virtual mobile function on high (CPS/OPS) and low power consumption processor. In addition, virtualising mobile core functions and hosting them close to the users will reduce the amount of traffic flows in the network toward mobile centre office which calls for less traffic induced power consumption. This chapter provides a review of virtualisation and network function virtualisation. Also, it makes an important review the efforts that have been made so far to improve the energy efficiency in 5G. In addition, it introduces the approach that is used to formulate and solve the optimisation problem.

3.2 Virtualisation

Virtualisation is a technique for hardware abstraction which hides the physical resources from other systems and users [156]. In most of the virtualisation techniques, the resources abstraction is realised by a software layer that lies between the hardware and the operating system which is called Virtual Machine Monitoring (VMM) or a hypervisor [157]. The VMM logically divides the hardware platform into one or more logical units called Virtual Machines (VMs) [158]. VMs were introduced by IBM in 1960s to provide interactive and concurrent access to mainframes [159]. The first official VM product was announced by IBM on August 2, 1972 and it was called VM/370 [160].

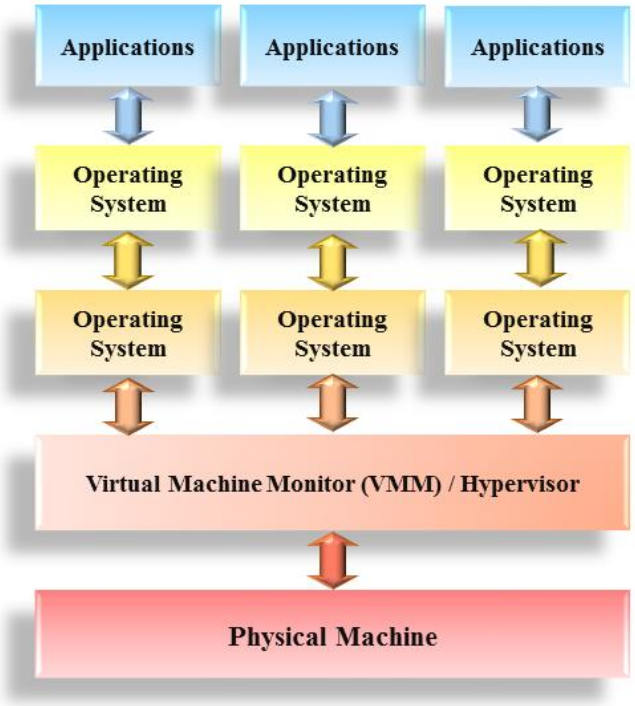


Figure 3.1 Virtualisation principles

Figure 3.1 illustrates the principles of virtualisation, VMM, and VMs [161]. Virtualisation provides flexible resources provisioning by managing and viewing

these resources as a pool that provides the user with the exact amount of needed resources [162]. The following are among the main benefits of virtualisation:

- **Consolidation:** Virtualisation aims to merge and bring workloads together on physical platforms to increase the hardware utilisation [163, 164].
- **Reliability and resilience:** In virtualisation, each VM is isolated from others, which means any failure or breakdown in one VM does not influence other VMs. Therefore, virtualisation ensures system operation and function availability for users and operators [164, 165]. Failure could be categorised into: VM failure, host failure, and link failure [166]. In case of VM failure, another VM can migrate and replace the faulty VM whilst in host failure, the VM could be migrated to a healthy node. In link failure, either the VM is migrated to another node through different link or another VM could be triggered to take over the job. Therefore, virtualisation is considered to be a reliability and resilience enabler [167].
- **Cost efficiency:** With consolidation characteristics and combined workload, virtualisation can reduce the hardware cost [168] by a factor of 29% to 64% [165].
- **Energy efficiency:** Energy efficiency is one of the main concerns in both industry and academia. According to the authors of [169], the Information and Communications Technology (ICT) contributes around 2-4% of the total carbon footprint produced by human activities. Accordingly, a number of techniques have been suggested and adopted in the literature to confine the expected growth in power consumption and virtualisation among these techniques. Virtualisation and NFV contribute to the energy-efficiency

through two main ways: (1) function abstraction and virtualisation, and (2) VM consolidation in a single hardware. In function abstraction and virtualisation, a mobile function that runs on high power consumption and low number of cycles per second processor could be virtualised and run alongside with another virtual mobile function on low power consumption and high number of cycles per second processor. Consolidation is a technique that reduces the number of active physical machines by packing virtual machines into one physical machine. However, consolidation reduces the number of active physical machines which calls for less power consumption.

According to [170] and [171], virtualisation can improve the energy efficiency of servers by packing more than one VM in one sever. In [172], the authors optimised virtual network embedding and the placement of VMs and contents in distributed clouds to minimise the energy consumption in IP over WDM networks.

3.2.1 Types of Virtualisation:

As alluded earlier, virtualisation is not a new concept, in fact it was traced back to the 1960s. Since that time, virtualisation has been adopted in many sectors and it gave birth to new technologies such as “Cloud Computing”. In computer science, virtualisation played a very important role where “memory virtualisation” in the 1970s was the first adoption of the virtualisation concept in computers as the memory was the most expensive part of the computers at that time [173]. The virtualisation of computer memory inspired researchers to look forward to

virtualising other parts of the computer which resulted in different flavours of virtualisation. The following are among the main types of virtualisation:

3.2.1.1 Storage virtualisation

In storage virtualisation, multiple physical storage devices are pooled into a single virtual storage resource that is centrally managed. The resulting storage resources appear as one logical disk for users, machines, or operating systems [164, 174] as shown in Figure 3.2. Storage virtualisation simplifies the storage administration, for instance, the administrators are able to compensate the amount of storage for each user online, which means there is no need for power cycle (OFF/ON) in order to manage the storage capacity like in conventional storage units [175].

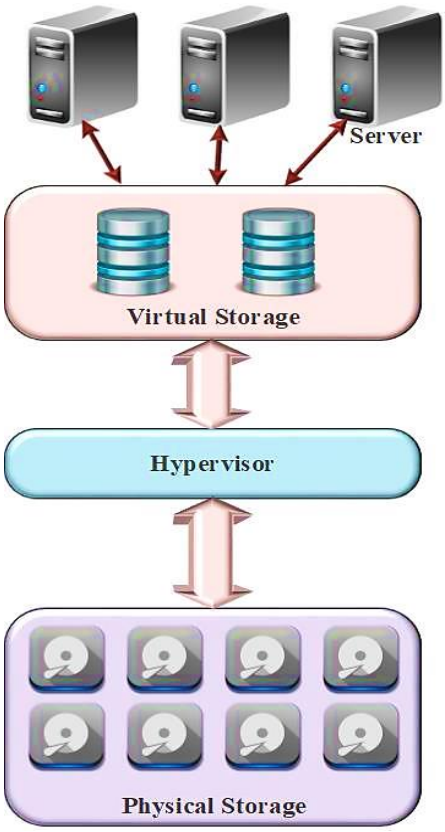


Figure 3.2 Storage virtualisation

3.2.1.2 Network Input-Output (I/O) virtualisation

This is a member of I/O virtualisation (IOV) family which aims to virtualise the Network Interface Card (NIC). Network I/O virtualisation plays a significant role in cost saving, performance improvement, and managements of servers [176]. With aggregation (bonding) capabilities provided by IOV, where one logical network interface can combine multiple physical interfaces [177]. Network bridging and bonding are strongly exploited by VMs for network IOV. VMs can have more than one virtual network interface card (VNIC) that communicate with outer network through bridging whilst bonding aggregates the physical NICs to provide an aggregate logic to the VMs as shown in Figure 3.3 [178].

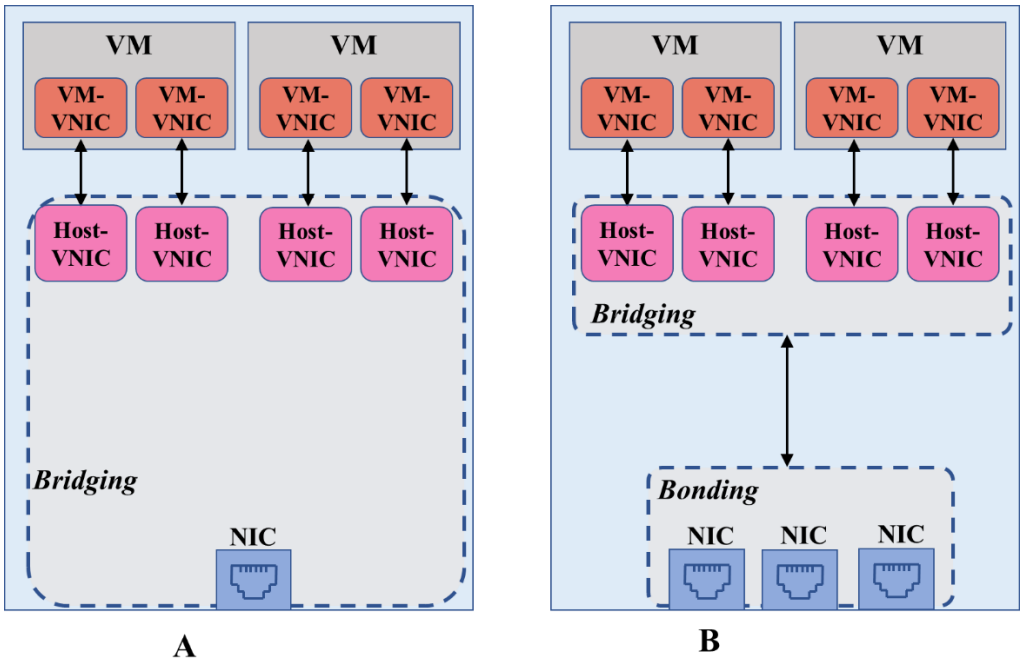


Figure 3.3 NIC, VNIC A) Bridging, B) Bridging and Bonding

3.2.1.3 Operating system virtualisation:

In operating system virtualisation; or as it is known sometimes as “a container-based virtualisation” [179], multiple isolated user-space instances (or containers) are permitted to run on the top of the host operating system and interact with applications through a set of libraries. These libraries are provided by the containers to delude the applications that they are running on their dedicated machines [180]. The design of operating system virtualisation is schematically depicted in Figure 3.4 [181].

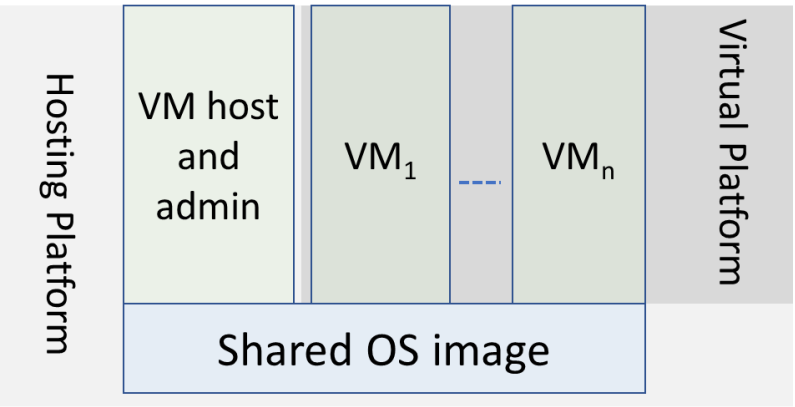


Figure 3.4 OS virtualisation

3.2.1.4 Application virtualisation

The growing needs to improve the security and availability of IT services and to deliver applications in any-device anywhere results in introducing application virtualisation [182]. In application virtualisation, the applications are encapsulated into containers along with a set of system files associated with these application in order to provide isolation and portability to these applications across different

computers [183, 184]. The application virtualisation basic concept is illustrated in Figure 3.5 [185].

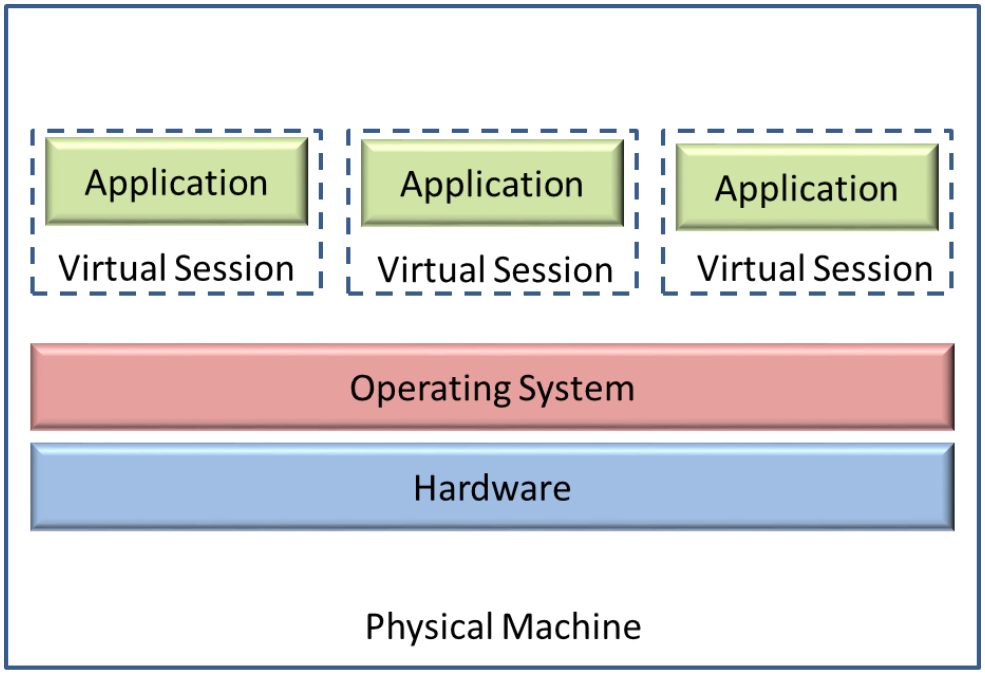


Figure 3.5 Application Virtualisation concept

3.2.2 Network Function Virtualisation

In October 2012, an Industry Specification Group (ISG) within the European Telecommunication Standards Institute (ETSI) published a white paper at "SDN and OpenFlow World Congress" conference in Darmstadt-Germany titled "***Network functions virtualisation****: An introduction, benefits, enablers, challenges and call for action*" [102]. Later in November 2012 ISG NFV was founded within ETSI to become the home of the industry specification group for NFV [186]. NFV aims to consolidate many network equipment types onto standard IT servers, network elements, and storage that could be located anywhere in the network [102]. Simply,

NFV converts the network function from hardware-based functions to software-based function as shown in Figure 3.6.

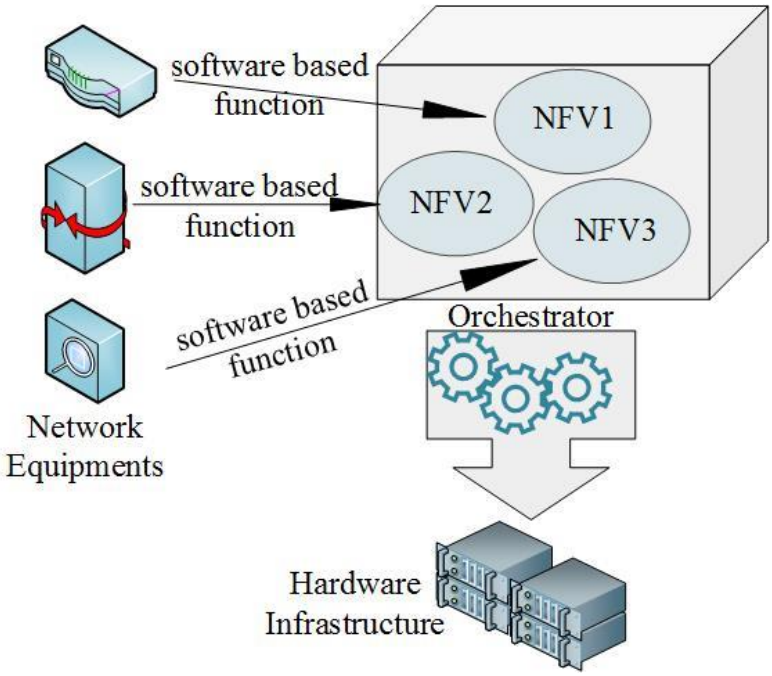


Figure 3.6 NFV concept

Network function virtualisation has been identified as a promising technology to improve the network service in the existing networks. Since it allows multiple heterogeneous network architecture to share the physical resource, NFV provides service flexibility and increases the network service scalability [187]. NFV is still in its infancy compared with other techniques such as server virtualisation which has enabled notable computing functions efficiencies to the benefit of the service providers. Therefore, in few years NFV is expected to bring all the benefits of other virtualisation techniques to the world of networking.

Researchers in both academia and industry embraced NFV at unprecedented speed although it is still at its early stages. This resulted in enormous work that could be categorised into three groups:

The first group of researchers were concerned with NFV reviews and surveys such as the work done in [155] where the authors compared NFV with other fields such as SDN and cloud computing and reviewed a number of research activities and suggested combining them.

The second group of researchers were concerned with integrating NFV with other paradigms such as the authors of [188]. They illustrated that due to the highly specialised hardware, interfaces, and control of the optical devices, the classical optical networks are becoming inefficient regarding the scalability and resource utilisation. They explained the need for flexible and programmable optical devices that are introduced by the combination of NFV and SDN to tackle the scalability and resource utilisation problems in optical networks. On the other side, the authors of the [189] showed in their work that the optical/electrical/optical OEO conversion can be effectively minimised in optical data centres by the employment of NFV chaining.

The last group of researchers were concerned with algorithms development for NFV such as NVF placement, migration and scheduling such as the authors of [190]. They developed a distributed orchestrator prototype in order to manage the virtual infrastructure in highly dynamic networks and services where the virtual nodes and links are added and removed according to the traffic and service requests.

Apart from these categories or the way the researchers introduced their ideas or tackled the problems; they all agreed that the deployment of NFV will bring many merits. In optical networks, the authors of [191] elaborated the implementation challenges of joint scheduling of computing and network resources in grid computing over optical network. They proposed a virtual optical network framework to tackle the joint scheduling challenges. In [192] the authors applied the concept of network virtualisation to the optical network of the data centres and they elaborated that the optical network virtualisation provides the data centres users with high data rate, low latency, and on demand service provisioning.

There are also opportunities for introducing NFV in 5G networks. The authors of [193] focused in their work on improving the Quality of Experience (QoE) of the 5G users. They proposed a video quality assessment method (VQA) as a virtualised network function to tackle the quality degradation caused by the small cells backhaul bottleneck. They used Long Term Evolution (LTE) infrastructure to evaluate the proposed virtualised VQA method. In [20] the authors proposed a hierarchal layered SDN architecture for 5G networks. In this architecture, the hardware functionalities of the baseband computation are provided as NFV and are pooled in a centralised architecture. In addition, the multiple remote radio head (RRH) in this architecture are connected to the baseband unit pool by a virtualised radio aggregation units (RAU). On the other side, the author of [194] proposed NFV for the mobile core of the future 5G networks. They explained that the mobile core network experiences scalability limitations as it is populated by hardware-based functions. Therefore, to cope with these limitations, the mobile core functions need to be freed from the hardware bands. To do so, NFV was suggested to mitigate the

dependency on the hardware. In [195] the authors suggested an integrated NFV/SDN orchestrator for dynamic backhaul deployment on a multi-layer optical network to support a number of mobile operators to flexibly manage and expand their network on a unified architecture. In their proposal, they virtualised the mobile core network and provided it as a virtual network function VNF to support a number of mobile network operators (MNO) with different services requirements. In addition, they virtualised the control plane of the SDN that controls the provisioning of backhaul network resources to the mobile operators. The authors in [196] explained that the current mobile networks suffer from lack of elasticity to deal with the highly dynamic traffic which might result in resources wastage. They proposed NFV as a credible enabler to tackle the elasticity problem in the mobile networks and they provided the mobility and management entity (MME) as a VNF. They carried out a scalability analysis to the virtualised MME to determine the minimal number of VMs required at the data centre to ensure the required system response time is satisfied.

3.3 Energy efficiency

As alluded earlier, 5G networks will serve an enormous number of devices providing them with ubiquitous connectivity, high data rate and low latency. By 2020 there will be around 50 billion connected devices [48] including IoT devices. The vision is to have a connected society where everything is connected to the Internet even possibly cows and shoes [197]. In order to serve such a huge number of connected devices, the capacity provided by 5G networks is required to be 1000 times higher than the capacity provided by the current networks [198]. To achieve

such an ambitious goal, a revolutionary architecture for the next generation of mobile networks is needed. However, bandwidth-hungry applications and massive number of connected devices call for increase in the power consumption in 5G networks. Therefore, energy efficiency is among the primary concerns in the design and operation of future 5G networks.

There are many studies around the world that focus on improving the energy efficiency in 5G networks. These studies can be grouped into main four categories [80] listed as follows

3.3.1 Energy harvesting related research

The process of harvesting the energy from natural resources to operate the communication system is a promising approach to enable the mobile networks to be run on clean and renewable energy sources. The authors of [199] investigated green energy enabled mobile networks and they focused in their work on powering the mobile base stations on green power. They elaborated how to design green energy powered base stations and how to optimise their resource management. The authors of [200] developed an algorithm that integrates an intelligent renewable energy system with renewable energy-based ON/OFF mechanism for the network base stations in order to save energy consumption.

3.3.2 Network planning and deployment

Network planning and deployment aim to ensure that the future mobile network is able to cope with the implications of the huge number of connected devices where the energy efficiency is among these implications.

Adding more cells to the mobile network in order to increase the network capacity (network densification) is one of the promising approaches in 5G to cope

with the large number of users. In the current mobile network, network densification is based on the deployment of small cells in the coverage of macro-base stations. However, as macro-base stations consume high power, the vision in 5G is to drastically increase the number of heterogeneous energy-efficient infrastructure nodes such as femto and picocells and eliminate the macro-base stations. Network densification reduces the distances between the users and the small base stations (femto and pico base stations) which ultimately results in low transmission power consumption. The major challenge of network densification is interference. This challenge was discussed and addressed in the work in [201] where the authors analyse the trade-off between energy efficiency and interference.

Other key technologies to boost the energy-efficiency in 5G mobile networks are traffic offloading techniques such as device-to-device (D2D) communication techniques and local caching techniques. The authors of [202] proposed an approach to improve the D2D communication through a resource allocation scheme to optimise the terminals battery life. The authors of [203] developed an auction based algorithm for D2D communication to improve the energy efficiency. In this algorithm, the devices act as bidders for channel resources whilst the cellular network acts as the auctioneer.

3.3.3 Resources allocation

The system radio resources allocation is considered a promising technique to increase the energy efficiency in mobile networks [82]. In [204] the authors proposed an energy-efficient resource allocation algorithm that optimises the power allocation and channel allocation separately for IoT in 5G networks.

3.3.4 Hardware solution

Hardware-based energy efficient solutions include a broad category of approaches such as energy efficient design of power amplifiers and hardware virtualisation. Virtualisation in computing has showed the network operators and service providers how the abilities of hosting multiple VMs onto a single standard server can improve the energy efficiency. Importing the same concept to networking, instead of having only one network function in a single physical device, this device could have several virtual network function (VNF). Virtualisation is one of the enabling technologies in 5G networks that can improve the energy efficiency [205]. The authors in [206] proposed a workload consolidation framework using virtualised general purpose processors in the radio access cloud to minimise the power consumption in radio access networks.

3.4 Mixed Integer Linear Programming (MILP) and Network Modelling Problem

3.4.1 Mixed Integer Linear Programming (MILP)

Linear Programming (LP) is a mathematical optimisation technique and a special case of mathematical programming [207]. It is one of three classes of constrained optimisation (linear, non-linear, and integer) where all its mathematical expressions (equations and inequalities) are linear [208] as its name implies. The standard form of linear programming model generally comprises four constituents:

- The optimisation outcome typified by the objective function where the optimisation problem seeks to minimise or maximise it depending on whether the model outcome is a cost or a reward;

- ❖ Model variables where their values represent the feasible solution of the objective function when all the constraints are met and optimal solution if they are the best objective values [209]. When one of these variables is a non-integer, the optimisation technique is called Mixed Integer Linear Programming (MILP);
- ❖ Set of linear mathematical expressions (equations and inequalities) known as constrains that draw the shape of feasible region of solution (polyhedron region);
- ❖ Variables' boundaries that control the upper and lower limits of each variable in the model.

To put all these constituents together in a problem formulation, consider the following problem [210, 211]

Objective function

$$z = c^T \cdot x \tag{3.1}$$

Subject to constraints

$$A \cdot x \geq b \tag{3.2}$$

and non-negativity constraints

$$x \geq 0 \tag{3.3}$$

where:

z is the objective function

c^T is the transport vector of the cost / reward coefficients vector given by

$$c = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{bmatrix} \tag{3.4}$$

x is the vector of decision variables given by

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \tag{3.5}$$

A is the constraint matrix given by

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \tag{3.6}$$

b is the right-hand side vector which represents the minimal requirement to be satisfied and is given by

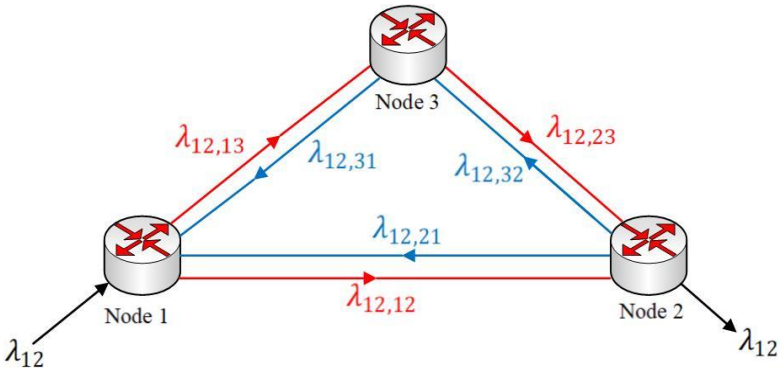
$$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix} \tag{3.7}$$

Whilst there are many approaches to solve constrained optimisation problems, systematic approaches are used to solve linear programming problems due to convexity of these problems. The most known approach among these approaches is *Branch-and-Bound* (B&B) approach [212, 213]. B&B as a systematic approach is a methodological approach that repeatedly and intelligently searches the polyhedron region of all feasible solutions. The B&B approach repeatedly divides the polyhedron region into smaller subsets. The upper or lower bound is calculated each time within each subset and any subset that exceeds the cost / reward of a feasible

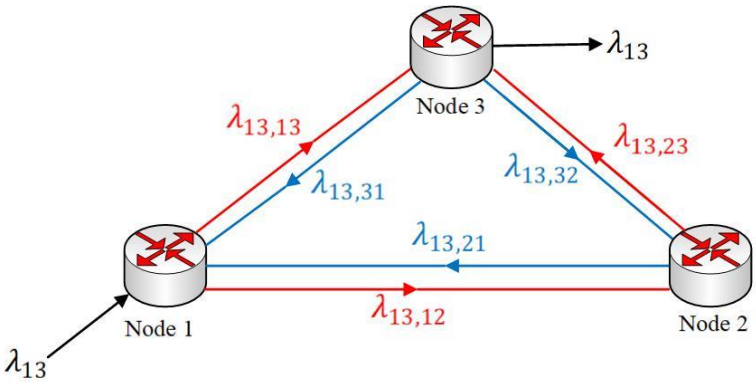
solution is excluded. Once there is a feasible solution whose cost /reward does not exceed the bound of any subset, then the partitioning process halts. It is worthy to mention that B&B is referred to by different names such as *divide and conquer*, *implicit enumeration*, or *separation and evaluation*.

3.4.2 Network Modelling Problem

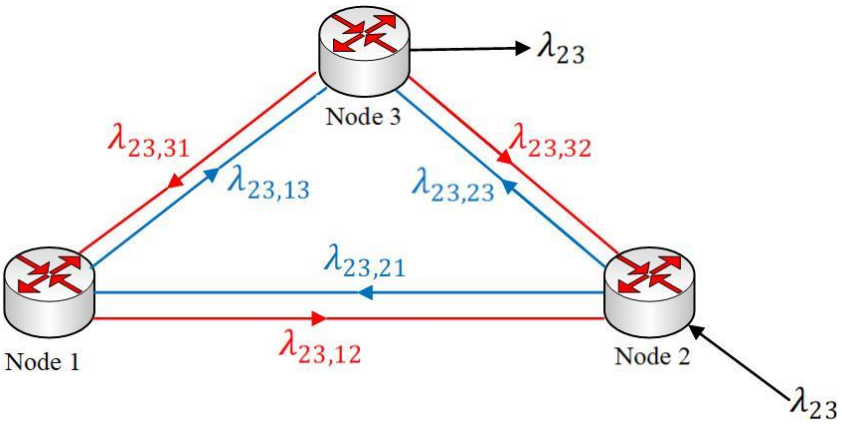
In communication networks, there are many ways to formulate the network optimisation problem using linear programming. In this work, we use node-link formulation to design the models. In this formulation both demands and links are usually directed, and the total link flow is considered on each link. The total traffic flow leaving a node minus that entering a node is considered zero except at the source and destination nodes [214]. The nodes between each two end nodes of a considered demand are referred to as *intermediate* or *transit* nodes. For each intermediate node in the considered demand, the link flow enters the node from the incoming links and are sent out on the node outgoing link. In other words, the total link flow at the incoming links equals to the total flow at the outgoing links for each intermediate node of the considered demand and this is called *flow conservation* law. According to that, if the node is the source node of the considered demand, then the demand volume is equal to the total outgoing flow, whilst it equals to the total incoming flow if the node is a sink node.



a) Demand flow from node 1 to node 2



b) Demand flow from node 1 to node 3



c) Demand flow from node 2 to node 3

Figure 3.7 Demand flows example in three nodes network

To illustrate the node-link formulation, consider a three node network shown in Figure 3.7 with three demand volumes: from node 1 to node 2 (λ_{12}), from 1 to node 3 (λ_{13}) and from node 2 to node 3 (λ_{23}). By applying flow conservation at node 1 for demand volume λ_{12} and using the convention that anything which goes into the node is negative and anything that leaves is positive, the following equation is obtained:

$$-\lambda_{12} + \lambda_{12,13} + \lambda_{12,12} - \lambda_{12,31} - \lambda_{12,21} = 0 \tag{3.8}$$

where:

$\lambda_{12,13}$ is the demand volume from node 1 to node 2 that flows through the link from node 1 to node 3;

$\lambda_{12,12}$ is the demand volume from node 1 to node 2 that flows through the link from node 1 to node 2;

$\lambda_{12,31}$ is the demand volume from node 1 to node 2 that flows through the link from node 3 to node 1;

$\lambda_{12,21}$ is the demand volume from node 1 to node 2 that flows through the link from node 2 to node 1;

λ_{12} is the total demand volume from node 1 to node 2.

The following equations are obtained by applying flow conservation at both node 2, and 3 for the demand volume λ_{12}

at node 2

$$-\lambda_{12,12} - \lambda_{12,32} + \lambda_{12} + \lambda_{12,21} + \lambda_{12,23} = 0 \tag{3.9}$$

at node 3

$$-\lambda_{12,13} - \lambda_{12,23} + \lambda_{12,32} + \lambda_{12,31} = 0 \tag{3.10}$$

For the demand volume λ_{12}, note that node 1 is the source node, node 3 is an intermediate node, whilst node 2 is the sink node. Therefore:

$$\lambda_{12,21} = \lambda_{12,31} = \lambda_{12,23} = 0 \tag{3.11}$$

and the previous flow conservation equations at nodes 1, 2, and 3 for the demand volume λ_{12} could be rewritten as:

$$\begin{array}{cccc} \lambda_{12,12} & +\lambda_{12,13} & = & \lambda_{12} \\ & -\lambda_{12,13} \quad +\lambda_{12,32} & = & 0 \\ -\lambda_{12,12} & -\lambda_{12,32} & = & -\lambda_{12} \end{array} \tag{3.12}$$

By using the same methodology, the system equations could be written for the demand volumes λ_{13} as

$$\begin{array}{cccc} \lambda_{13,12} & +\lambda_{13,13} & = & \lambda_{13} \\ -\lambda_{13,12} & +\lambda_{13,23} & = & 0 \\ & -\lambda_{13,13} \quad -\lambda_{13,23} & = & -\lambda_{13} \end{array} \tag{3.13}$$

and for the demand volume λ_{23}

$$\begin{array}{cccc} \lambda_{23,21} & +\lambda_{23,23} & = & \lambda_{23} \\ -\lambda_{23,21} & \lambda_{23,13} & = & 0 \\ \lambda_{23,13} & -\lambda_{23,23} & = & -\lambda_{23} \end{array} \tag{3.14}$$

Assume that each undirected link has two arcs, and the demand flows in one of these two arcs only. Hence, we can add extra constraints to the model by considering the capacity of each link as in Figure 3.8. If the capacity of the link from node 1 to node 2 is c, then the capacity constraint is expressed as:

$$\lambda_{12,12} + \lambda_{13,12} \leq c_{12} \tag{3.15}$$

By writing the analogous inequalities for other links in the network and putting everything together, we will have the following system model:

$$F = \lambda_{12,12} + \lambda_{12,13} + \lambda_{12,32} + \lambda_{13,12} + \lambda_{13,13} + \lambda_{13,23} + \lambda_{23,21} + \lambda_{23,13} + \lambda_{23,23} \tag{3.16}$$

where F is the model objective function to be minimised (here as it is the cost of routing traffic in the network)

subject to

$$\begin{array}{cccccccccccc}
\lambda_{12,12} & +\lambda_{12,13} & & & & & & & & = & \lambda_{12} \\
 & -\lambda_{12,13} & +\lambda_{12,32} & & & & & & & = & 0 \\
-\lambda_{12,12} & & -\lambda_{12,32} & & & & & & & = & -\lambda_{12} \\
 & & & \lambda_{13,12} & +\lambda_{13,13} & & & & & = & \lambda_{13} \\
 & & & -\lambda_{13,12} & & +\lambda_{13,23} & & & & = & 0 \\
 & & & & -\lambda_{13,13} & -\lambda_{13,23} & & & & = & -\lambda_{13} \\
 & & & & & & \lambda_{23,21} & & +\lambda_{23,23} & = & \lambda_{23} \\
 & & & & & & -\lambda_{23,21} & +\lambda_{23,13} & & = & 0 \\
 & & & & & & & -\lambda_{23,13} & +\lambda_{23,23} & = & -\lambda_{23} \\
\lambda_{12,12} & & & +\lambda_{13,12} & & & & & & \leq & c_{12} \\
 & \lambda_{12,13} & & & +\lambda_{13,13} & & & & & \leq & c_{21} \\
 & & & & & & & \lambda_{23,13} & & \leq & c_{13} \\
 & & & & & \lambda_{13,23} & & & +\lambda_{23,23} & \leq & c_{23} \\
 & & \lambda_{13,23} & & & & & & & \leq & c_{32}
\end{array} \tag{3.17}$$

and all λ are none negative

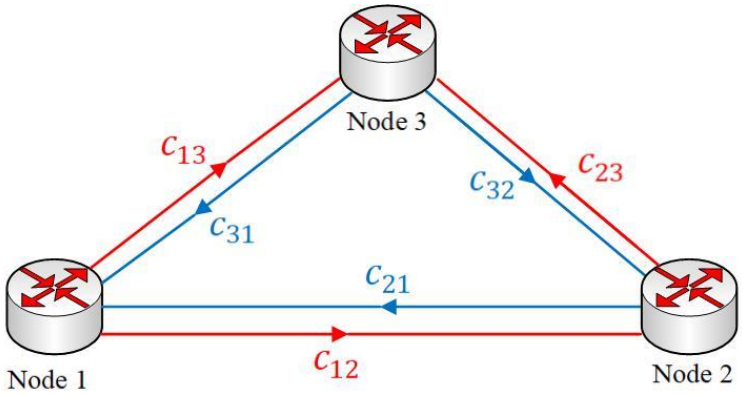


Figure 3.8 Link capacity example in three nodes network

3.4.3 Optimisation Modelling Languages

There exists a large number of programming languages that can be used to model and solve optimisation problems. There are several modelling language providers that have developed modelling and optimisation languages such as *Advanced Interactive Multidimensional Modelling System* (AIMMS) and *Optimisation Programming Language* (OPL). A List of modelling languages and their providers could be found in [215]. In this work, *A Mathematical Programming Language* (AMPL) is exploited to model the optimisation problems for generality of its syntax and the similarity of its statements to the modeller's algebraic notation [216]. The AMPL programme user feeds their codes and data into the software and the software convert the codes into an intermediate file that can be read by a solver. A solver is a mathematical software which reads the intermediate file and applies an appropriate algorithm. There exist many options of solvers such as *GNU Linear Programming Kit* (GLPK) [217], *Coin-or Branch and Cut* (CBC) [218], and the widely used IBM CPLEX [219]. In this thesis CPLEX is used as the MILP optimisation problem solver.

3.5 Summary

This chapter gave a review of virtualisation and Network Function Virtualisation (NFV). The concept of virtualisation and its types were introduced in this chapter. This chapter also has discussed the emergence of NFV and reviewed the work concerned with the deployment of NFV in 5G. A review of the current research efforts undertaken to enhance the energy efficiency in 5G has also been presented. Finally, an overview of the optimisation tools of linear programming and their use in network optimisation was presented in this chapter.

Chapter 4
A Framework for Energy Efficient NFV in 5G Networks

4.1 Introduction

As alluded to earlier, network function virtualisation is an emerging IT technology that transfers the hardware-dedicated function to a software that can be run on virtualised computational environments. In other words, NFV separates the hardware based function from its underlying hardware and converts it to a software-based function that can be run on emulated hardware. This definition implies that the available resources of physical machines can be shared by more than one virtual machine [102, 220].

Several critical benefits can be provided by the concept of NFV when this concept is brought under the roof of 5G [15] such as scalability, high level of flexibility, efficient utilisation of network resources, cost and power reduction [221, 222], and on demand allocation of network resources [31]. A number of functions in the mobile network can be implemented by virtual machines and provided on demand. NFV can be deployed in both mobile core networks and radio access networks (RAN) [220]. Nevertheless, RAN has drawn the interest of the network operators and service providers, as it is the highest energy consuming part of the network [79]. Therefore, by consolidating as many RAN functions as possible in standard hardware using NFV, power consumption in the access network can be reduced.

This chapter introduces a framework for designing an energy efficient architecture for 5G mobile network function virtualisation. In the proposed

architecture, the main functionalities of the mobile core network are virtualised and provisioned on demand. In addition, the function of baseband processing of the mobile base station “eNodeB” is virtualised and offloaded from the mobile radio side. The capabilities of gigabit passive optical networks have been leveraged as the radio access technology to connect the remote radio head to the new virtualised BBU. IP over WDM network is considered as a backbone networks. IP over WDM and PON nodes (ONU and OLT) are considered as the hosts of virtual machine where network function will be implemented. Mainly two scenarios are investigated: virtualisation in PON and virtualisation in PON and IP over WDM network. The main key is that virtualisation contributes to the energy-efficiency through two main ways: (1) function abstraction and virtualisation, and (2) VM packing in a single hardware. In function abstraction and virtualisation, a mobile function that runs on high power consumption and low number of cycles per second processor could be virtualised and run alongside with another virtual mobile function on low power consumption and high number of cycles per second processor. Consolidation or packing VMs is a technique that reduces the number of active physical machines by packing virtual machines into one physical machine. However, consolidation reduces the number of active physical machines which calls for less power consumption. Therefore, by selecting a proper location to host a VM the energy efficiency could be improved.

4.2 NFV in 5G networks

In the literature, a number of studies has investigated NFV deployment in mobile networks; however, only a few focused on saving energy. The authors of [15] and

[223] have shown that the deployment of NFV in 5G will resiliently support the functional demands and the implementation of new and more network services. The authors of [224] have shown that some of the 5G networks' requirements and needs could be met by the integration of a software defined network (SDN) with NFV. To verify this, they have proposed an SDN and NFV integration-based architecture and deployed it on a testbed. Building upon the management point of view, the authors of [225] have argued that the deployment of NFV in 5G towards the edge cloud could bring many merits to both virtual and traditional operators as well as network users.

According to the third generation partnership project (3GPP) the latest mobile core network is the evolved packed core (EPC) [226]. There are four main functions in the EPC [227, 228] illustrated in Figure 4.1: the packet data network gateway (PGW), the serving gateway (SGW), the mobility and management entity (MME), and the policy control and charging role function (PCRF).

On the other hand, the evolved node base station (eNB), which represents the RAN of the current mobile system, consists of two entities: Base Band Unit (BBU) and Remote Radio Unit (RRU), as shown in Figure 4.1.

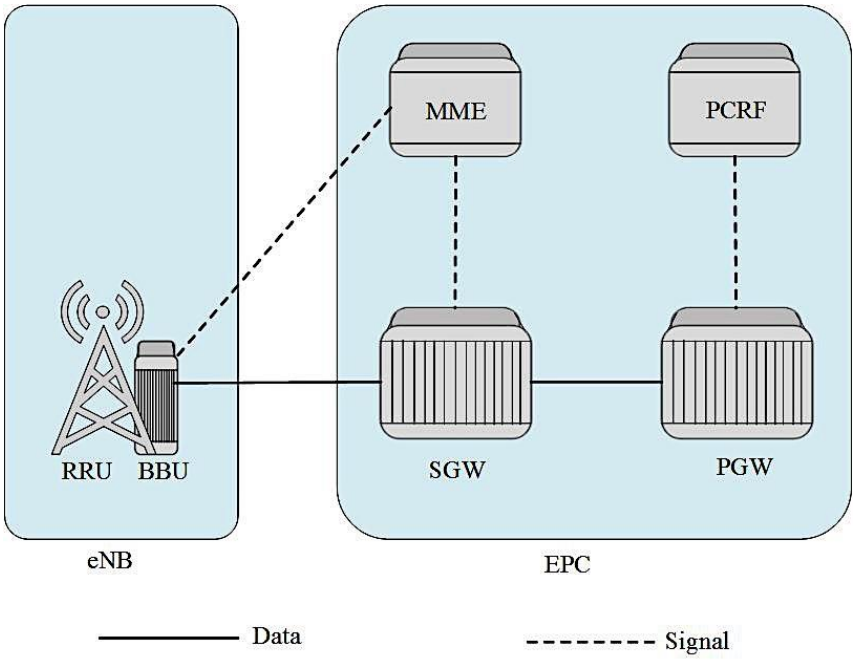


Figure 4.1 Evolved Packet System Architecture

In this chapter, an optical-based framework is introduced for energy efficient NFV deployment in 5G networks. In this framework, the functions of the four entities of mobile core network are virtualised and provided as one virtual machine, which has been dubbed “core network virtual machine” (CNVM). For the radio access side, the BBU and RRU are split up and the function of BBU is virtualised and provisioned as a “BBU virtual machine” (BBUVM). Consequently, the wireless access network of the mobile system will encompass only the RRU units that remain after the RRU-BBU decoupling. RRU has been called “RRH” as in a number of studies after it has been separated from BBU, such as in [29, 229, 230]. The traffic from CNVM to RRH is compelled to pass through BBUVMs for baseband processing, as in Figure 4.2. Moreover, the capabilities of Passive Optical Network (PON) are leveraged as an energy-efficient access network to connect the IP over

WDM network to RRH nodes, and to typify the wired access network of our proposed system. In addition to this, the PON is linked to an IP over WDM network, which acts as the backbone of the proposed architecture. Figure 4.3 shows three locations that can accommodate virtual machines (VMs) of any type (BBUVMs or CNVMs), which are ONU, OLT, and the IP over WDM nodes. For simplicity, the nodes where the hosted servers are accommodated are referred to as “Hosting Nodes”.

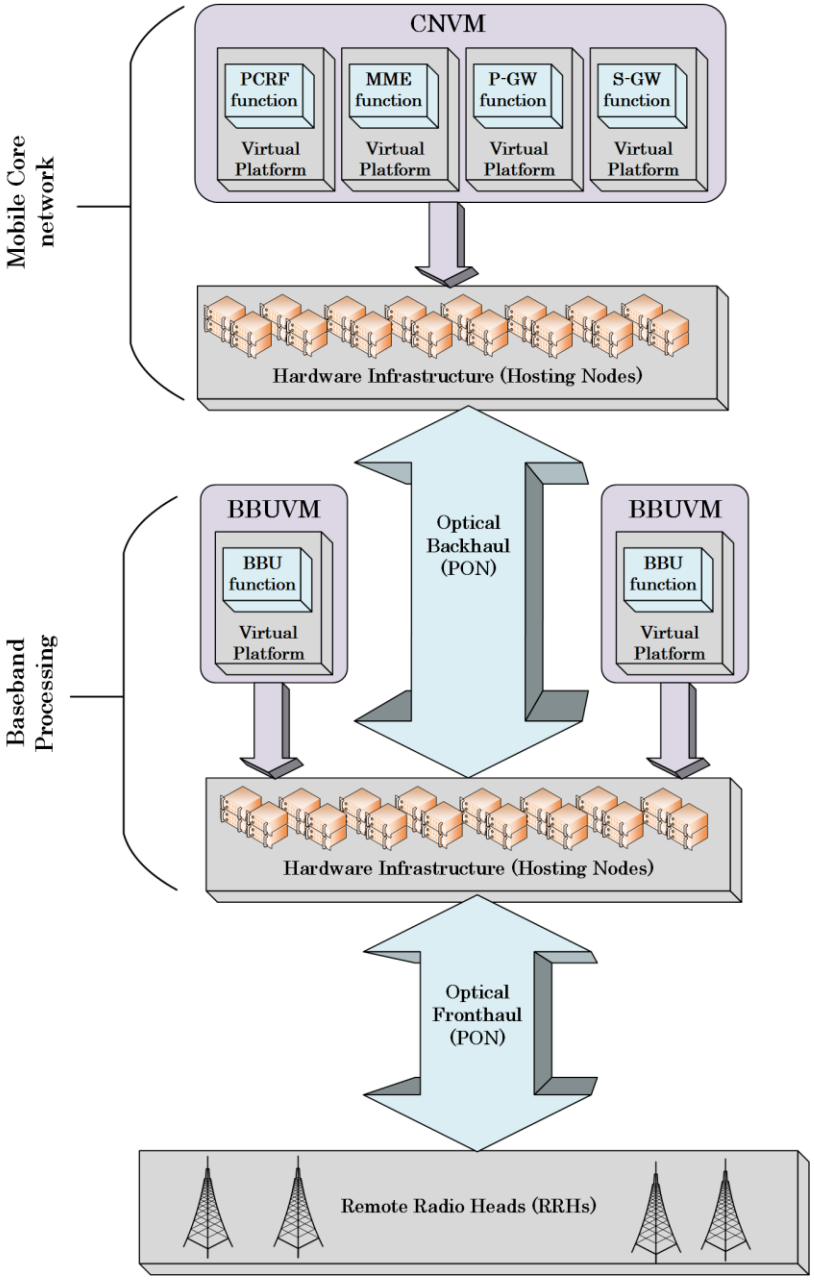


Figure 4.2 The proposed architecture for NFV in 5G

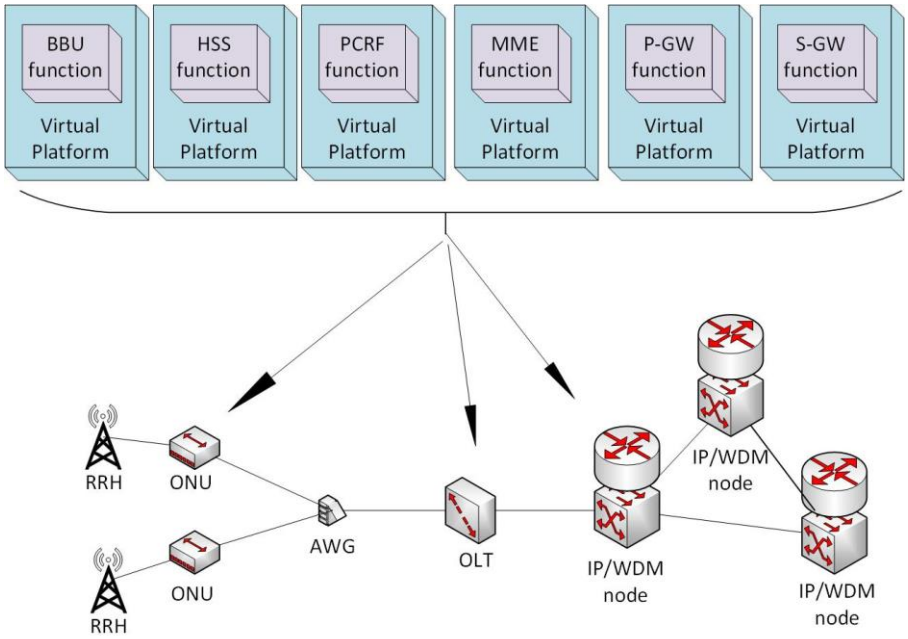


Figure 4.3 The candidate location for hosting virtual machines in the proposed architecture

The hosting nodes (ONU, OLT and IP over WDM nodes) might host one VM or more than one VM of the same or different types, bringing forth the creation of small clouds, or "Cloudlets". Therefore, the proposed architecture will provide an agile allotment of services and processes through a flexible distribution of VMs over the optical network (PON and IP over WDM network), which is one of the main concerns of this work in minimising the total power consumption. Based on this architecture, an MILP formulation has been developed with the overall aim of minimising power consumption.

4.3 MILP model for Energy Efficient NFV in 5G

An MILP model for energy efficient virtualisation in 5G networks has been developed to minimise the total power consumption associated with the RRH requests and optimise the location of the virtual machines. Two different virtual machine types have been considered: mobile core network functions virtual

machines (CNVMs) and BBU functions virtual machines (BBUVMs). The power consumption of a virtual machine is modelled as a function of the normalised workload. In modelling the power consumption of the network, four power-consuming parts have been considered: IP over WDM network, PON, RRH, and the processing clouds power consumption.

The following indices, parameters, and variables are defined Table 4.1, Table 4.2 and Table 4.3 respectively to represent the developed model:

Table 4.1 MILP model indices

Indices	Comments
c	Index of a core network virtual machine CNVM
b	Index of a baseband unit virtual machine BBUVM
r	Index of a remote radio head node RRH
h, u, v	Indices of the nodes that may host any virtual machine.
x, y	Indices of any two nodes in the network
s, d	Indices of source and destination nodes in the IP layer of the IP over WDM network
i, j	Indices of any two nodes in the IP layer of the IP over WDM network
m, n	Indices of any two nodes in the physical fibre link of the optical layer of the IP over WDM network

Table 4.2 MILP model parameters

Parameters	Comments
TN	Set of total nodes
RH	Set of RRH nodes
ONU	Set of ONU nodes
OLT	Set of OLT nodes
N	Set of IP over WDM nodes

NB_x	Set of neighbours of x
NN_m	Set of neighbours of m in the physical layer of the IP over WDM network
$BBUVM$	Set of BBU virtual machines
$CNVM$	Set of mobile core network CN virtual machine
H	Set of hosting nodes (ONU ∪ OLT ∪ N)
LB_{br}	Traffic from the BBUVM b to the RRH node r
LC_{cb}	Traffic from the CNVM c to the BBUVM b
β	Large number
$BBUvmWL_b$	The normalised workload of the BBUVM b
$CNvmWL_c$	The normalised workload of the CNVM c
$OLchP$	OLT chassis power consumption
$OLupP$	OLT uplink card power consumption
$OLponP$	OLT PON card power consumption
$OLswP$	OLT switching card power consumption
$OLTcap$	OLT switching capacity
ONT_MPCh	OLT maximum power consumption calculated as $OLT\_MPCh = (OLchP + 2 \cdot OLupP + 16 \cdot OLponP + 2 \cdot OLswP)$ This calculation is based on the OLT rack configuration [231] [Ref].
$OLTepb$	OLT energy per bit calculated as: $OLTepb = (OLT\_MPCh - OLchP)/OLTcap$
ONU_MPCh	ONU maximum power consumption
$ONTcap$	ONU switching capacity (Gbps)
$ONUepb$	ONU energy per bit calculated as $ONUepb = ONU\_MPCh/ONTcap$
$rruPC$	RRH power consumption
$SmaxPC$	Cloud VM server maximum power consumption
$SMWL$	Cloud VM server maximum normalised workload
$Sepb$	Cloud VM server energy per bit
$CswC$	Cloud LAN switch capacity
CrC	Cloud LAN router capacity
$CswPC$	Cloud LAN switch power consumption

$CrPC$	Cloud LAN router power consumption
Rd	Cloud LAN redundancy factor
$CswEPB$	Cloud LAN switch energy per bit calculated as $CswPC/CswC$
$CrEPB$	Cloud LAN router energy per bit calculated as $CrPC/CrC$
B	Capacity of the wavelength channel (Gb/s)
W	The number of wavelengths per fibre
$RPPC$	IP over WDM router port power consumption
PT	Transponder power consumption
PMD	Power consumption of the MUX and DMUX
PE	EDFA power consumption
S	Maximum span distance between EDFA
D_{mn}	Distance between node pair (m, n) in the IP/WDM network
A_{mn}	Number of EDFA between node pair (m, n) in the IP/WDM network calculated as: $A_{mn} = \left((D_{mn}/S) - 1\right) + 2 \quad \forall m, n \in N$

Table 4.3 MILP model variables

Variables	Comments
LB_{hbr}	Traffic from the BBUVM b hosted in node h to the *RRH* node r
γ_{hb}	Binary indicator, set to 1 if the BBUVM b is hosed in node h
LR_{hr}	Traffic from the BBUVM hosted in node h to the RRH node r
LC_{hcb}	Traffic from the CNVM c hosted in the hosting node h to the BBUVM virtual machine b
γ_{hc}	Binary indicator, set to 1 if the CNVM c is hosted in the node h
$intIND_{ubvc}$	Binary indicator, it is set to 1 if the BBUVM b is hosted in the node u and the CNVM c is hosted in the node v
DUM_{ubvc}	The complement of $intIND_{ubvc}$
$intL_{sd}$	Traffic from s to d in IP/WDM network due to BBUVMs

LR_{xy}^{hr}	Traffic from node h to node r traversing the link (x, y)
$extL_{sd}$	Traffic from s to d in IP/WDM network due to CNVMs
L_{sd}	Total traffic from the node s to the node d in the IP/WDM network
LD_{uv}	Traffic from the hosting node u to the hosting node v due to CNVMs
LD_{xy}^{uv}	Traffic from the hosting node u to the hosting node v due to CNVMs $d$traversing the link (x, y)
γ_{hbr}	Binary indicator, set to 1 if the BBUVM b is hosted in hosting node h to serve the RRH node r
$BBUPC$	Total power consumption of the BBUs
$OLTpc$	Total power consumption of OLTs
$ONUpc$	Total power consumption of ONUs
$CLDWL_c$	Total normalised workload of the cloud c
$CLDPC_c$	Power consumption of the cloud c
C_{ij}	Number of wavelength channels in the virtual link (i, j)
w_{mn}^{ij}	Number of the wavelength channels between node pairs (i, j) that traverse the physical link (m, n)
f_{mn}	Number of fibres on the physical link (m, n)
w_{mn}	Total number of wavelengths in the physical link (m, n)
AGP_x	Number of aggregation ports of the router x
$IPWDM\_PC$	IP/WDM network power consumption
DL_{ij}^{sd}	Traffic from node s to node d in IP network that traverses the virtual link (i, j)

The total power consumption is composed of

1) IP over WDM power consumption is composed of [131, 232]:

a) Aggregation ports power consumption calculated as the total number of aggregation ports multiplied by power consumption of a single port:

$$\sum_{s \in N} (AGP_s \cdot RPPC)$$

b) High speed ports power consumption calculated as the multiplication of total number of high speed ports by the power consumption of a single port

$$\left(RPPC \cdot \sum_{m \in N} \sum_{n \in NN_m} w_{mn} \right)$$

c) Transponders power consumption calculated as the multiplication of total number of wavelength by the power consumption of a single transponder:

$$\left(\sum_{m \in N} \sum_{n \in NN_m} w_{mn} \cdot PT \right)$$

d) EDFAs power consumption calculated as the multiplication of the total number of EDFAs by the total number of fibres by the power consumption of a single EDFA

$$\left(\sum_{m \in N} \sum_{n \in NN_m} PE \cdot f_{mn} \cdot A_{mn} \right)$$

e) MUX/DMUX power consumption calculated as the total number of fibres multiplied by the power consumption of a single MUX/DMUX

$$\left(\sum_{m\in N}\sum_{n\in NN_m} PMD\cdot f_{mn}\right)$$

2) ONUs and RRHs nodes power consumption calculated as the total traffic of passing through ONUs multiplied by the ONU energy per transmitted bit added to the total power consumption of RRHs

$$\sum_{x\in ONU}\left(rruPC + ONUepb \cdot\left(\sum_{h\in H}\sum_{r\in RH}\sum_{y\in NB_x} LR_{xy}^{hr} + \sum_{u\in H}\sum_{v\in H:v\neq u}\sum_{y\in NB_x} LD_{xy}^{uv}\right)\right)$$

3) OLTs power consumption calculated as the total traffic of passing through OLTs multiplied by the OLT energy per transmitted bit

$$\sum_{x\in OLT}\left(OLchP + OLTepb \cdot\left(\sum_{h\ H}\sum_{r\in RH}\sum_{y\in NB_x} LR_{xy}^{hr} + \sum_{u\in H}\sum_{v\in H:v\neq u}\sum_{y\in NB_x} LD_{xy}^{uv}\right)\right)$$

4) Processing cloud power consumption consists of

   a) VM severs power consumption calculated as the total nodes workload multiplied by the energy per processing bit

$$\sum_{h\in H}(CLDWL_h\cdot SmaxPC/SMWL)$$

b) Local networks power consumptions calculated as the total traffic passing through VM servers local network multiplied by the energy per bit of the network router and switches

$$\sum_{h\in H}\left((Rd\cdot CswEPB+CrEPB+Sepb)\cdot\left(\sum_{r\in RH}LR_{hr}+\sum_{v\in N:v\neq h}LD_{hv}\right)\right)$$

The model objective is to minimise the total power consumption as follows:

Minimise

$$\left[\sum_{s\in N}(AGP_s\cdot RPPC)+\left(RPPC\cdot\sum_{m\in N}\sum_{n\in NN_m}w_{mn}\right)+\left(\sum_{m\in N}\sum_{n\in NN_m}w_{mn}\cdot PE\right)\right.$$
$$\left.+\left(\sum_{m\in N}\sum_{n\in NN_m}PE\cdot f_{mn}\cdot A_{mn}\right)+\left(\sum_{m\in N}\sum_{n\in NN_m}PMD\cdot f_{mn}\right)\right]$$
$$+\left[\sum_{x\in ONU}\left(rruPC+ONUepb\right.\right.$$
$$\left.\left.\cdot\left(\sum_{h\in H}\sum_{r\in RH}\sum_{y\in NB_x}LR_{xy}^{hr}+\sum_{u\in H}\sum_{v\in H:v\neq u}\sum_{y\in NB_x}LD_{xy}^{uv}\right)\right)\right]$$

$$+\left[\sum_{h\in H}\left(CLDWL_h\cdot SmaxPC/SMWL+(Rd\cdot CswEPB+CrEPB+Sepb)\right.\right.$$
$$\left.\left.\cdot\left(\sum_{r\in RH}LR_{hr}+\sum_{v\in N:v\neq h}LD_{hv}\right)\right)\right]$$

$$+\left[\sum_{x\in OLT}\left(OLchP+OLTepb\right.\right.$$

$$\left.\left.\cdot\left(\sum_{h\in BBUL}\sum_{r\in RH}\sum_{y\in NB_x}LR_{xy}^{hr}+\sum_{u\in H}\sum_{v\in H:v\neq u}\sum_{y\in NB_x}LD_{xy}^{uv}\right)\right)\right]$$

Subject to the following constraints:

1) Traffic from the virtual machine b in all the hosting node should meet the demand of the RRH node r

$$\sum_{h\in H}LB_{hbr}=LB_{br}$$
$$\forall b\in BBUVM,\forall r\in RH \tag{4.1}$$

2) BBUVM location

$$\beta\cdot\sum_{r\in RH}LB_{hbr}\geq\gamma_{hb}$$
$$\forall b\in BBUVM,\forall h\in H \tag{4.2}$$

$$\sum_{r\in RH}LB_{hbr}\leq\beta\cdot\gamma_{hb}$$
$$\forall b\in BBUVM,\forall h\in H \tag{4.3}$$

Constraint (4.1) represents the traffic from BBUVM b in all hosting nodes to the RRH node r. It also enables the distribution of BBUVM b over more than one hosting node (VM slicing). Constraints (4.2) and (4.3) determine the location of BBUVM b.

3) Traffic from the hosting node h to the RRH node r due to the hosting the BBUVM in node h

$$\sum_{b \in BBUVM} LB_{hbr} = LR_{hr} \quad \forall r \in RH, \forall h \in H \tag{4.4}$$

4) Flow conservation due to serving BBUVM

$$\sum_{y \in NB_x} LR_{xy}^{hr} - \sum_{y \in NB_x} LR_{yx}^{hr} = \begin{cases} LR_{hr} \; if \; x = h \\ -LR_{hr} \; if \; x = r \\ 0 \; otherwise \end{cases} \quad \forall h \in H, \forall r \in RH, \forall x \in TN \tag{4.5}$$

Constraint (4.4) represents the total traffic from the hosting node *h* toward RRH node *r*, whilst its flow conservation is represented by constraint (4.5).

5) Traffic from CNVM in all the nodes should satisfy BBUVM demand

$$\sum_{h \in H} LC_{hcb} = LC_{cb} \quad \forall c \in CNVM, \forall b \in BBUVM \tag{4.6}$$

6) CNVM location

$$\beta \cdot \sum_{b \in BBUVM} LC_{hcb} \geq \gamma_{hc} \quad \forall h \in H, \forall c \in CNVM \tag{4.7}$$

$$\sum_{b \in BBUVM} LC_{hcb} \leq \beta \cdot \gamma_{hc} \quad \forall h \in H, \forall c \in CNVM \tag{4.8}$$

Constraint (4.6) represents the traffic from CNVM *c* in all hosting nodes to the BBUVM *b*. It also slices CNVM *c* over a number of hosting nodes. Constraints (4.7) and (4.8) determine the location of CNVM.

7) Determine whether BBUVM and CNVM are at the same node or not

$$\gamma_{ub} + \gamma_{vc} = 2 \cdot intIND_{ubvc} + DUM_{ubvc}$$
$$\forall u, v \in H: u \neq v, \forall b \in BBUVM, \forall c \in CNVM \tag{4.9}$$

8) Traffic from node u to node v due to host CNVM in u and BBUVM in v

$$\sum_{b \in BBUVM} \sum_{c \in CNVM} intIND_{ubvc} \cdot LC_{cb} = LD_{uv}$$
$$\forall u, v \in H, : u \neq v \tag{4.10}$$

9) Flow conservation due to serving CNVM

$$\sum_{y \in NB_x} LD_{xy}^{uv} - \sum_{y \in NB_x} LD_{yx}^{uv} = \begin{cases} LD_{uv} \; if \; x = u \\ -LD_{uv} \; if \; x = v \\ 0 \;\; otherwise \end{cases}$$
$$\forall u, v \in H, \forall x \in TN, u \neq v \tag{4.11}$$

Constraint (4.9) determines whether BBUVM b and CNVM c are hosted at the same place or not. It is equivalent to the logical ANDing of the two binary variables γ_{ub} and γ_{vc}; i.e $intIND_{ubvc} = \gamma_{ub} \; AND \; \gamma_{vc}$ whilst DUM_{ubvc} is a dummy binary variable. Constraints (4.10) represents the total traffic between two hosting node u and v due to CNVMs and BBUVMs communication whilst its flow conservation is represented in constraint (4.11).

10) Traffic in IP over WDM network due to BBUVM

$$\sum_{h \in H} \sum_{r \in RH} LR_{xy}^{hr} = extL_{xy}$$
$$\forall x \in N, \forall y \in NB_x \cap N \tag{4.12}$$

11) Traffic in IP over WDM network due to CNVM

$$\sum_{u \in H} \sum_{v \in H: v \neq u} LD_{xy}^{uv} = intL_{xy}$$
$$\forall x \in N, \forall y \in NB_x \cap N \tag{4.13}$$

12) Total traffic in IP over WDM network

$$L_{sd} = extL_{sd} + intL_{sd}$$
$$\forall s, d \in N: s \neq d \tag{4.14}$$

Constraint (4.12) represents the traffic from BBUVMs to RRH nodes that flows in the IP over WDM networks. Constraint (4.13) represents the traffic from CNVMs to BBUVMs that flows in the IP over WDM network whilst the total traffic flows in the IP over WDM network is represented by (4.14).

13) Cloud location

$$\sum_{c \in CNVM} \gamma_{hc} + \sum_{b \in BBUVM} \gamma_{hb} \geq CLD_h \quad \forall h \in H \tag{4.15}$$

$$\sum_{c \in CNVM} \gamma_{hc} + \sum_{b \in BBUVM} \gamma_{hb} \leq \beta \cdot CLD_h \quad \forall h \in H \tag{4.16}$$

14) Cloud total normalised workload

$$CLDWL_h = \sum_{c \in CNVM} \gamma_{hc} \cdot CNVM\_WL_c + \sum_{b \in BBUVM} \gamma_{hb} \cdot BBUVM\_WL_b \quad \forall h \in H \tag{4.17}$$

Constraints (4.15) and (4.16) determine the location of the processing cloud formed by VMs whilst constraint (4.17) represents its total normalised workload.

15) Flow conservation in IP layer of IP/WDM network

$$\sum_{j \in N: i \neq j} L_{ij}^{sd} - \sum_{j \in N: i \neq j} L_{ji}^{sd} = \begin{cases} L_{sd} \; if \; x = s \\ -L_{sd} \; if \; x = d \\ 0 \; otherwise \end{cases} \quad \forall s, d \in N: s \neq d \tag{4.18}$$

16) Number of wavelength channels

$$\sum_{s \in N} \sum_{d \in N: s \neq d} L_{ij}^{sd} \leq C_{ij} \cdot B \quad \forall i, j \in N: i \neq j \tag{4.19}$$

17) Flow conservation in the physical layer

$$\sum_{n \in NN_m} w_{mn}^{ij} - \sum_{n \in NN_m} w_{nm}^{ij} = \begin{cases} C_{ij} \; if \; m = i \\ -C_{ij} \; if \; m = j \\ 0 \; otherwise \end{cases} \quad \forall i, j, m \in N : i \neq j \tag{4.20}$$

Constraint (4.18) represents the flow conservation in the IP over WDM network. Constraint (4.19) determines the number of wavelength channels. Constraint (4.20) represents the flow conservation in the physical (optical) layer of the IP over WDM network.

18) Number of fibres

$$\sum_{i \in N} \sum_{j \in N : i \neq j} w_{mn}^{ij} \leq W \cdot f_{mn} \quad \forall m \in N, \forall n \in NN_m \tag{4.21}$$

19) Total number of wavelengths

$$\sum_{i \in N} \sum_{j \in N : i \neq j} w_{mn}^{ij} = w_{mn} \quad \forall m \in N, \forall n \in NN_m \tag{4.22}$$

20) Number of aggregation ports

$$AGP_s = \sum_{d \in N : d \neq s} L_{sd} / B \quad \forall s \in N \tag{4.23}$$

Constraints (4.21) and (4.22) determine the number of fibres and wavelengths respectively between any two nodes in the WDM network, whilst constraint (4.23) determines the total number of aggregation ports in each IP over WDM router.

4.4 MILP model setup and results

The MILP model considers the network topology illustrated in Figure 4.4. In this topology, two groups of 15 ONUs are considered (30 in total) where each ONU is connected to one RRH whilst each group of the ONUs is connected to one OLT. The two OLTs are connected to only one of 5 IP over WDM nodes (typically a connection is established to one of the nearest core nodes, here 5 core nodes were considered).

This architecture represents the virtualisation infrastructure of the MILP model. The situation where each RRH requests one BBUVM, and each BBUVM in turn requests one CNVM was considered. The developed MILP model considers 14 BBUVMs uniformly and randomly distributed (we also considered a larger and smaller number of BBUVMs and observed similar trends, but report here the case of 14 BBUVMs) over the 30 RRH nodes. The same distribution method was used with the distribution of 14 CNVMs over the BBUVMs. The traffic requested by each RRH node is randomly and uniformly generated with a maximum of 2 Gbps. To capture a range of processing scenarios and account for factors such as variation in the BBU processor type or number of cycles per instruction, five scenarios were examined with different ranges of normalised workload (NWL) of BBUVMs and CNVMs as shown in Table 4.4 with the parameters listed in Table 4.5.

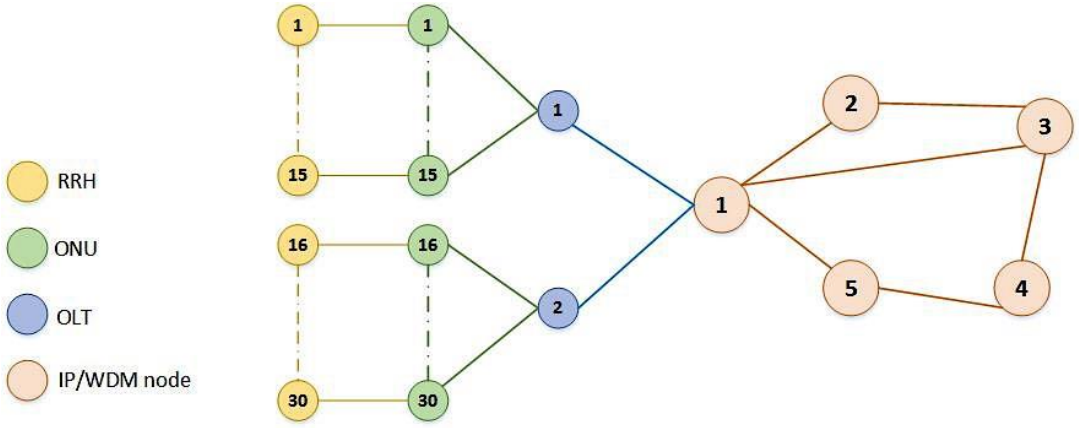


Figure 4.4 Network topology considered in the developed MILP model

Table 4.4 Virtual machines workloads (%) of different scenarios

Scenario	BBUVM NWL (%)	CNVM NWL (%)
Scenario 1	0.25 – 2.5%	1 – 10%
Scenario 2	0.5 – 5%	2 – 20%
Scenario 3	0.75 – 7.5%	3 – 30%
Scenario 4	1 – 10%	4 – 40%
Scenario 5	1.25 – 12.5%	5 – 50%

Table 4.5 MILP model input parameters

ONU maximum power consumption (ONU_MPCh)	15 (W) [233]
OLT maximum power consumption (ONT_MPCh)	1940 (W) [231]
OLT idle power ($OLchP$)	60 (W) [231]
OLT maximum capacity ($OLTcap$)	8600 (Gbps) [231]
ONU maximum capacity ($ONTcap$)	10 (Gbps) [233]
RRH node power consumption ($rruPC$)	1140 (W) [234]
Cloud VM server maximum power consumption ($SmaxPC$)	300 (W) [235]
Cloud VM server maximum normalised workload ($SMWL$)	100%
Cloud VM server energy per bit ($Sepb$)	211.1 [235]
Cloud LAN switch capacity ($CswC$)	320 Gbps [235]

Cloud LAN router capacity (CrC)	66 Gbps [235]
Cloud LAN switch power consumption ($CswPC$)	3800 (W) [235]
Cloud LAN router power consumption ($CrPC$)	5100 (W) [235]
Cloud LAN redundancy factor (Rd)	2
Capacity IP over WDM wavelength channel (B)	40 (Gbps) [236]
Number of wavelength per fibre in IP over WDM (W)	16 [232]
Transponder power consumption (PT)	75 (W) [237]
Router port power consumption ($RPPC$)	1000 (W) [232]
Power consumption of the MUX and DMUX (PMD)	16 (W) [232]
EDFA power consumption (PE)	8 (W) [232]
Maximum span distance between EDFAs (S)	80 (km) [236]

Figure 4.5 illustrates the total power consumption of the five scenarios. It is clearly seen that the power consumption increases as the virtual machine workloads increase from one scenario to another. The difference in power consumption is small between any two successive scenarios for two reasons: same number of VMs in each scenario and overlapped CNVMs workloads.

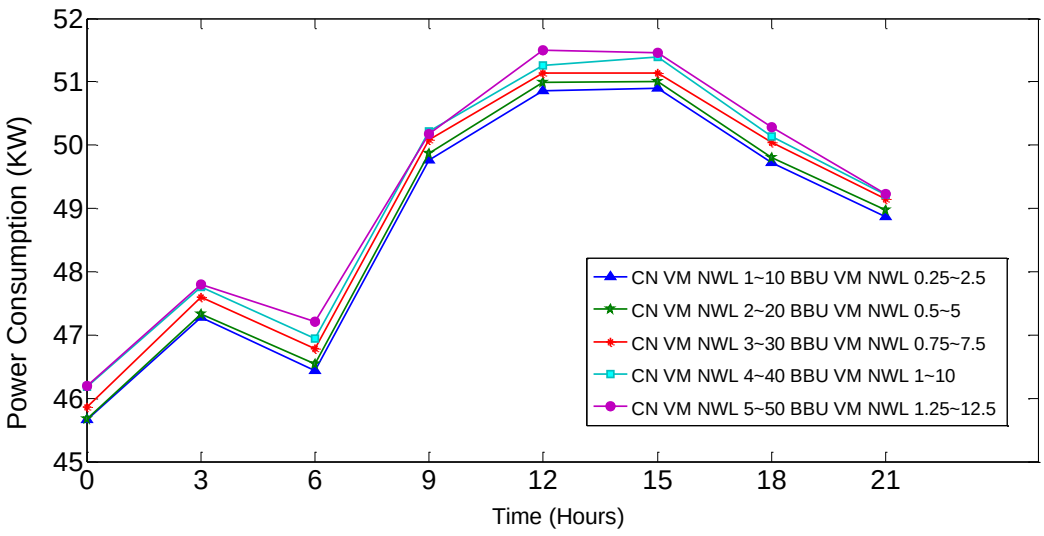


Figure 4.5 Total power consumption under different VM workloads at different times of the day

The case where the virtualisation is restricted in only IP over WDM network was also considered for the above five scenarios but here only the scenario with the highest power consumption is reported as the others have the same behaviour. To restrict the virtualisation in IP over WDM network only, the following two constraints are added:

$$\sum_{b \in BBUVM} \gamma_{hb} \leq 0 \qquad \forall h \in OLT \cup ONU \tag{4.24}$$

$$\sum_{b \in CNVM} \gamma_{hc} \leq 0 \qquad \forall h \in OLT \cup ONU \tag{4.25}$$

Figure 4.6 compares the power consumption of the virtualisation only in IP over WDM network with the power consumption of the virtualisation in IP over WDM and PON.

Virtualisation in the IP over WDM and GPON networks approach shows an average saving in power consumption of 22% compared to the approach where virtualisation is restricted in the IP over WDM network. Virtualisation in the PON network adds a high level of flexibility due to virtual machine behaviour. The virtual machines can be migrated, replicated, or distributed according to the demand and the satisfaction of the optimisation goal, which is the minimisation of total power consumption. For a particular RRH, the demand does not need to travel along the network to be served or processed if the virtual machine is placed close to it in ONU or in OLT. As a result, the power consumption decreases due to the shorter route taken by the demand. In addition, if the BBUVM and its serving CNVM are placed

in the same node, the power consumption due to the internal traffic between them will be zero as they are processed at the same node.

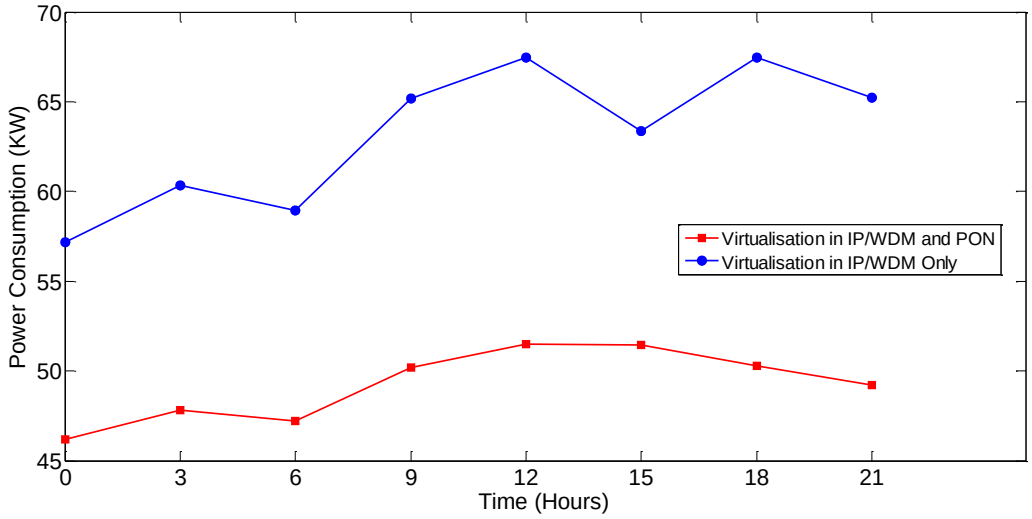


Figure 4.6 Power consumption comparison of virtualisation in IP over WDM only and IP over WDM with PON

Virtualisation in PON extends the range of candidate locations to host VMs. This calls for a shorter traffic path and lower traffic induced power compared with virtualisation in IP over WDM network only. Figure 4.7 illustrates an example of VMs distribution in PON network where two BBUVMs are hosted in ONU nodes close to the users and each BBUVM is served by one CNVM. The total power consumption, in this example, equals to the sum of RRH (r_1 and r_2), ONUs (h_1 and h_2), and VMs processing power consumptions. Another case of virtualisation in PON is shown in Figure 4.8. In this example two BBUVMs are hosted at the same node that is node h_3 (OLT). The power consumption of this case equals to sum of RRH (r_1 and r_2), ONUs (h_1 and h_2), VMs processing, and OLT (h_3) power consumptions. Compared to the previous case, this case leverages the bin packing technique to accommodate VMs in the same node as much as the VM server can

host. Although bin packing techniques reduces the processing power consumption "VM servers"; it increases the traffic induced power. To back VMs in one node, a proper location is needed that is close to all targeted RRH nodes; so that VMs can serve them efficiently. Figure 4.9 illustrate another case of virtualisation which is virtualisation in IP over WDM network only. In this case all BBUVMs and CNVM are hosted in the closest IP over WDM node to the RRH nodes. The total power consumption is calculated as the sum of the power consumption of RRH (r_1 and r_2), ONUs (h_1 and h_2), OLT (h_3), VMs processing, and the part of IP over WDM network that deliver the traffic to RRH nodes. This case has the most traffic induced power compared to the previous two cases illustrated in Figure 4.7 and Figure 4.8. However, by selecting the proper close location to RRH nodes to pack VMs; the power consumption could be optimised.

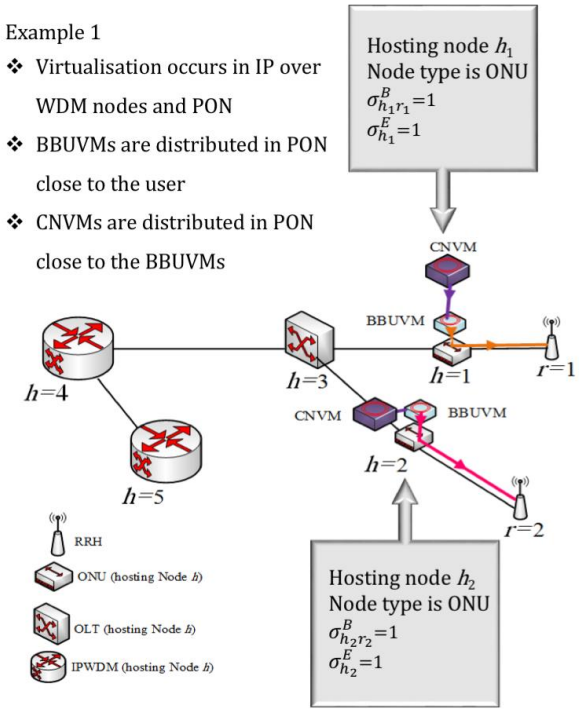


Figure 4.7 Example of VMs distribution in PON

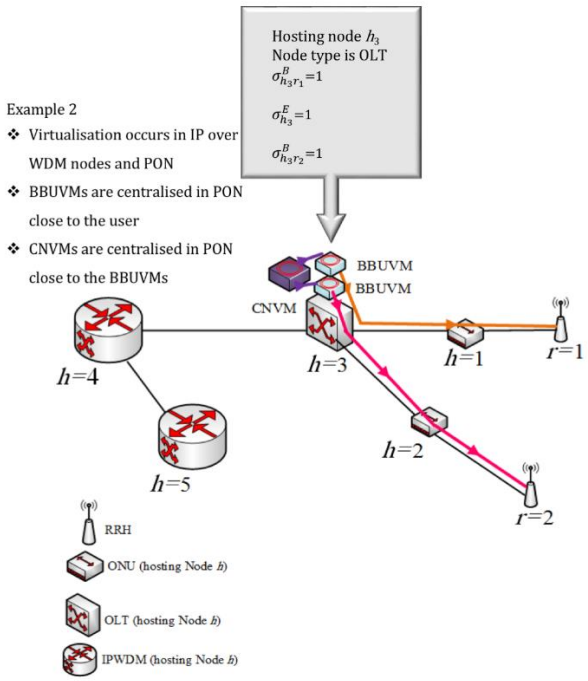


Figure 4.8 Example of VMs packing in PON

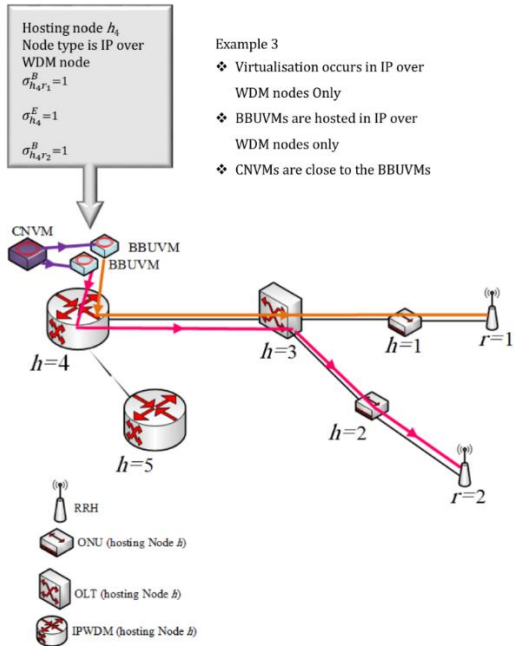


Figure 4.9 Example of VMs packing in IP over WDM nodes

4.5 MILP model for Energy-Efficient NFV with the impact of BBUVM processing and CNVMs communication

This section presents an extended version of the MILP model introduced in Section 4.3 to include a large range of virtual machine workloads in the presence of mobile core network virtual machines (CNVMs) communication. In addition, a wide range of traffic expansion /reduction factors that are caused by BBUVM processing have been considered and their impact on the power consumption has been investigated. Two approaches were considered: virtualisation in both IP over WDM and PON and virtualisation in the IP over WDM network only. The following indices, parameters, and variables are defined to represent the developed model:

Table 4.6 MILP model indices

Indices	Comment
x, y	Indices of any two nodes in the proposed model
m, n	Indices of any two nodes in the physical layer of the IP over WDM network
i, j	Indices of any two nodes in the IP layer of the IP over WDM network.
r	Index of RRH node
h, u, p, q	Indices of the nodes where the VM could be hosted

Table 4.7 MILP model parameters

Parameters	Comment
R	Set of RRH nodes
U	Set of ONU nodes
L	Set of OLT nodes
N	Set of IP over WDM nodes
T	Set of all nodes (RRH, ONU, OLT, and IP over WDM nodes)

NN_m	Set of neighbours of node m in the IP over WDM network, $\forall\ m\ \in N$
TN_x	Set of neighbours of node x, $\forall\ x\ \in T$
H	Set of hosting nodes (ONU, OLT, and IP over WDM nodes)
λR_r	RRH node r traffic demand (Gbps)
$\nabla_{p,q}$	Intra-traffic between core network VMs (CNVM) at hosting nodes p, and q (Gbps)
$1/(1-\alpha)$	is the traffic expansion due to overheads and possibly digitisation of BBUVM to RRH traffic (unitless)
ΩU	ONU maximum power consumption (W)
ΩL	OLT maximum power consumption (W)
ΩLd	OLT idle power (W)
CL	OLT maximum capacity (Gbps)
CU	ONU maximum capacity (Gbps)
ΩR_x	Power consumption of the Remote Radio Head (RRH) connected to ONU node *x* (W)
ΩP_U	Power consumption of hosting VMs at ONU node
ΩP_L	Power consumption of hosting VMs at OLT node
ΩP_N	Power consumption of hosting VMs at IP over WDM node
ΨM_h	Maximum workload at hosting node h
β	Large number (unitless)
η	Very small number (unitless)
B	Capacity of the wavelength channel (Gbps)
w	Number of wavelengths per fiber
ΩT	Transponder power consumption (W)
ΩRP	Router power consumption per port (W)
ΩG	Regenerator power consumption (W)
ΩE	EDFA power consumption (W)
$NG_{m,n}$	Number of regenerators in the optical link (m,n)
S	Maximum span distance between EDFAs (km)
$D_{m,n}$	Distance between node pair (m,n) in the IP over WDM network (km)
$A_{m,n}$	Number of EDFAs between node pair (m,n) calculated as $A_{m,n} = \left((D_{mn}/S) - 1\right) + 2$

Table 4.8 MILP model variables

Variables	Comment
$\lambda B_{p,h}$	Traffic from CNVMs in node p to the BBUVMs in node h (Gbps)
$\lambda R_{h,r}$	Traffic from BBUVMs in node h to the RRH node r (Gbps)
$\sigma B_{h,r}$	Binary indicator, set to 1 if the node h hosts BBUVMs to serve the RRH node r, 0 otherwise
σB_{h}	Binary indicator, set to 1 if the node h hosts a BBUVM, 0 otherwise
$\sigma E_{p,h}$	Binary indicator, set to 1 if the node h hosts CNVMs to serve the BBUVMs at hosting node h, 0 otherwise
σE_{p}	Binary indicator, set to 1 if the hosting node p hosts CNVMs is, 0 otherwise
$\psi_{p,q}$	Binary indicator, set to 1 if two different hosting nodes p and q host CNVMs, 0 otherwise. It is equivalent to the ANDing of the two binary variables (σE_p, σE_q).
$\lambda E_{p,q}$	Traffic between hosting nodes due to CNVMs communication (Gbps)
$\lambda T_{p,q}$	Total traffic from node p to node q caused by CNVM to CNVM traffic and CNVM to BBUVM traffic (Gbps)
$\lambda R_{x,y}^{h,r}$	Traffic from hosting node h to RRH node r that traverses the link between the nodes (x, y) in the network in Gb/s
$\lambda T_{x,y}^{p,q}$	Total traffic from node p to node q that traverses the link between the nodes (x, y) in the network (Gbps)
ΨC_h	Total workload at node h
$W_{i,j}$	Number of wavelength channels in the virtual link (i, j)
$W_{m,n}^{i,j}$	Number of wavelength channels in the virtual link (i, j) that traverse the physical link (m, n)
$f_{m,n}$	Number of fibres in the physical link (m, n)
$W_{m,n}$	Total number of wavelengths in the physical link (m, n)
Λ_m	Number of aggregation ports of the router at node m

The total power consumption is composed of

1) The power consumption of RRHs and ONUs

$$\sum_{x\in U}\left[\Omega R_x + \frac{\Omega U}{CU}.\left(\sum_{h\in H}\sum_{r\in R}\sum_{y\in TN_x}\lambda R_{x,y}^{h,r} + \sum_{p\in H}\sum_{q\in H:p\neq q}\sum_{y\in TN_x\cap H}\lambda T_{x,y}^{p,q}\right)\right]$$

2) The power consumption of the OLTs

$$\sum_{x\in L}\left[\Omega Ld + \frac{\Omega L - \Omega Ld}{CL}.\left(\sum_{h\in H}\sum_{r\in R}\sum_{y\in TN_x}\lambda R_{x,y}^{h,r} + \sum_{p\in H}\sum_{q\in H:p\neq q}\sum_{y\in TN_x\cap H}\lambda T_{x,y}^{p,q}\right)\right]$$

3) The power consumption of IP over WDM network

$$\left(\Omega RP\cdot\sum_{m\in N}\Lambda_m\right) + \left(\Omega RP\cdot\sum_{m\in N}\sum_{n\in NN_m}W_{m,n}\right)$$

$$+\left(\Omega T\cdot\sum_{m\in N}\sum_{n\in NN_m}W_{m,n}\right) + \left(\Omega E\cdot\sum_{m\in N}\sum_{n\in NN_m}A_{m,n}\cdot f_{m,n}\right)$$

$$+\left(\Omega G\cdot\sum_{m\in N}\sum_{n\in NN_m}NG_{m,n}\cdot W_{m,n}\right)$$

4) The total power consumption of VMs.

$$\sum_{h\in U}(\Omega P_U\cdot\Psi C_h/\Psi M_h) + \sum_{h\in L}(\Omega P_L\cdot\Psi C_h/\Psi M_h) + \sum_{h\in N}(\Omega P_N\cdot\Psi C_h/\Psi M_h)$$

The model objective is to minimise the total power consumption as follows:

Minimise

$$\sum_{x\in U}\left[\Omega R_x+\frac{\Omega U}{CU}.\left(\sum_{h\in H}\sum_{r\in R}\sum_{y\in TN_x}\lambda R_{x,y}^{h,r}+\sum_{p\in H}\sum_{q\in H:p\neq q}\sum_{y\in TN_x\cap H}\lambda T_{x,y}^{p,q}\right)\right]$$

$$+\sum_{x\in L}\left[\Omega Ld+\frac{\Omega L-\Omega Ld}{CL}.\left(\sum_{h\in H}\sum_{r\in R}\sum_{y\in TN_x}\lambda R_{x,y}^{h,r}\right.\right.$$

$$\left.\left.+\sum_{p\in H}\sum_{q\in H:p\neq q}\sum_{y\in TN_x\cap H}\lambda T_{x,y}^{p,q}\right)\right]+\left(\Omega RP\cdot\sum_{m\in N}\Lambda_m\right)$$

$$+\left(\Omega RP\cdot\sum_{m\in N}\sum_{n\in NN_m}W_{m,n}\right)$$

$$+\left(\Omega T\cdot\sum_{m\in N}\sum_{n\in NN_m}W_{m,n}\right)+\left(\Omega E\cdot\sum_{m\in N}\sum_{n\in NN_m}A_{m,n}\cdot f_{m,n}\right)$$

$$+\left(\Omega G\cdot\sum_{m\in N}\sum_{n\in NN_m}NG_{m,n}\cdot W_{m,n}\right)$$

$$+\left(\sum_{h\in U}(\Omega P_U\cdot\Psi C_h/\Psi M_h)+\sum_{h\in L}(\Omega P_L\cdot\Psi C_h/\Psi M_h)\right.$$

$$\left.+\sum_{h\in N}(\Omega P_N\cdot\Psi C_h/\Psi M_h)\right)$$

Subject to:

1) Traffic to RRH nodes

$$\sum_{h\in H}\lambda R_{h,r}=\lambda R_r \qquad (4.26)$$

$$\forall r\in R.$$

2) Location of BBUVMs

$$\beta \cdot \lambda R_{h,r} \geq \sigma B_{h,r} \quad \forall r \in R, \forall h \in H \tag{4.27}$$

$$\lambda R_{h,r} \leq \beta \cdot \sigma B_{h,r} \quad \forall r \in R, \forall h \in H \tag{4.28}$$

$$\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r} \geq \sigma B_h \quad \forall h \in H \tag{4.29}$$

$$\sum_{\forall r \in R} \lambda R_{h,r} \leq \beta \cdot \sigma B_h \quad \forall h \in H \tag{4.30}$$

Constraint (4.26) represents the traffic from BBUVMs in all hosting nodes to the RRH node *r*. Constraints (4.27) and (4.28) ensure that the RRH node r is served by the BBUVM that is hosted at the node h as illustrated in Figure 4.10. Constraints (4.29) and (4.30) determine the location of BBUVM; β is a large enough number to ensure that σB_{hr} and σB_h are equal to 1 when $\sum_{\forall r \in R} \lambda R_{hr} > 0$. In constraint (4.29) there are two possibilities for the value of $(\sum_{\forall r \in R} \lambda R_{h,r})$ which are either zero (no traffic from *h* to *r*) or greater than zero (there is a traffic from *h* to *r*). When the value of $\sum_{\forall r \in R} \lambda R_{h,r}$ is zero, the left-hand side of the inequality $(\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r})$

should be zero and this sets the value of σB_h to zero. In the second case when the value of $\sum_{\forall r \in R} \lambda R_{h,r}$ is greater than zero, the left-hand side of the inequality ($\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r}$) will be much greater than 1 because of the large value β. Here the value of σB_h may be set to 1 or zero. In the same way constraint (4.30) sets the value of σB_h. Table 4.9 illustrates the operation of constraints (4.29) and (4.30).

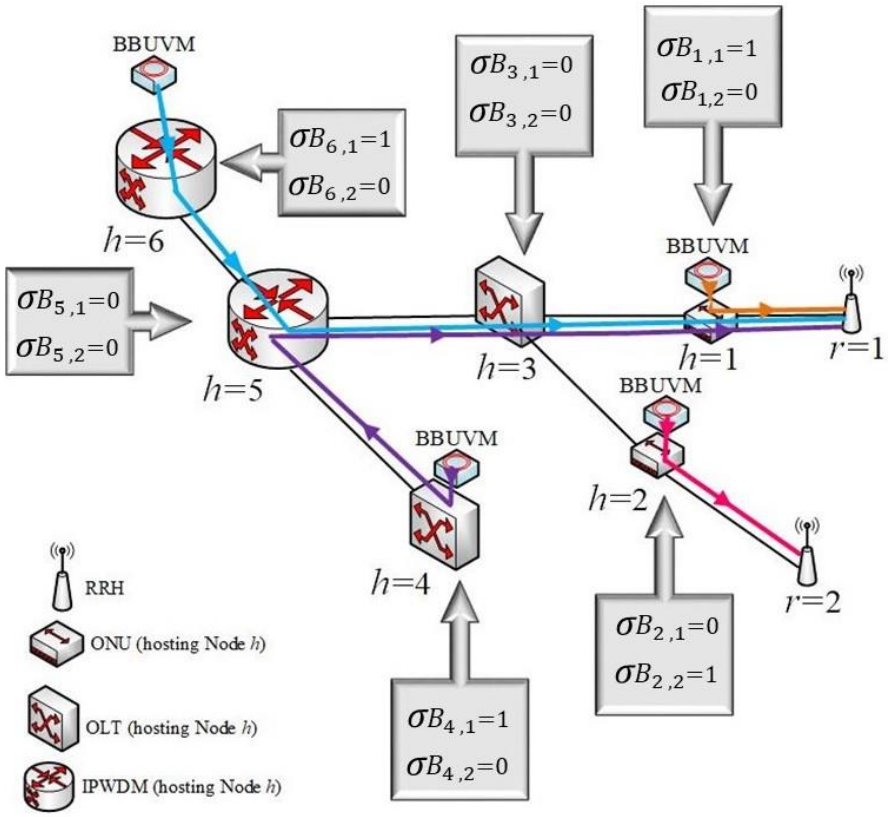


Figure 4.10 Locations of BBUVMs and the traffic toward RRH nodes

Table 4.9 BBUVM location constraints illustration

Input	Constraint	Outcome	σB_h	Value of σB_h that satisfies both constraints
$\sum_{\forall r \in R} \lambda R_{h,r} > 0$	$\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r} \geq \sigma B_h$	$\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r} \gg 1$	0 or 1	1
	$\sum_{\forall r \in R} \lambda R_{h,r} \leq \beta \cdot \sigma B_h$	$\beta \cdot \sigma B_h \gg 1$	1	
$\sum_{\forall r \in R} \lambda R_{h,r} = 0$	$\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r} \geq \sigma B_h$	$\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r} = 0$	0	0
	$\sum_{\forall r \in R} \lambda R_{h,r} \leq \beta \cdot \sigma B_h$	$\beta \cdot \sigma B_h = 0$	0 or 1	

3) Traffic from CNVMs to BBUVM

$$(1-\alpha) \cdot \sum_{r \in R} \lambda R_{h,r} = \sum_{p \in H} \lambda B_{p,h} \quad \forall h \in H \tag{4.31}$$

4) Core network virtual machine location

$$\sigma E_p \geq \eta \cdot \sum_{h \in H} \lambda B_{p,h} \quad \forall p \in H \tag{4.32}$$

$$\sigma E_p \leq 1 + \sum_{h \in H} \lambda B_{ph} - \eta \quad \forall p \in H \tag{4.33}$$

5) CNVMs common location

$$\psi_{pq} \leq \sigma E_p \quad \forall p, q \in H, p \neq q \tag{4.34}$$

$$\psi_{p,q} \leq \sigma E_q \quad \forall p, q \in H, p \neq q \tag{4.35}$$

$$\psi_{p,q} \geq \sigma E_p + \sigma E_q - 1 \quad \forall p, q \in H, p \neq q \tag{4.36}$$

6) Inter-traffic between CNVMs

$$\lambda E_{p,q} = \nabla_{p,q} \cdot \psi_{p,q} \quad \forall p, q \in H: p \neq q \tag{4.37}$$

Constraint (4.31) represents the traffic from CNVMs to the BBUVM in node h where α is a unitless quantity; less than 1 and represents the reduction in the traffic caused by BBUVM processing. Constraints (4.32) and (4.33) determine the location of the CNVMs by setting the binary variable σE_p to 1 if there is a CNVM hosted at node p, where η is very small number. Figure 4.11 illustrates the functions of constraints (4.32) and (4.33) whilst Table 4.10 illustrates their operation. Constraints (4.34), (4.35) and (4.36) ensure that the CNVMs communicate with each other if they are hosted at different nodes p and q, and this is equivalent to the logical operation $\psi_{p,q} = \sigma E_p \; AND \; \sigma E_q$. Figure 4.12 illustrates the function of constraints

(4.34), (4.35) and (4.36). Constraint (4.37) represents the traffic between CNVMs at hosting nodes p and q.

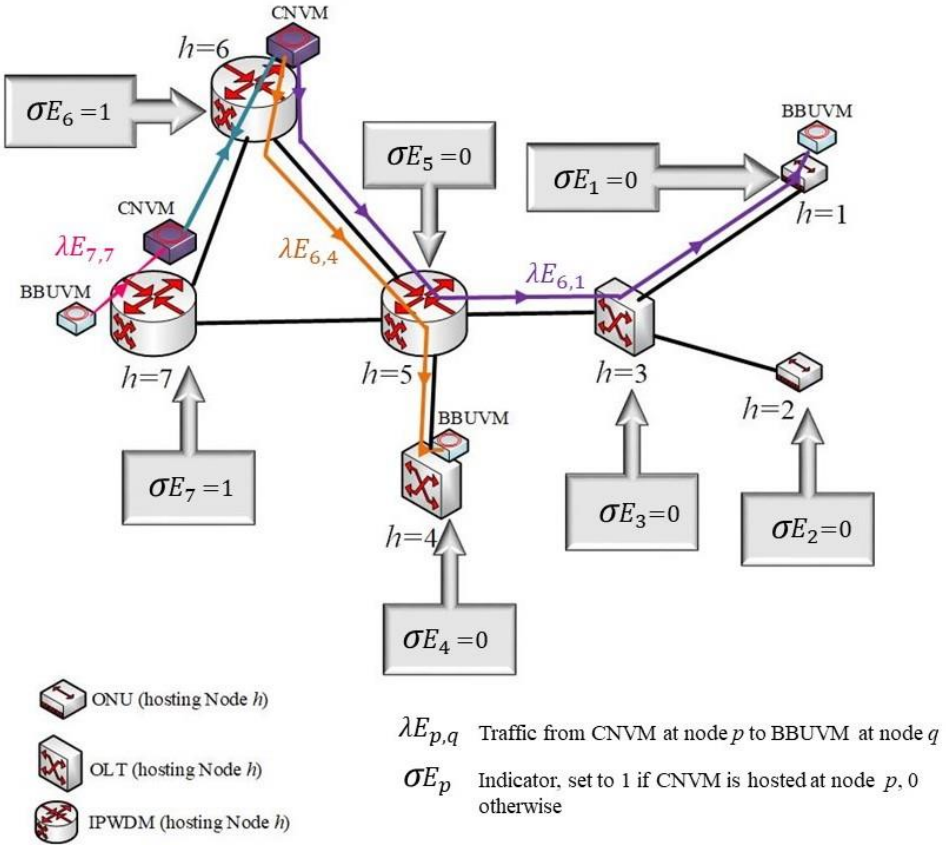


Figure 4.11 Traffic from CNVMs to BBUVM and the location of CNVMs

Table 4.10 operation of CNVM location constraints

Input	Constraints	Outcome	σE_p	The value of σE_p that satisfies both constraints
$\sum_{h\in H} \lambda B_{ph}$	$\sigma E_p \geq \eta \cdot \sum_{h\in H} \lambda B_{p,h}$	$\eta \cdot \sum_{h\in H} \lambda B_{p,h} \ll 1$	1	1

> 0	$\sigma E_p \leq 1 + \sum_{h \in H} \lambda B_{ph} - \eta$	$1 + \sum_{h \in H} \lambda B_{ph} - \eta > 1$	0 or 1	
$\sum_{h \in H} \lambda B_{ph} = 0$	$\sigma E_p \geq \eta \cdot \sum_{h \in H} \lambda B_{p,h}$	$\eta \cdot \sum_{h \in H} \lambda B_{p,h} = 0$	0 or 1	0
	$\sigma E_p \leq 1 + \sum_{h \in H} \lambda B_{ph} - \eta$	$1 + \sum_{h \in H} \lambda B_{ph} - \eta < 1$	0	

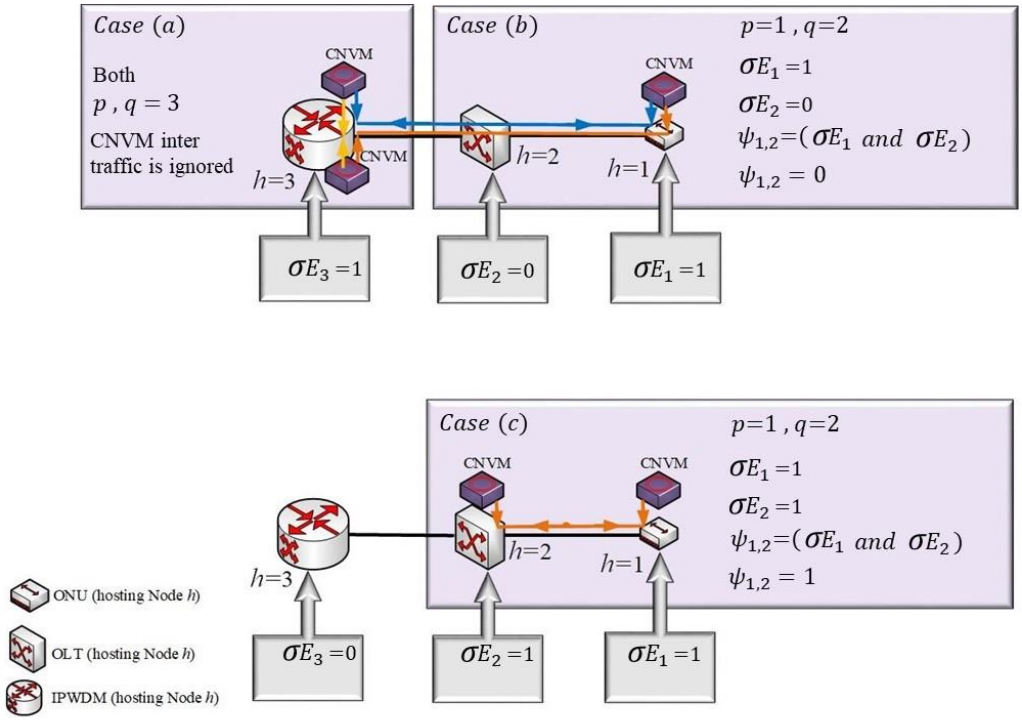


Figure 4.12 CNVMs common location

7) Flow conservation of the total traffic to the RRH nodes

$$\sum_{y \in TN_x} \lambda R_{x,y}^{h,r} - \sum_{y \in TN_x} \lambda R_{y,x}^{h,r} = \begin{cases} \lambda R_{h,r} & if\ x = h \\ -\lambda R_{h,r} & if\ x = r \\ 0 & otherwise \end{cases} \quad (4.38)$$

$$\forall r \in R, \forall h \in H, \forall x \in T$$

8) Total traffic between any two hosting nodes

$$\lambda T_{p,q} = \lambda E_{p,q} + \lambda B_{p,q}$$
$$\forall p, q \in H: p \neq q \tag{4.39}$$

9) Flow conservation of processing nodes communication traffic

$$\sum_{y \in TN_x \cap H} \lambda T_{x,y}^{p,q} - \sum_{y \in TN_x \cap H} \lambda T_{y,x}^{p,q} = \begin{cases} \lambda T_{p,q} & if\ x = p \\ -\lambda T_{p,q} & if\ x = q \\ 0 & otherwise \end{cases} \tag{4.40}$$
$$\forall p, q, x \in H: p \neq q$$

10) Cloud total workload

$$\Psi C_h = \Psi B_h \cdot \sigma B_h + \Psi E_h \cdot \sigma E_h$$
$$\forall h \in H \tag{4.41}$$

Constraint (4.38) represents the flow conservation of the total traffic to the RRH nodes. Figure 4.13 illustrates the principle of flow conservation, and for clarification purposes, it is applied to constraint (4.38). Constraint (4.39) represents the total traffic between any two hosting nodes (p, q) which is caused by virtual machines communication whilst its flow conservation is represented by constraint (4.40). Constraint (4.42) determines the total cloud workload at node *h*.

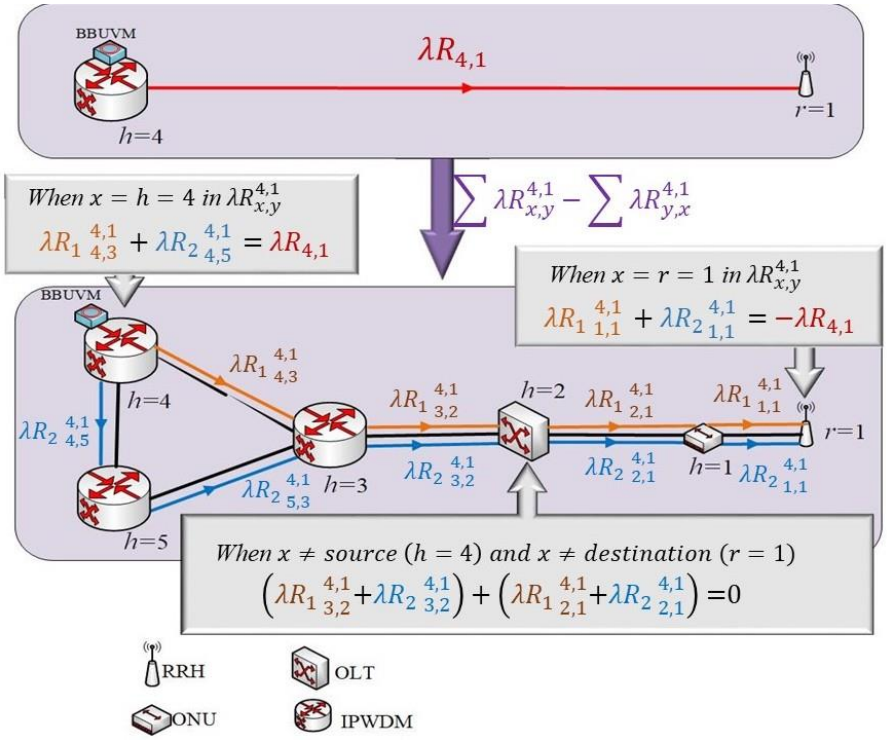


Figure 4.13 Principle of traffic flow conservation

11) Total traffic between IP over WDM nodes

$$\lambda N_{i,j} = \sum_{p \in H} \sum_{q \in H, q \neq p} \lambda T^{p,q}_{i,j} + \sum_{h \in H} \sum_{r \in R} \lambda R^{h,r}_{i,j} \quad \forall i, j \in N: i \neq j \tag{4.42}$$

12) Flow conservation in the IP layer of the IP over WDM network

$$\sum_{j \in N} \lambda N^{s,d}_{i,j} - \sum_{j \in N} \lambda N^{s,d}_{j,i} = \begin{cases} \lambda N_{i,j} & if\ i = s \\ -\lambda N_{i,j} & if\ i = d \\ 0 & otherwise \end{cases} \quad \forall s, d, i, j \in N, s \neq d \tag{4.43}$$

13) Virtual link capacity of the IP over WDM network

$$\sum_{s \in N} \sum_{d \in N: s \neq d} \lambda N^{s,d}_{i,j} = W_{ij} \cdot B \quad \forall i, j \in N, i \neq j \tag{4.44}$$

14) Flow conservation in the optical layer of the IP over WDM network

$$\sum_{n \in NN_m} W_{m,n}^{i,j} - \sum_{n \in NN_m} W_{n,m}^{i,j} = \begin{cases} W_{i,j} & if\ n = i \\ -W_{i,j} & if\ n = j \\ 0 & otherwise \end{cases} \quad \forall i, j, m \in N, i \neq j \tag{4.45}$$

15) Number of wavelength channels

$$\sum_{i \in N} \sum_{j \in N: i \neq j} W_{m,n}^{i,j} = w \cdot f_{mn} \quad \forall m \in N, \forall n \in NN_m \tag{4.46}$$

16) Total number of wavelength channels

$$W_{mn} = \sum_{i \in N} \sum_{j \in N: i \neq j} W_{m,n}^{i,j} \quad \forall m \in N, \forall n \in NN_m \tag{4.47}$$

17) Number of aggregation ports

$$\Lambda_i = \left(\sum_{i \in N, i \neq j} \lambda N_{i,j} \right) / B \quad \forall i \in N \tag{4.48}$$

Constraint (4.42) determines the total traffic flows between any two IP over WDM nodes. Constraint (4.43) represents the flow conservation of the traffic in the IP layer of the IP over WDM network. Constraint (4.44) determines the link capacity between any two IP over WDM nodes. Constraint (4.45) represents the flow conservation in the optical layer of the IP over WDM network. It ensures that the total expected number of incoming wavelengths to the IP over WDM nodes of the virtual link (i, j) is equal to the total number of outgoing wavelengths of that

link. Constraint (4.46) and (4.47) are the constraints of the physical link (m, n). Constraint (4.46) ensures that the total number of wavelength channels in the logical link (i, j) that traverse the physical link (m, n) does not exceed the fiber capacity. Constraint (4.47) determines the number of wavelength channels in the physical link and ensures it is equal to the total number of wavelength channels in the virtual link traversing that physical link. Constraint (4.48) determines the required number of aggregation ports in each IP over WDM router.

4.6 MILP model setup and results

Five IP over WDM nodes were considered constituting the optical backbone network of the proposed architecture. Each IP over WDM node in turn has been attached to two PONs with one OLT and two ONUs for each. Accordingly, the network topology has 10 OLTs and 20 ONUs. In addition, each ONU is connected to one RRH node as shown in Figure 4.14. The case where each RRH unit requests a BBUVM and each BBUVM in turn request a CNVM was considered. Also, a fixed traffic between CNVMs was considered as an internal communication traffic. Beside this, the traffic between CNVM and BBUVM was investigated with different reduction factors (from 10% to 90%). This reduction is due to the BBUVM processing. Moreover, we considered ranges of BBUVM, and CNVM workloads. The BBUVM workload (WL) was set to (10%, 30%, and 50%) of the ONU processor capacity and for each value a range of CNVM workload and reduction factor were considered. The traffic requested by each RRH node is uniformly and randomly generated with a maximum of 10 Gbps. All these factors were investigated with the impact of CNVMs communication where the traffic between

two CNVMs was considered 10% of the traffic between CNVMs and BBUVMs. This traffic was considered to maintain the handover from one CNVM to another CNVM. Other input parameters are listed in Table 4.11. Generally, two scenarios were considered: virtualisation in both IP over WDM and PON and virtualisation in IP over WDM network only; where the results of the two scenarios were compared.

Figs 4.12 - Fig. 4.14 illustrate the total power saving recorded at BBUVM workload 10%, 30% and 50% respectively of the ONU node processing capacity recoded under different CNVM workloads and traffic reductions. In all trends, the main contributor in the power saving is the traffic reduction/ expansion. The virtualisation in both IP over WDM and PON shows a maximum power saving of 7% compared to the virtualisation in IP over WDM only recorded when high traffic (low reduction factor) flows in the network from CNVMs toward BBUVMs. This is due to the extension in the range of candidate locations that might host the virtual machines in the case of virtualisation in both IP over WDM and PON. This case allows accommodating BBUVMs close to the RRH nodes which ultimately causes a short path for the expanded traffic that flows from BBUVMs to RRHs nodes. At high traffic reduction (for example 90% traffic reduction), the traffic between the BBUVM and RRH is very high compared to the traffic between the BBUVM and the CNVM. Therefore, this scenario is very sensitive to the optimal placement of BBUVM and generates maximum savings when the BBUVM is placed closest to the RRH. As the traffic reduction becomes smaller, the savings drop below 7%. When the BBUVM consumes more power as a proportion of the CNVM power consumption (x axis in Fig 4.12 – Fig. 4.14), the savings increase. Comparing Fig. 4.12 – Fig. 4.14, the savings are comparable, but are slightly smaller in Fig. 4.13

compared to Fig. 4.12 and even smaller in Fig. 4.14 compared to Fig. 4.13. This is due to the increase in the BBUVM power consumption and therefore more power out of the total power is used in processing. The power saving due to traffic reduction remains the same. As a result, the percentage reduction in power consumption due to lower power usage attributed to traffic leads to lower overall saving. These changes are however very small.

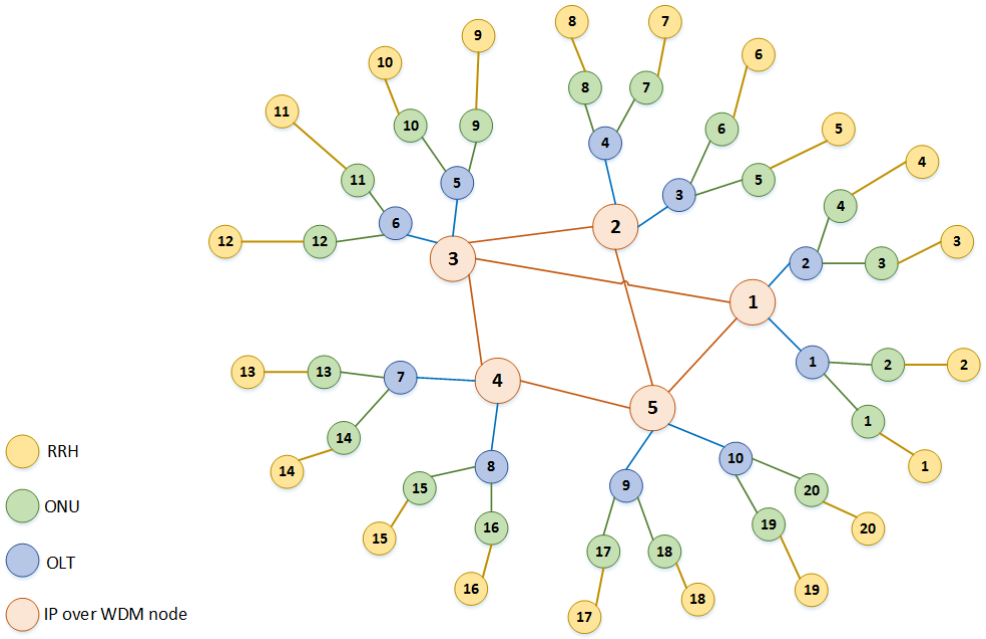


Figure 4.14 Tested network topology

Table 4.11 Input parameter for the MILP model

ONU maximum power consumption (ΩU)	15 (W) [233]
OLT maximum power consumption (ΩL)	3000 (W) [238]
OLT idle power (ΩLd)	60 (W) [231]
OLT maximum capacity (CL)	1440 (Gbps) [238]
ONU maximum capacity (CU)	10 (Gbps) [233]
RRH node power consumption (ΩR_x)	1140 (W) [234]
Maximum power consumption for hosting VMs at ONU	5.9 (W) [239]

node (ΩP_U)	
Maximum power consumption for hosting VMs at OLT node (ΩP_L)	27 (W) [240]
Maximum power consumption for hosting VMs at IP over WDM node (ΩP_N)	86 (W) [241]
Capacity IP over WDM wavelength channel (B)	40 (Gbps) [236]
Number of wavelength per fibre in IP over WDM (w)	16 [232]
Transponder power consumption (ΩT)	75 [232]
Router port power consumption (ΩRP)	825 (W) [172]
Regenerator power consumption (ΩG)	335 (W) [172]
EDFA power consumption (ΩE)	55 (W) [172]
Maximum span distance between EDFAs (S)	80 (km) [236]

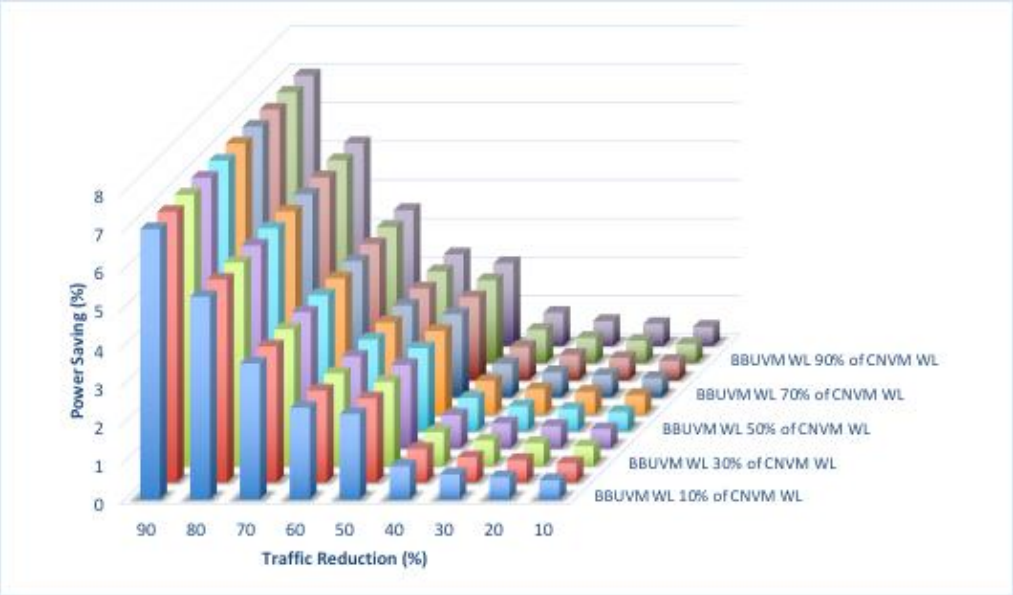


Figure 4.15 Total power saving at BBUVM workload 10% of the ONU under different traffic reduction and CNVM workloads

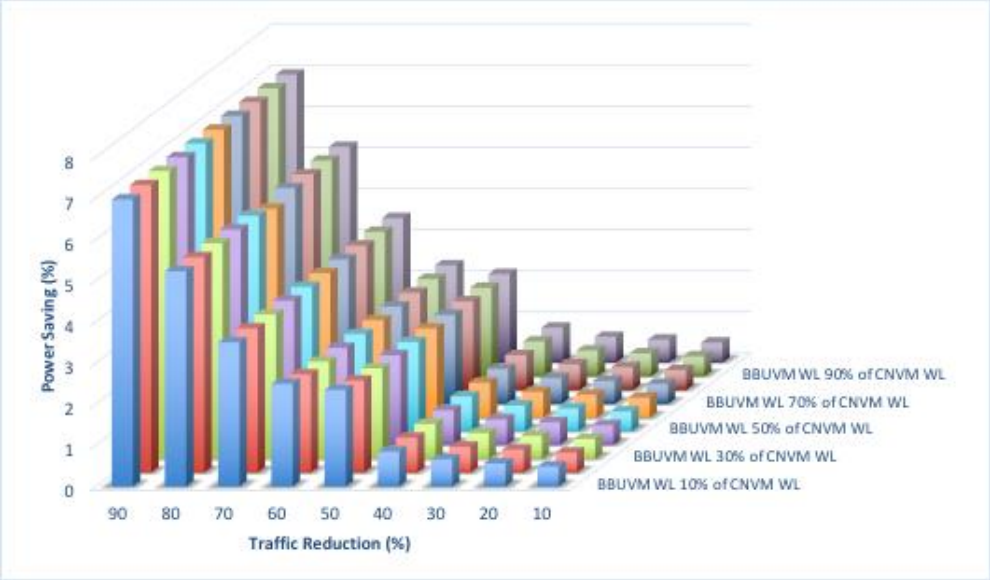


Figure 4.16 Total power saving at BBUVM workload 30% of the ONU under different traffic reduction and CNVM workloads

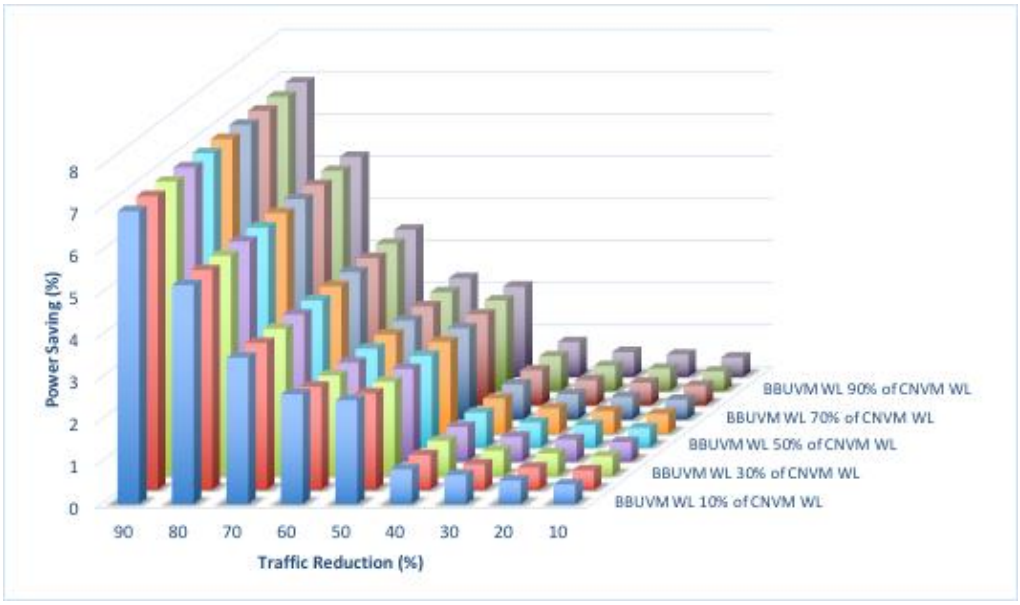


Figure 4.17 Total power saving at BBUVM workload 50% of the ONU under different traffic reduction and CNVM workloads

4.7 Summary

This chapter has introduced an optical-based framework for energy efficient NFV in 5G networks. In this framework the main functionalities of the mobile core

network are virtualised and provided as a core network virtual machine (CNVM). In addition, the base band unit of eNodeB (BBU) is virtualised and provisioned as a "BBU virtual machine" (BBUVM). An MILP optimisation model was developed with the objective of optimising the total power consumption associated with the requests of RRH. The developed model considers 14 VMs of both BBUVMs and CNVMS with different ranges of normalised workloads that have been uniformly and randomly distributed over 30 RRH nodes. Two main approaches were investigated by the MILP models: virtualisation in only IP over WDM network and virtualisation in both PON and IP over WDM network. Virtualisation in the IP over WDM and GPON networks approach shows an average saving in power consumption of 22% compared to the approach where virtualisation is restricted in the IP over WDM network. Virtualisation in the PON network adds a high level of flexibility due to virtual machine behaviour. The virtual machines can be migrated, replicated, or distributed according to the demand and the satisfaction of the optimisation goal, which is the minimisation of total power consumption.

To investigate the impact of the traffic between CNVMs and the traffic reduction / expansion caused by BBUVM, an MILP model was developed to extend the previous model by considering these factors with a wide range for VM workloads. The MILP model results show that scenario of virtualization in both IP over WDM and PON has less power consumption than the virtualisation in IP over WDM network only. It general, the virtualisation in both IP over WDM and PON shows a maximum power saving of 7% compared to the virtualisation in IP over WDM only recorded when high traffic (low reduction factor) flows in the network from CNVMs toward BBUVMs.

Chapter 5
NFV in 5G Mobile Networks: The impact of number of users, backhaul / fronthaul traffic and base-band workload

5.1 Introduction

A framework for energy efficient NFV in 5G networks has been introduced in the previous chapter where different VMs with different workloads were investigated as well as the traffic reduction caused by BBUVM processing for the purpose of energy efficiency. In addition, the impact of virtualisation in PON was investigated and compared to the case where the virtualisation is restricted in IP over WDM network only.

This chapter extends to the work done in Chapter 4 by investigating the impact of the backhaul to the fronthaul traffic ratio and the influence of the inter-traffic between VMs of the mobile core functions. In addition, it investigates the effect of variation of the total number of users during the day on the energy-efficiency and the optimisation problem. The MILP model is validated by two heuristic models: An Energy-Efficient NFV without Inter-traffic (EENFVnoITr) heuristic for no CNVMs inter-traffic and an Energy-Efficient NFV with CNVMs inter-traffic (EENFVwithITr) heuristic for the case where CNVMs inter-traffic is considered. In addition, different values of CNVMs inter-traffic are considered.

5.2 Fronthaul, Backhaul configuration and the amount of BBU workload

This section illustrates the configuration of the fronthaul and backhaul used in the proposed network; so that the ratio of the backhaul to the fronthaul data rate could be calculated. Fronthaul is the network segment that connects the remote radio head (RRH) to the baseband unit (BBU) [242], whilst the network segment that connects the BBU to the mobile core network (CN) is called "backhaul" [243]. The internal interface of the fronthaul is defined as a result of the digitisation of the radio signal according to a number of specifications. The well-known and most used specification among radio access network (RAN) vendors is the Common Public Radio Interface (CPRI) specification [244] which is implemented using digital radio over fibre (D-RoF) techniques. On the other hand, the backhaul interface leverages Ethernet networks as they are the most cost effective network for transporting the backhaul IP packets [245, 246].

In order to adequately determine the data rate in each network segment (backhaul and fronthaul), we will start with the physical layer of the current mobile network which is the Long-Term Evolution (LTE). LTE network uses single-carrier frequency-division multiple access (SC-FDMA) uplink (UL), whilst orthogonal frequency-division multiple access (OFDM) is used in downlink (DL) [247]. In both techniques, the transmitted data are turbo coded and modulated using one of the following modulation: QPSK, 16QAM, or 64QAM with 15 kHz subcarriers spacing [248]. A generic frame is defined in LTE which has 10 ms duration and 10 equal-size subframes. Each subframe is divided into two slot periods of 0.5 ms duration

[249]. Depending on the cyclic prefix (CP), slots in OFDMA have either 7 symbols for normal CP or 6 symbols for extended CP [250]. Figure 5.1 illustrates an LTE downlink frame with normal CP. In the LTE frames, a resource element (RE) is the smallest modulation structure which has one subcarrier of 15 kHz by one symbol [251]. Resource elements are grouped into a physical resource block (PRB) which has dimensions of 12 consecutive subcarriers by one slot (6 or 7 symbols). Therefore, one PRB has a bandwidth of 180 kHz (12×15 kHz). Different transmission bandwidths use different number of physical resource blocks (PRBs) per time slot (0.5 ms) which are defined by 3GPP [252]. Figure (2) illustrates the LTE downlink resource grid. For instance, 10 MHz transmission bandwidth has 50 PRBs whilst 20 MHz has 100 PRBs [253]. If 10 MHz bandwidth is used with 64QAM (6 bits/symbol) and 7 OFDM symbols (short CP), we have:

$$\frac{\left(50\ RB\ \times \frac{12\ subcarriers}{RB} \times \frac{7\ symbols}{subcarrier} \times \frac{6\ bits\ (QAM)}{symbol}\right)}{0.5\ ms\ time\ slot} = 50.4\ Mbps \qquad (5.1)$$

and by considering (12.6%) the protocol overhead (or 87.4 system efficiency) [254] , the backhaul (IP) data rate is calculated as:

$$50.4\ Mbps\ \times\ 0.874\ system\ efficiency = 44.0496\ Mbps \qquad (5.2)$$

It is worth mentioning that for each transmission antenna, there is one resource grid (50 PRBs for 10 MHz); therefore in 2×2 MIMO the previous data rate is double (100.8) [255].

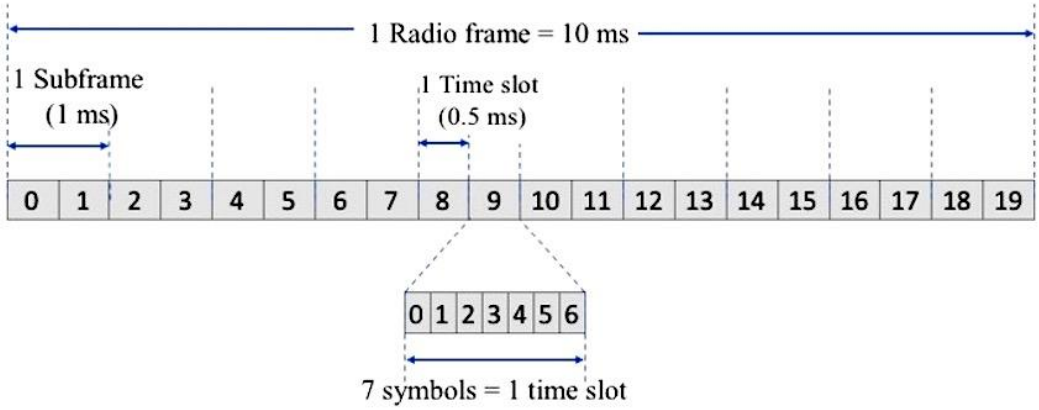


Figure 5.1 LTE downlink frame with normal CP

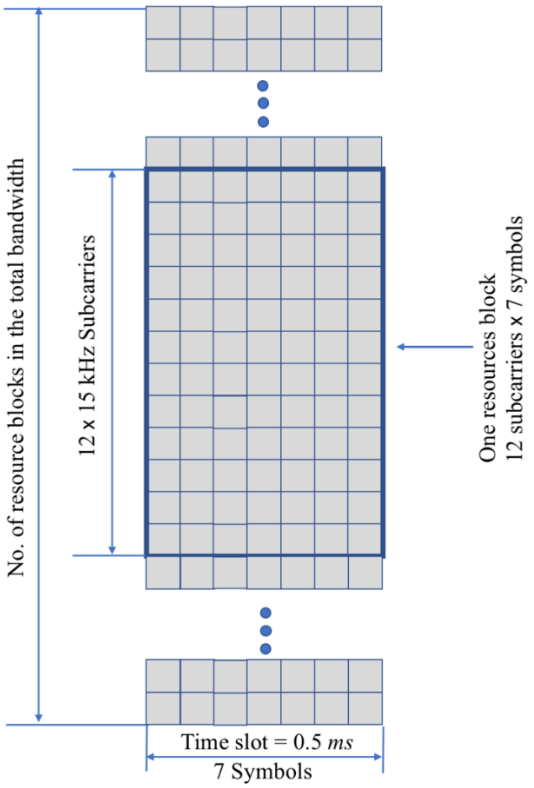


Figure 5.2 LTE downlink resource grid

The transmission of user plane data is achieved in the form of In-phase and quadrature (IQ) components that are sent via one CPRI physical link where each IQ data flow represents the data of one carrier for one antenna that is called Antenna-Carrier (AxC) [256]. A number of parameters affect the data carried by AxC, [255]:

Sampling frequency which is calculated as: subcarrier BW (15 kHz) times the FFT window (size). The FFT size is chosen to be the least multiple of 2 that is greater than the ratio of the radio signal bandwidth to the subcarrier BW. For instance, if the radio bandwidth is 10 MHz, the FFT size is the least multiple of 2 number that is greater than 666.67 (10 MHz / 15 kHz) which is 1028 (2^{10}). In this case the sampling frequency will be calculated as $15\,kHz \times 1024 = 15.36\,MHz$. Using the same approach, the sampling frequency at 20 MHz radio bandwidth system will be 30.72 MHz.

IQ sample width (M-bits per sample): According to the CPRI specification, the IQ sample width supported by CPRI is between 4 and 20 bits per sample for I and Q in the uplink and it is between 8 and 20 in the downlink [256]. For instance, with M = 15 bits per sample; one AxC contains 15 bits per sample for I and 15 bits per sample for Q which are 30 ($2 \times M$) bits per sample I and Q which are transported in sequence: $I_0 Q_0 I_1 Q_1 \dots I_{14} Q_{14}$. The IQ sample data rate can be calculated by multiplying the number of bits per samples by the sampling frequency. For instance; for a radio bandwidth of 10 MHz (fs =15.36 MHz) and IQ samples 15 (M = 15) the IQ data rate will be:

$$(2 \times M) \times f_s = (30\ bits/sample) \times 15.36\,MHz = 0.4608\,Gbps.$$

CPRI data rate is designed based on Universal Mobile Telecommunications System (UMTS) chip rate [256] which is 3.84 MHz [257, 258]. Therefore, one basic CPRI frame is created every Tc= 260.416 ns (1/3.84 MHz) and this duration should remain constant for all CPRI options and data rate. According to CPRI specification in [256], one basic CPRI frame consists of 16 words indexed (W=0…15), where the first word is reserved for control. The length of the frame word (T) depends on the CPRI line rate as specified by CPRI specification in [256]. Accordingly, the transmission of AxC data will be expanded by a factor of 16/15 (15 bits payload, 1 bit control and management). In addition to the sampling rate f_s that is calculated earlier, AxC data needs to be coded using either 8B/10B or 64B/66B.

To put all these calculations together, let's start with the number of bits per word in the CPRI frame. The number of bits per word is equal to the total number of bits per frame divided by the frame payload words (15 words). Recall that the frame duration should be constants (260.416 ns); therefore:

$$\frac{total\ number\ of\ bits\ in\ CPRI\ frame\ (no\ of\ bit\ per\ word\ \times\ 15\ words)}{samples\ of\ IQ\ f_{IQ}} = 260.416\ ns \tag{5.3}$$

$$no\ of\ bits\ per\ word\ (Nbpw) = \frac{f_{IQ} \times 260.416\ ns}{15}. \tag{5.4}$$

One CPRI frame word has Nbpw bits, since the CPRI frame has 16 words:

$$NbpF = Nbpw\ \times\ (15\ payload\ words) + Nbpw\ \times\ 1\ control\ word \tag{5.5}$$

$$NbpF = Nbpw \times (15 + 1) = \frac{f_{IQ} \times 260.416\, ns}{15} \times 16 \tag{5.6}$$

to calculate the data rate in one CPRI frame

$$\frac{NbpF}{260.416\, ns} = \frac{\frac{f_{IQ} \times 260.416\, ns}{15} \times 16}{260.416\, ns} = f_{IQ} \times \frac{16}{15} \tag{5.7}$$

by replacing f_{IQ} with $f_{IQ} = 2 \times M \times f_s$

where M is defined earlier as the number of IQ bits.

In addition, AxC data are coded by either 8B/10B or 64B/66B. By putting these together, the CPRI data rate is calculated as:

$$2 \times M \times f_s \times \frac{16}{15} \times L_{coding} \tag{5.8}$$

note that the previous equation is for one AxC grid, and by considering more than one, the above equation is rewritten as:

$$\begin{aligned} &mobile\ fronthaul\ data\ rate\ (CPRI\ data\ rate) \\ &= 2 \times M \times f_s \times \frac{16}{15} \times L_{coding} \times N_{AxC} \end{aligned} \tag{5.9}$$

A 10 MHz radio signal bandwidth is used in the calculation of backhaul data rate (IP data rate), and for consistency, the calculation of the fronthaul data rate is based on the same bandwidth.

In addition, 10 bits IQ is used with 8B/10B line coding and one AxC grid.

$$\begin{aligned} &mobile\ fronthaul\ data\ rate\ (CPRI\ data\ rate) \\ &= 2 \times 10 \times 15.36\ Mbps \times \frac{16}{15} \times \frac{10}{8} \times 1 = 327.68\ Mbps \end{aligned} \tag{5.10}$$

Finally, the ratio of the backhaul to fronthaul data rate is calculated as:

$$\frac{backhaul\ (IP) data\ rate}{fronthaul\ (CPRI) data\ rate} = \frac{44.0496\ Mbps}{327.68\ Mbps} \times 100\ \% = 13.44\ \% \qquad (5.11)$$

Therefore, depending on coding, sampling, quantisation, and other parameters; the baseband processing adds overheads to the backhaul traffic as it passes through the BBU. In this work the ratio (13.44) calculated in (5.11) is used in our model, whilst the amount of workload in Giga Operation Per Second (GOPS) needed to process one user traffic is used based on the following equation which is explained in [259]:

$$wl = \left(30 \cdot A + 10 \cdot A^2 + 20\frac{M}{6} \cdot C \cdot L\right) \cdot \frac{R}{50} \qquad (5.12)$$

where:

wl: is the baseband workload in (GOPS) needed to process one user traffic,

A: number of antennas used,

M: modulation bits,

C: the code rate,

L: number of MIMO layers

R: number of physical resource blocks allocated for the user.

5.3 MILP model

This section introduces the MILP model that has been developed to minimise the power consumption due to both processing by virtual machines (hosting servers) and the traffic flow through the network. As in the previous chapter, the MILP model considers the same optical-based architecture with two types of VMs (BBUVM and CNVMs) that could be accommodated in ONU, OLT and/or IP over WDM as in Figure 5.3. The maximum number of VM-hosting servers was considered to be 1, 5, and 20 in ONU, OLT, and IP over WDM nodes respectively. All VM-hosting servers were considered as sleep-capable servers for the purpose of VM consolidation (bin packing)

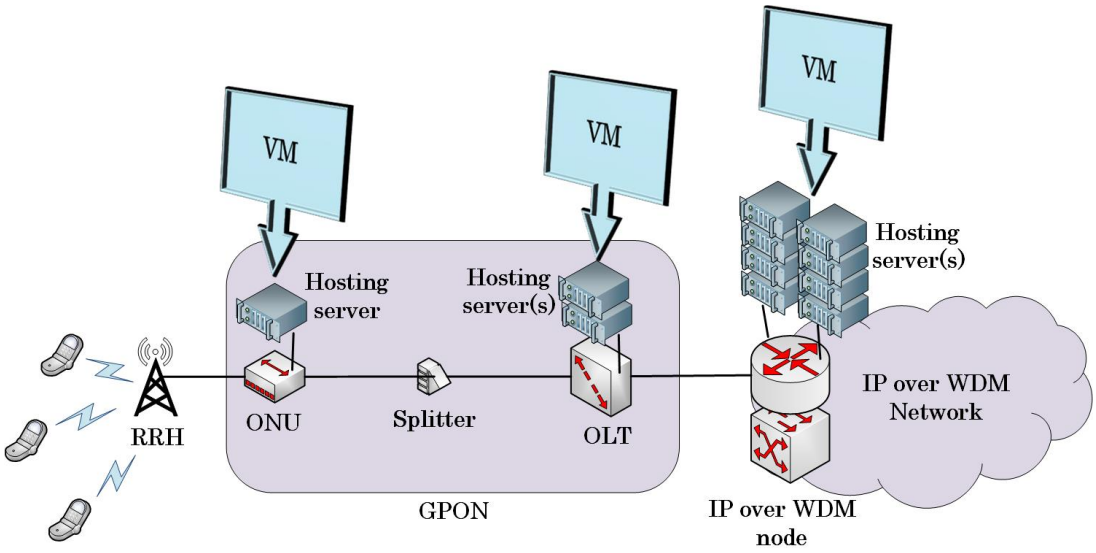


Figure 5.3 The candidate location for hosting virtual machines in the proposed architecture

For a given request, the MILP model responds by selecting the optimum number of virtual machines and their location so that the total power consumption is minimised.

The following indices, parameters, and variables are defined to represent the developed model:

Table 5.1 Energy-efficient NFV MILP model indices

Indices	Comment
x, y	Indices of any two nodes in the proposed model
m, n	Indices of any two nodes in the physical layer of the IP over WDM network
i, j	Indices of any two nodes in the IP layer of the IP over WDM network.
r	Index of RRH node
h, u, p, q	Indices of the nodes where the VM could be hosted

Table 5.2 Energy-efficient NFV MILP model parameters

Parameters	Comment
R	Set of RRH nodes
U	Set of ONU nodes
L	Set of OLT nodes
N	Set of IP over WDM nodes
T	Set of all nodes (RRH, ONU, OLT, and IP over WDM nodes)
NN_m	Set of neighbours of node m in the IP over WDM network, $\forall\, m \in N$
TN_x	Set of neighbours of node x, $\forall\, x \in T$
H	Set of hosting nodes (ONU, OLT, and IP over WDM nodes)
l	Line coding rate (bits per sample)
y	Number of MIMO layers (ie number of data streams)
q	Number of bits used in QAM modulation
a	Number of antennas in a cell
cp	CPRI link data rate
ΨX	Maximum BBU workload needed for fully loaded RRH (GOPS); calculated as: $30 \cdot a + 10 \cdot a^2 + 20 \cdot q \cdot l \cdot y$

ΨS	Server CPU maximum workload (GOPS)
ΨC_h	Workload needed for hosting one CNVM (GOPS)
ρ_r	Number of active users connected to RRH node r
n	Maximum number of physical resources blocks for cell (r)
pb	Number of physical resource blocks per user
λR_r	RRH node r traffic demand (Gbps); calculated as: $[(pb/n) \cdot cp \cdot \rho_r]$, where $r \in R$
$\nabla_{p,q}$	Intra-traffic between core network VMs (CNVM) at hosting nodes p, and q (Gbps)
α	The ratio of the backhaul to the fronthaul traffic (unitless)
ΩU	ONU maximum power consumption (W)
ΩL	OLT maximum power consumption (W)
ΩLd	OLT idle power (W)
CL	OLT maximum capacity (Gbps)
CU	ONU maximum capacity (Gbps)
ΩR_x	Power consumption of the Remote Radio Head (RRH) connected to ONU node x (W)
ΩS	Server maximum power consumption (W)
ΩSd	Server idle power (W)
ΩH_h	Maximum power consumption of hosting VMs at not h
β	Large number (unitless)
η	Very small number (unitless)
B	Capacity of the wavelength channel (Gbps)
w	Number of wavelengths per fiber
ΩT	Transponder power consumption (W)
ΩRP	Router power consumption per port (W)
ΩG	Regenerator power consumption (W)
ΩE	EDFA power consumption (W)
$NG_{m,n}$	Number of regenerators in the optical link (m, n)
S	Maximum span distance between EDFAs (km)
$D_{m,n}$	Distance between node pair (m, n) in the IP over WDM network (km)
$A_{m,n}$	Number of EDFAs between node pair (m, n) calculated as $A_{m,n} =$

	$\left((D_{mn}/S)-1\right)+2$

Table 5.3 Energy-efficient NFV MILP model variables

Variables	Comment
$\lambda B_{p,h}$	Traffic from CNVMs in node p to the BBUVMs in node h (Gbps)
$\lambda R_{h,r}$	Traffic from BBUVMs in node h to the RRH node r (Gbps)
$\sigma B_{h,r}$	Binary indicator, set to 1 if the node h hosts BBUVMs to serve the RRH node r, 0 otherwise
σB_h	Binary indicator, set to 1 if the node h hosts a BBUVM, 0 otherwise
$\sigma E_{p,h}$	Binary indicator, set to 1 if the node h hosts CNVMs to serve the BBUVMs at hosting node h, 0 otherwise
σE_p	Binary indicator, set to 1 if the hosting node p hosts CNVMs is, 0 otherwise
$\psi_{p,q}$	Binary indicator, set to 1 if two different hosting nodes p and q host CNVMs, 0 otherwise. It is equivalent to the ANDing of the two binary variables (σE_p, σE_q).
$\sigma \chi_h$	Binary indicator, set to 1 if the hosting node h hosts any virtual machine of any type, 0 otherwise. It is equivalent to the ORing of the two binary variables (σB_h, σE_h).
$\lambda E_{p,q}$	Traffic between hosting nodes due to CNVMs communication (Gbps)
$\lambda T_{p,q}$	Total traffic from node p to node q caused by CNVM to CNVM traffic and CNVM to BBUVM traffic (Gbps)
$\lambda R_{x,y}^{h,r}$	Traffic from hosting node h to RRH node r that traverses the link between the nodes (x, y) in the network in Gb/s
$\lambda T_{x,y}^{p,q}$	Total traffic from node p to node q that traverses the link between the nodes (x, y) in the network (Gbps)
ΨB_h	BBU workload at node h (GOPS)
Ψi_h	The integer part of the total normalised workload at node h.
Ψf_h	The fractional part of the total normalised workload at node h.
$W_{i,j}$	Number of wavelength channels in the virtual link (i, j)
$W_{m,n}^{i,j}$	Number of wavelength channels in the virtual link (i, j) that traverse the physical link (m, n)
$f_{m,n}$	Number of fibres in the physical link (m, n)

$W_{m,n}$	Total number of wavelengths in the physical link (m, n)
Λ_m	Number of aggregation ports of the router at node m

The total power consumption is composed of:

5) The power consumption of RRHs and ONUs calculated as the total traffic of passing through ONUs multiplied by the ONU energy per transmitted bit added to the total power consumption of RRHs

$$\sum_{x \in U} \left[\Omega R_x + \frac{\Omega U}{CU} . \left(\sum_{h \in H} \sum_{r \in R} \sum_{y \in TN_x} \lambda R_{x,y}^{h,r} + \sum_{p \in H} \sum_{q \in H : p \neq q} \sum_{y \in TN_x \cap H} \lambda T_{x,y}^{p,q} \right) \right]$$

6) The power consumption of the OLTs calculated as the total traffic of passing through OLTs multiplied by the OLT energy per transmitted bit

$$\sum_{x \in L} \left[\Omega L d + \frac{\Omega L - \Omega L d}{CL} . \left(\sum_{h \in H} \sum_{r \in R} \sum_{y \in TN_x} \lambda R_{x,y}^{h,r} + \sum_{p \in H} \sum_{q \in H : p \neq q} \sum_{y \in TN_x \cap H} \lambda T_{x,y}^{p,q} \right) \right]$$

7) The power consumption of the IP over WDM network is composed of :

   a) Aggregation ports power consumption calculated as the total number of aggregation ports multiplied by power consumption of a single port:

$$\left(\Omega RP \cdot \sum_{m \in N} \Lambda_m \right)$$

b) High speed ports power consumption calculated as the multiplication of total number of high speed ports by the power consumption of a single port

$$\left(\Omega RP \cdot \sum_{m \in N} \sum_{n \in NN_m} W_{m,n}\right)$$

c) Transponders power consumption calculated as the multiplication of total number of wavelength by the power consumption of a single transponder:

$$\left(\Omega T \cdot \sum_{m \in N} \sum_{n \in NN_m} W_{m,n}\right)$$

d) EDFAs power consumption calculated as the multiplication of the total number of EDFAs by the total number of fibres by the power consumption of a single EDFA

$$\left(\Omega E \cdot \sum_{m \in N} \sum_{n \in NN_m} A_{m,n} \cdot f_{m,n}\right)$$

e) Regenerators power consumption calculated as the total number of regenerators time the power consumption of a single generator

$$\left(\Omega G \cdot \sum_{m \in N} \sum_{n \in NN_m} NG_{m,n} \cdot W_{m,n}\right)$$

8) The total power consumption of VMs and hosting servers

$$\sum_{h \in H} \left(\Omega Sd \cdot (\Psi i_h + \sigma \chi_h) + \Psi f_h \cdot (\Omega S - \Omega Sd)\right)$$

The model objective is to minimise the total power consumption as follows:

Minimise

$$\sum_{x\in U}\left[\Omega R_x + \frac{\Omega U}{CU}.\left(\sum_{h\in H}\sum_{r\in R}\sum_{y\in TN_x}\lambda R_{x,y}^{h,r} + \sum_{p\in H}\sum_{q\in H:p\neq q}\sum_{y\in TN_x\cap H}\lambda T_{x,y}^{p,q}\right)\right]$$

$$+\sum_{x\in L}\left[\Omega Ld + \frac{\Omega L - \Omega Ld}{CL}.\left(\sum_{h\in H}\sum_{r\in R}\sum_{y\in TN_x}\lambda R_{x,y}^{h,r} + \sum_{p\in H}\sum_{q\in H:p\neq q}\sum_{y\in TN_x\cap H}\lambda T_{x,y}^{p,q}\right)\right]$$

$$+\left(\Omega RP\cdot\sum_{m\in N}\Lambda_m\right)+\left(\Omega RP\cdot\sum_{m\in N}\sum_{n\in NN_m}W_{m,n}\right)$$

$$+\left(\Omega T\cdot\sum_{m\in N}\sum_{n\in NN_m}W_{m,n}\right)+\left(\Omega E\cdot\sum_{m\in N}\sum_{n\in NN_m}A_{m,n}\cdot f_{m,n}\right)$$

$$+\left(\Omega G\cdot\sum_{m\in N}\sum_{n\in NN_m}NG_{m,n}\cdot W_{m,n}\right)$$

$$+\sum_{h\in H}\left(\Omega Sd\cdot(\Psi i_h+\sigma\chi_h)+\Psi f_h\cdot(\Omega S-\Omega Sd)\right)$$

Subject to the following constraints:

1) Traffic from CNVM to BBUVM

$$\sum_{p\in H}\lambda B_{p,h} = \alpha\cdot\sum_{r\in R}\lambda R_{h,r} \qquad (5.13)$$

$$\forall h\in H$$

2) Traffic to RRH nodes

$$\sum_{h \in H} \lambda R_{h,r} = \lambda R_r \qquad \forall r \in R \tag{5.14}$$

Constraint (5.13) represents the traffic from CNVMs to the BBUVM in node h where α is a unitless quantity which represents the ratio of backhaul to fronthaul traffic. Note that this constraint allows a BBUVM to receive traffic from more than a single CNVM, which may occur for example in network slicing.

Constraint (5.14) represents the traffic to RRH nodes from all BBUVMs that are hosted in hosting nodes. This enables an RRH to receive traffic from more than a single BBUVM (network slicing).

3) The served RRH nodes and the location of BBUVM

$$\beta \cdot \lambda R_{h,r} \geq \sigma B_{h,r} \qquad \forall r \in R, \forall h \in H \tag{5.15}$$

$$\lambda R_{h,r} \leq \beta \cdot \sigma B_{h,r} \qquad \forall r \in R, \forall h \in H \tag{5.16}$$

$$\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r} \geq \sigma B_h \qquad \forall h \in H \tag{5.17}$$

$$\sum_{\forall r \in R} \lambda R_{h,r} \leq \beta \cdot \sigma B_h \qquad \forall h \in H \tag{5.18}$$

Constraint (5.15) and (5.16) ensure that the RRH node r is served by the BBUVM that is hosted at node h as illustrated in Figure 5.4. Constraints (5.17) and (5.18) determine the location of BBUVM; β is a large enough number to ensure that σB_{hr} and σB_h are equal to 1 when $\sum_{\forall r \in R} \lambda R_{hr} > 0$. In constraint (5.17) there are two possibilities for the value of ($\sum_{\forall r \in R} \lambda R_{h,r}$) which are either zero (no traffic from *h* to *r*) or greater than zero (there is a traffic from *h* to *r*). When the value of $\sum_{\forall r \in R} \lambda R_{h,r}$ is zero, the left-hand side of the inequality ($\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r}$) should be zero and this sets the value of σB_h to zero. In the second case when the value of $\sum_{\forall r \in R} \lambda R_{h,r}$ is greater than zero, the left-hand side of the inequality ($\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r}$) will be much greater than 1 because of the large value β. In this, the value of σB_h may be set to 1 or zero. In the same way constraint (5.18) sets the value of σB_h. Table 5.4 illustrates the operation of constraints (5.17) and (5.18).

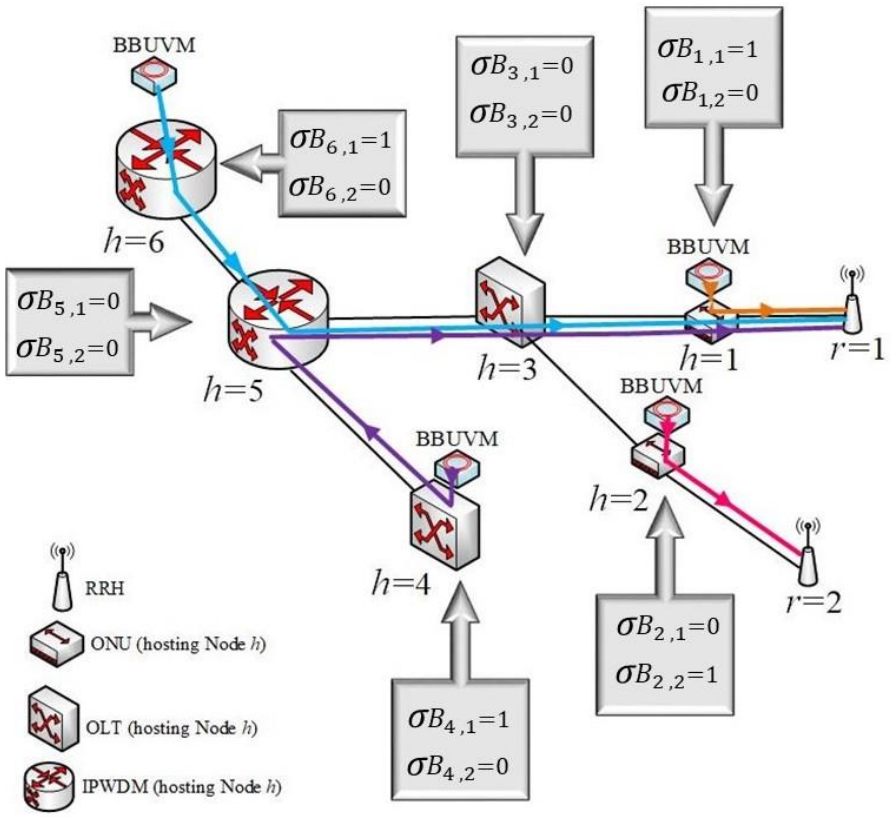


Figure 5.4 BBUVM locations and the traffic toward RRH nodes

Table 5.4 BBUVM constraints operation

Input	Constraint	Outcome	σB_h	Value of σB_h that satisfies both constraints
$\sum_{\forall r \in R} \lambda R_{h,r} > 0$	$\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r} \geq \sigma B_h$	$\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r} \gg 1$	0 or 1	1
	$\sum_{\forall r \in R} \lambda R_{h,r} \leq \beta \cdot \sigma B_h$	$\beta \cdot \sigma B_h \gg 1$	1	
$\sum_{\forall r \in R} \lambda R_{h,r} = 0$	$\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r} \geq \sigma B_h$	$\beta \cdot \sum_{\forall r \in R} \lambda R_{h,r} = 0$	0	0

	$\sum_{\forall r \in R} \lambda R_{h,r} \leq \beta \cdot \sigma B_h$	$\beta \cdot \sigma B_h = 0$	0 or 1	

4) CNVM locations

$$\beta \cdot \lambda B_{p,h} \geq \sigma E_{p,h} \quad \forall p, q \in H, p \neq q \tag{5.19}$$

$$\lambda B_{p,h} \leq \beta \cdot \sigma E_{p,h} \quad \forall p, q \in H, p \neq q \tag{5.20}$$

$$\sigma E_p \geq \eta \cdot \sum_{h \in H} \lambda B_{p,h} \quad \forall p \in H \tag{5.21}$$

$$\sigma E_p \leq 1 + \sum_{h \in H} \lambda B_{ph} - \eta \quad \forall p \in H \tag{5.22}$$

$$\psi_{pq} \leq \sigma E_p \quad \forall p, q \in H, p \neq q \tag{5.23}$$

$$\psi_{p,q} \leq \sigma E_q \quad \forall p, q \in H, p \neq q \tag{5.24}$$

$$\psi_{p,q} \geq \sigma E_p + \sigma E_q - 1 \quad \forall p, q \in H, p \neq q \tag{5.25}$$

5) Hosting any VM of any type

$$\sigma\chi_h \leq \sigma B_h + \sigma E_h \quad \forall h \in H \tag{5.26}$$

$$\sigma\chi_h \geq \sigma B_h \quad \forall h \in H \tag{5.27}$$

$$\sigma\chi_h \geq \sigma E_h \quad \forall h \in H \tag{5.28}$$

Constraints (5.19) and (5.20) ensure that the BBUVMs at node h are served by the CNVMs that are hosted at the node p. Constraints (5.21) and (5.22) determine the location of the CNVMs by setting the binary variable σE_p to 1 if there is a CNVM hosted at node p, where η is very small number. Figure 5.5 illustrates the functions of constraints (5.21) and (5.22) whilst Table 5.5 illustrates their operation. Constraints (5.23) - (5.25) ensure that the CNVMs communicate with each other if they are hosted at different nodes p and q, and this is equivalent to the logical operation $\psi_{p,q} = \sigma E_p \; AND \; \sigma E_q$. Figure 5.6 illustrates the function of constraints (5.23) - (5.25). Constraints (5.26) - (5.28) determine if the hosting node h hosts any VM of any type (BBUVM or CNVM). It is equivalent to the logical operation $\sigma\chi_h = \sigma E_p \; OR \; \sigma E_q$.

Table 5.5 CNVM constraints operation

Input	Constraints	Outcome	σE_p	The value of σE_p that satisfies both constraints
$\sum_{h\in H} \lambda B_{ph} > 0$	$\sigma E_p \geq \eta \cdot \sum_{h\in H} \lambda B_{p,h}$	$\eta \cdot \sum_{h\in H} \lambda B_{p,h} \ll 1$	1	1
	$\sigma E_p \leq 1 + \sum_{h\in H} \lambda B_{ph} - \eta$	$1 + \sum_{h\in H} \lambda B_{ph} - \eta > 1$	0 or 1	
$\sum_{h\in H} \lambda B_{ph} = 0$	$\sigma E_p \geq \eta \cdot \sum_{h\in H} \lambda B_{p,h}$	$\eta \cdot \sum_{h\in H} \lambda B_{p,h} = 0$	0 or 1	0
	$\sigma E_p \leq 1 + \sum_{h\in H} \lambda B_{ph} - \eta$	$1 + \sum_{h\in H} \lambda B_{ph} - \eta < 1$	0	

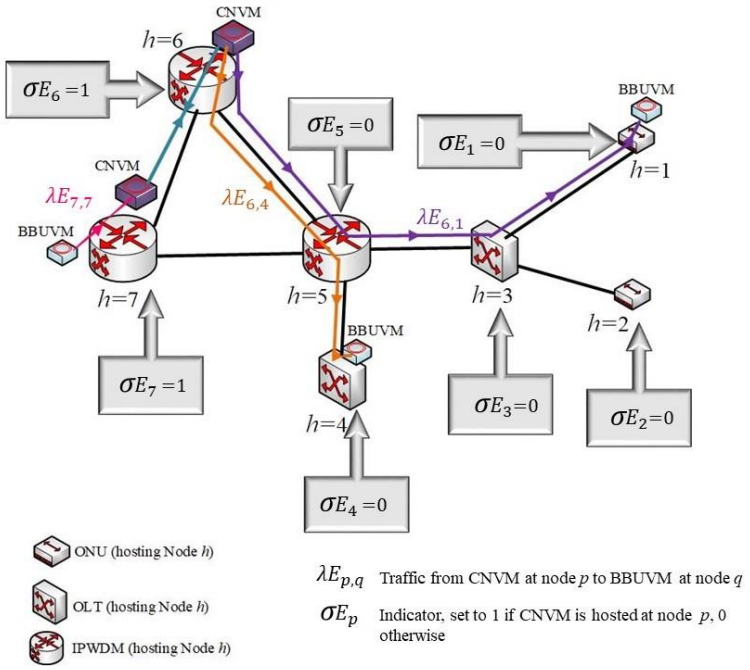


Figure 5.5 CNVM locus and the traffic toward BBUVMs

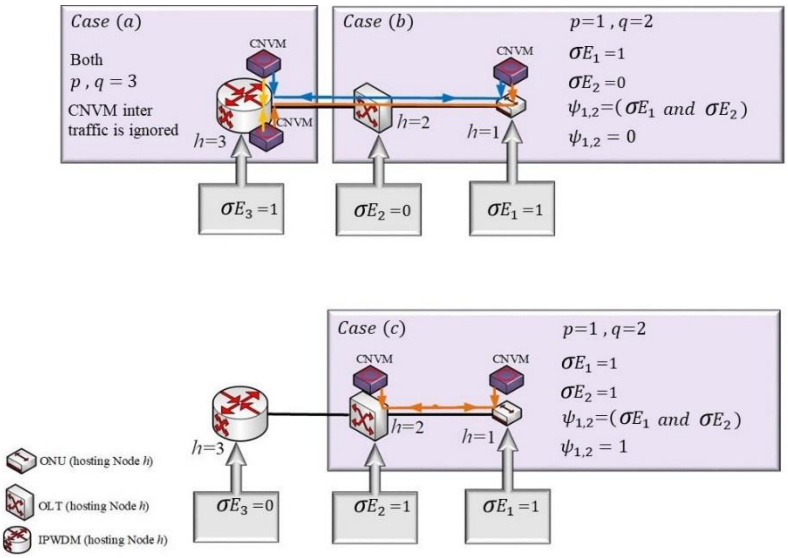


Figure 5.6 CNVM locus and the common locus

6) Communication traffic between CNVMs

$$\lambda E_{p,q} = \nabla_{p,q} \cdot \psi_{p,q} \qquad (5.29)$$

$$\forall p, q \in H : p \neq q$$

7) Total traffic between two hosting nodes

$$\lambda T_{p,q} = \lambda E_{p,q} + \lambda B_{p,q} \qquad (5.30)$$

$$\forall p, q \in H : p \neq q$$

8) Flow conservation of the total traffic to the RRH nodes

$$\sum_{y \in TN_x} \lambda R_{x,y}^{h,r} - \sum_{y \in TN_x} \lambda R_{y,x}^{h,r} = \begin{cases} \lambda R_{h,r} & if\ x = h \\ -\lambda R_{h,r} & if\ x = r \\ 0 & otherwise \end{cases} \qquad (5.31)$$

$$\forall r \in R, \forall h \in H, \forall x \in T$$

9) Flow conservation of hosting nodes communication traffic

$$\sum_{y \in TN_x \cap H} \lambda T_{x,y}^{p,q} - \sum_{y \in TN_x \cap H} \lambda T_{y,x}^{p,q} = \begin{cases} \lambda T_{p,q} & if\ x = p \\ -\lambda T_{p,q} & if\ x = q \\ 0 & otherwise \end{cases} \quad (5.32)$$

$$\forall p, q, x \in H : p \neq q$$

Constraint (5.29) represents the traffic between CNVMs at hosting nodes p and q. Constraint (5.30) represents the total traffic between any two hosting nodes (p, q) which is caused by virtual machines communication. Constraint (5.31) represents the flow conservation of the total fronthaul traffic to the RRH nodes. Figure 5.7 illustrates the principle of flow conservation, and for clarification purposes, it is applied to constraint (5.31). Constraint (5.32) represents the flow conservation of the total traffic between any two hosting nodes that might host virtual machines of any type (BBUVM or CNVM).

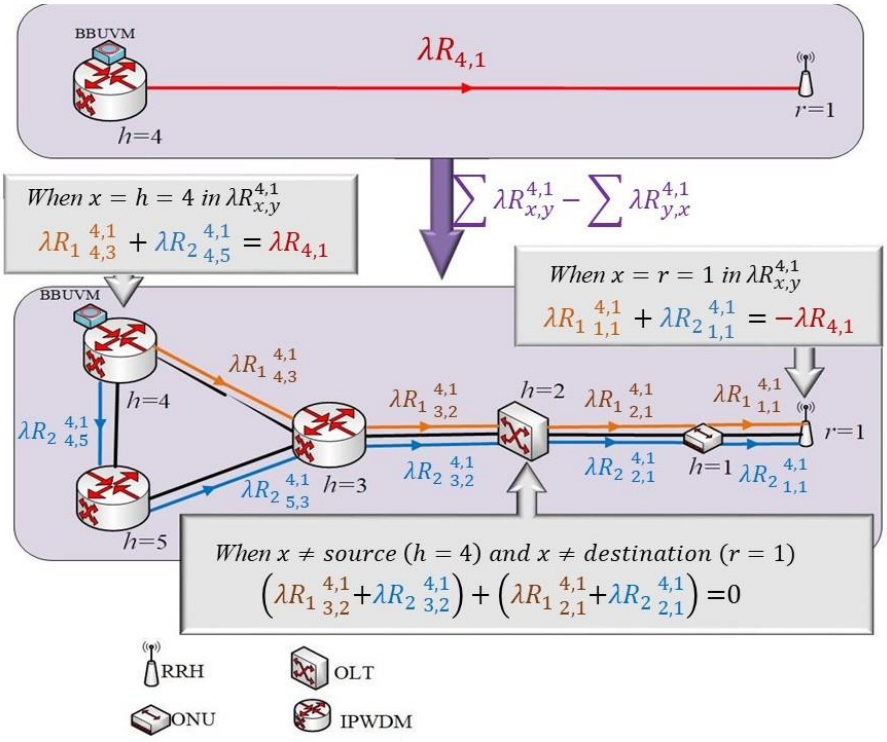


Figure 5.7 Flow conservation principle

10) Total BBU workload at any hosting node h

$$\Psi B_h = \left(\left(\sum_{\forall r \in R} \lambda R_{h,r} \right) \Big/ cp \cdot \right) \Psi \mathrm{X} \qquad \forall h \in H \tag{5.33}$$

11) Total normalized workload at hosting node h

$$\Psi i_h + \Psi f_h = (\Psi B_h + \Psi C_h)/\Psi S \qquad \forall h \in H \tag{5.34}$$

12) Hosting node capacity

$$\left(\Omega Sd \cdot (\Psi i_h + \sigma \chi_h) + \Psi f_h \cdot (\Omega S - \Omega Sd)\right) \leq \Omega H_h \qquad \forall h \in H \tag{5.35}$$

13) GPON link constraints

$$\sum_{h \in H} \sum_{r \in R} \sum_{j \in TN_i \cap L} \lambda R_{i,j}^{h,r} \leq 0 \qquad \forall i \in U \tag{5.36}$$

$$\sum_{p \in H} \sum_{q \in H, q \neq p} \sum_{j \in TN_i \cap L} \lambda T_{i,j}^{p,q} \leq 0 \qquad \forall i \in U \tag{5.37}$$

$$\sum_{h\in H}\sum_{r\in R}\sum_{j\in \mathrm{TN_i}\cap \mathrm{N}} \lambda R_{i,j}^{h,r} \leq 0 \qquad (5.38)$$

$$\forall i \in L$$

$$\sum_{p\in H}\sum_{q\in H, q\neq p}\sum_{j\in \mathrm{TN_i}\cap \mathrm{N}} \lambda T_{i,j}^{p,q} \leq 0 \qquad (5.39)$$

$$\forall i \in L$$

Constraint (5.33) represents the total BBU workload at any hosing node *h*. Constraint (5.34) calculates the total BBU and CNVM normalized workload at any hosting node. The workload is scaled and normalized relative to the server CPU workload and is separated into integer and fractional parts. Constraint (5.35) ensures that the total power consumption of hosting VMs does not exceed the maximum power consumption allocated for each host. Constraints (5.36) – (5.39) ensure that the total PON downlink traffic does not flow in the opposite direction.

14) Virtual Link capacity of the IP over WDM network

$$\sum_{p\in H}\sum_{q\in H, q\neq p} \lambda T_{i,j}^{p,q} + \sum_{h\in H}\sum_{r\in R} \lambda R_{i,j}^{h,r} \leq W_{i,j} \cdot B \qquad (5.40)$$

$$\forall i, j \in N, i \neq \mathrm{j}.$$

15) Flow conservation in the optical layer of IP over WDM network

$$\sum_{n\in NN_m} W_{m,n}^{i,j} - \sum_{n\in NN_m} W_{n,m}^{i,j} = \begin{cases} W_{i,j} & if\ n = i \\ -W_{i,j} & if\ n = j \\ 0 & otherwise \end{cases} \qquad (5.41)$$

$$\forall i, j, m \in N, i \neq j$$

Constraint (5.40) ensures that the total traffic traversing the virtual link (i, j) does not exceed its capacity, in addition it determines the number of wavelength channels that carry the traffic burden of that link. Constraint (5.41) represents the flow

conservation in the optical layer of the IP over WDM network. It ensures that the total expected number of incoming wavelengths for the IP over WDM nodes of the virtual link (i,j) is equal to the total number of outgoing wavelengths of that link.

16) Number of wavelength channels

$$\sum_{i \in N} \sum_{j \in N: i \neq j} W_{m,n}^{i,j} \leq w \cdot f_{m,n} \qquad (5.42)$$

$$\forall m \in N, \forall n \in NN_m$$

17) Total number of wavelength channels

$$W_{m,n} = \sum_{i \in N} \sum_{j \in N: i \neq j} W_{m,n}^{i,j} \qquad (5.43)$$

$$\forall m \in N, \forall n \in NN_m$$

18) Number of aggregation ports

$$\Lambda_i = \left(\sum_{j \in L \cap TN_i} \left(\sum_{p \in H} \sum_{q \in H, q \neq p} \lambda T_{i,j}^{p,q} + \sum_{h \in H} \sum_{r \in R} \lambda R_{i,j}^{h,r} \right) \right) \Big/ B \qquad (5.44)$$

$$\forall i \in N$$

Constraint (5.42) and (5.43) are the constraints of the physical link (m,n). Constraint (5.42) ensures that the total number of wavelength channels in the logical link (i,j) that traverse the physical link (m,n) does not exceed the fiber capacity. Constraint (5.43) determines the number of wavelength channels in the physical link and ensures that it is equals to the total number of wavelength channels in the virtual link traversing that physical link. Constraint (5.44) determines the required number of aggregation ports in each IP over WDM router.

5.4 MILP model setup and results

Five IP over WDM nodes are considered constituting the optical-based backbone network of the architecture. The distribution and topology of the IP over WDM nodes have been built upon the NSFNET network described in [235, 260]. Each IP over WDM node in turn is attached to two GPONs with one OLT and two ONUs for each GPON. Accordingly, the network topology has 10 OLTs and 20 ONUs. In addition, each ONU is connected to one RRH node as shown in Figure 5.8. Two GPONs for each IP over WDM node are enough to investigate the VM response for demands and power savings. To finalise the portrait of the network topology, we have concentrated on the distribution of the hosting nodes and the way in which they are connected to each other and for this reason the GPON splitters are not shown.

As alluded to earlier, two types of VMs have been considered: BBUVM, which realise the functions of the BBU, and CNVM to achieve the functions of the mobile core network. The amount of workload needed for BBUVMs is calculated in GOPS according to equation (5.12) [259] and based on the calculated workload, the hosting server CPU utilisation due to hosting BBUVMs is determined. On the other hand, the total workload needed for CNVMs is calculated based on the number BBUVMs group in each hosting node since we have allocated one CNVM for each group of BBUVMs in one hosting node. A single VM consumes around 18W [261] and by knowing the hosting server maximum power consumption (365W), idle power (112) and the maximum workload (368 GOPS), ΨC_h can calculated for a single VM. Therefore $\Psi C_h =$ corresponds is $(18 \times 368)/(365 - 112) = 26$ GOPS.

We have investigated the effect of the intra-traffic between the CNVMs by considering a range of intra-traffic relative to the total network traffic (0%, 1%, 5%,10%, and 16% of the total traffic) flows from CNVMs. Moving toward access network, each RRH node is considered to serve a small cell that operates on 10 MHz bandwidth and with a maximum capacity of 10 users. Each user in the small cell is allocated 5 physical resources blocks (PRB) as the users are assumed to request the same task from the network. Accordingly, the total downlink traffic to the RRH node depends on the total number of active users in the small cell. The input parameters to the developed MILP model are listed in Table 5.6. We have considered 17 time slots over all the day from 0 hour to 24 hour in steps of 1.5 hours using the average number of users daily profile shown in Figure 5.9.The MILP results are compared with the case where there is no NFV deployment. In "no virtualisation" scenario, the BBU is located close to the RRH where they are attached to each other, whilst the integrated platform ASR5000 is deployed to realise mobile core network functionalities and it is connected directly to the IP over WDM network. The ASR5000 maximum power consumption, idle power, and maximum capacity are 5760 (W), 800 (W), and 320 (Gbps) respectively [262], whilst the BBU maximum power consumption, idle power, and maximum capacity are 531 (W), 51 (W), 9.8 (Gbps) respectively [263].

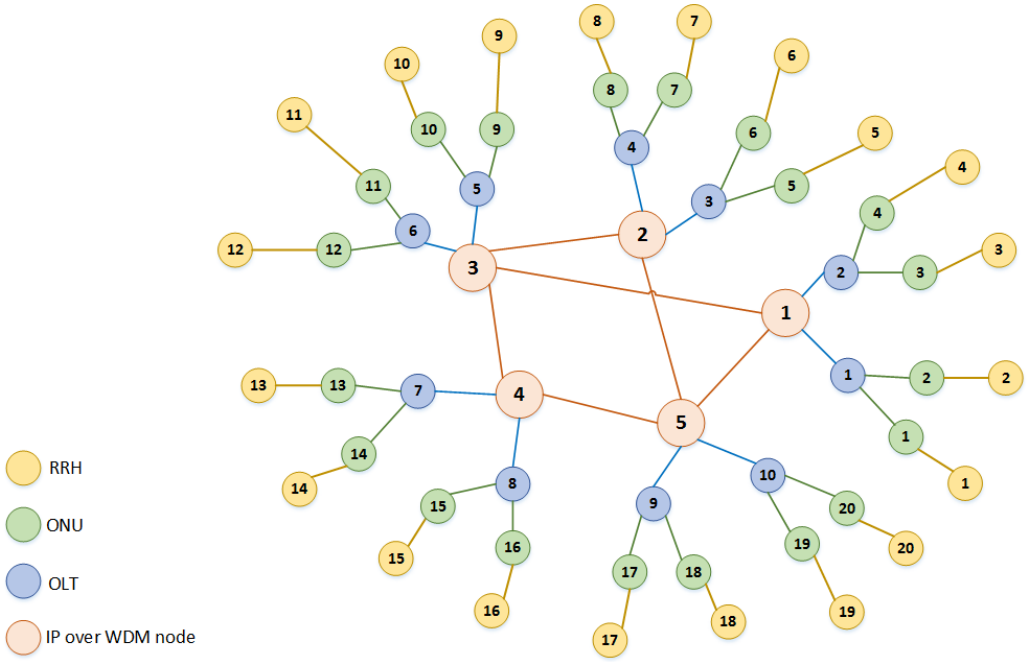


Figure 5.8 Tested Network topology

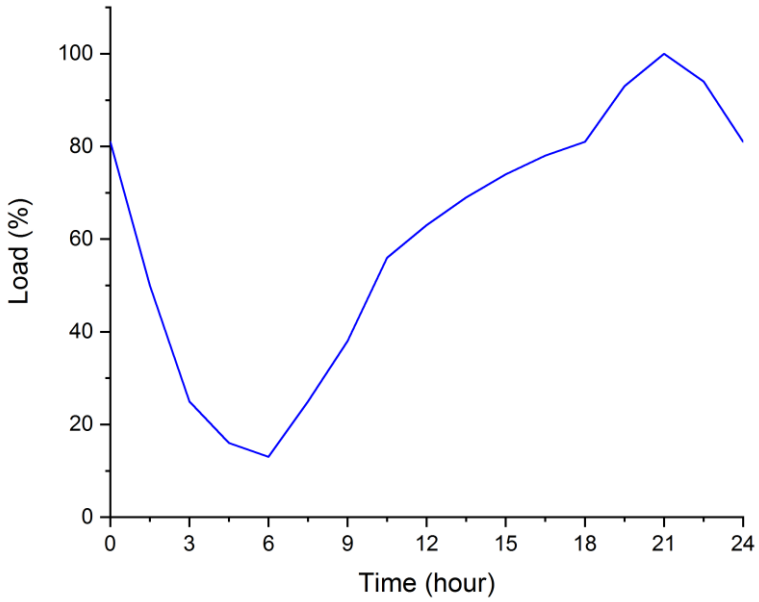


Figure 5.9 Illustration of average number of users daily profile [264]

Table 5.6 MILP model input parameters

Line coding rate for 8B/10B line coding (l)	10/8 (bit / sample)
Number of MIMO layers (y)	2
Number of bits used in QAM modulation for 64 QAM modulation (q)	6 (bits)
Number of antennas in a cell (a)	2
Maximum fronthaul (CPRI) data rate for CPRI line rate option 7 (cp)	9.8304 (Gbps) [256]
Maximum baseband processing workload needed for fully loaded RRH (ΨX) given by: $30 \cdot a + 10 \cdot a^2 + 20 \cdot q \cdot l \cdot y$	400 (GOPS)
Server CPU maximum workload (ΨS)	368 (GOPS) [265]
Workload needed for hosting one CNVM (ΨC)	26.17 (GOPS)
Number of active users in a small cell (ρ_r)	Uniformly distributed (1-10 users)
Maximum number of users per cell (n)	10 (users)
Number of physical resource blocks per user (pb)	5 (PRB)
The ratio of the backhaul to the front haul traffic (α)	0.1344 (unitless)
ONU maximum power consumption (ΩU)	15 (W) [233]
OLT maximum power consumption (ΩL)	1940 (W) [231]
OLT idle power (ΩLd)	60 (W) [231]
OLT maximum capacity (CL)	8600 (Gbps) [231]
ONU maximum capacity (CU)	10 (Gbps) [233]
RRH node power consumption (ΩR_x)	1140 (W) [234]
Hosting server maximum power consumption (ΩS)	365 (W) [266]
Hosting server idle power consumption (ΩSd)	112 (W) [266]
Capacity IP over WDM wavelength channel (B)	40 (Gbps) [236]
Number of wavelengths per fibre in IP over WDM (w)	32 [236]
Transponder power consumption (ΩT)	167 (W) [237]
Router port power consumption (ΩRP)	825 (W) [172]
Regenerator power consumption (ΩG)	334 (W) [172]

EDFA power consumption (ΩE)	55 (W) [172]
Maximum span distance between EDFAs (S)	80 (km) [236]

The results in Figure 5.10 show the total power consumption of the no virtualisation scenario as well as the virtualisation scenario under different CNVMs intra-traffic for different time slots in a day. Figure 5.11 shows the total power consumption of the same scenarios versus the total number of active users in the network. The virtualisation model has resulted in less power consumption compared to the no virtualisation scenario as it optimises the processing locations of the downlink traffic through optimum placement and consolidation of VMs. It is clearly seen in Figure 5.11 that the no virtualisation case has a rapid increase in the total power consumption as the total number of active users in the network increases. In contrast, all virtualisation cases (different CNVMs intra-traffic) try to cope with the increase in the total number of active users by optimising the VMs locations and consolidating the resources in the hosting servers.

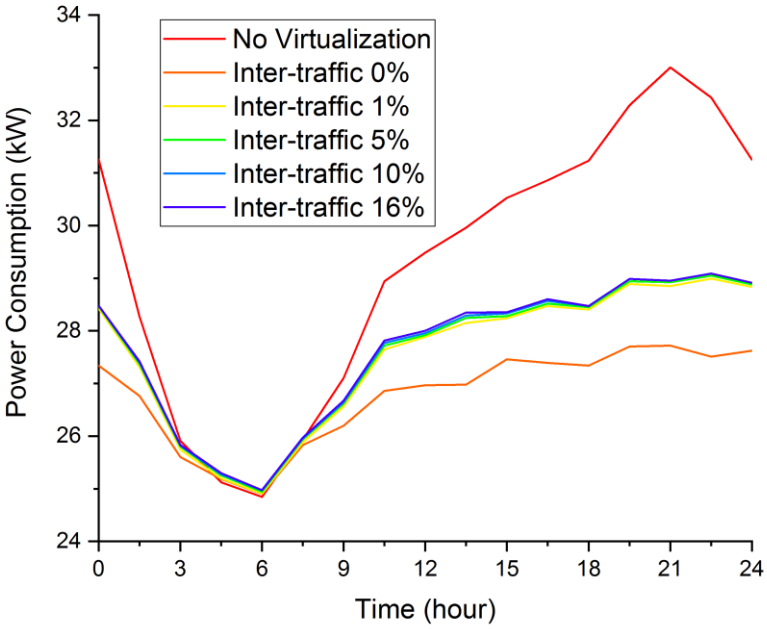


Figure 5.10 Total power consumption without and with virtualisation under different CNVMs intra-traffic at different time slots of one day**.**

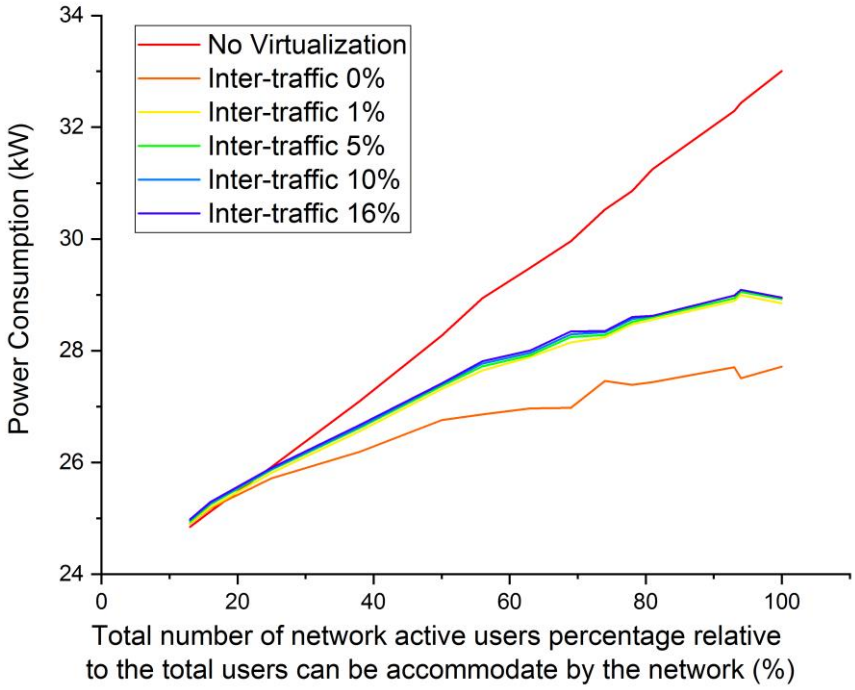


Figure 5.11 Total power consumption without and with virtualisation under different CNVMs intra-traffic vs total active users in the network

Figure 5.12 compares the virtualisation power saving under different CNVMs intra-traffic for one day whilst Figure 5.13 shows the virtualisation power saving under different CNVMs intra-traffic versus total number of active users. Compared to other virtualisation cases, virtualisation with 0% CNVMs intra-traffic (no intra-traffic) has saved a maximum of 16% (average 8%) of the total power consumption. This is because there is no power consumption produced by the CNVMs intra-traffic as this traffic is zero. The second highest saving was recoded for virtualisation with 1% CNVMs intra-traffic which is 12.6% (average 5.7%) whilst virtualisation with 16% CNVMs intra-traffic produced a saving of 12.3% (average 5.3%) which is the lowest saving recoded due to the power consumption induced by intra-traffic as shown in Figure 5.14. Virtualisation in the presence of CNVMs intra-traffic resulted in close values of total power consumption (and power saving) for all values of

CNVMs intra-traffic greater than zero. The main reason behind this is that the CNVMs intra-traffic produces relatively small amount of power consumption compared to the power consumption induced by the fronthaul traffic and hosting servers. As the intra-traffic increases, the MILP model tends to eliminate its effect by consolidating CNVMs in one place.

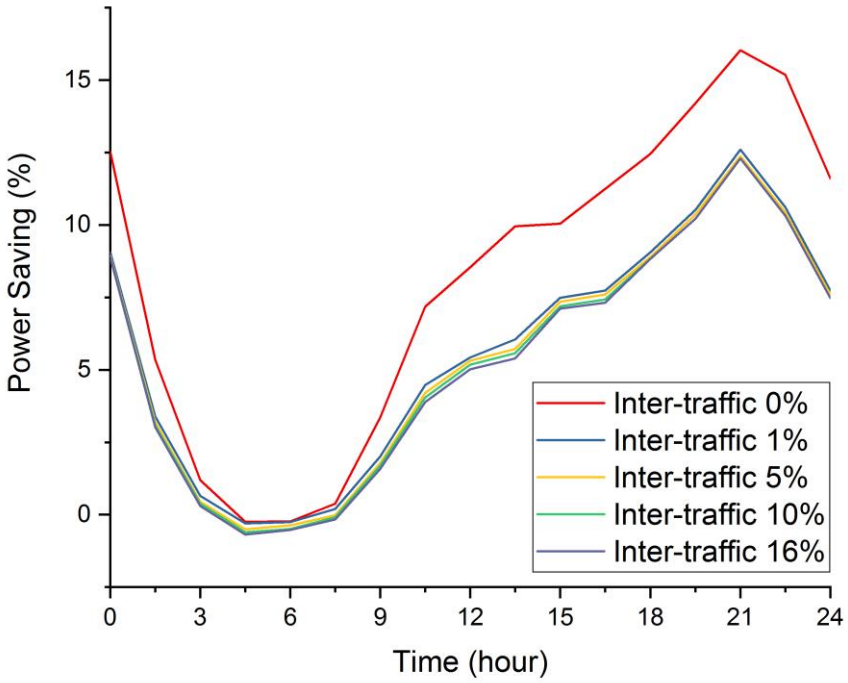


Figure 5.12 Power saving comparison of virtualisation under different CNVMs intra-traffic for one day

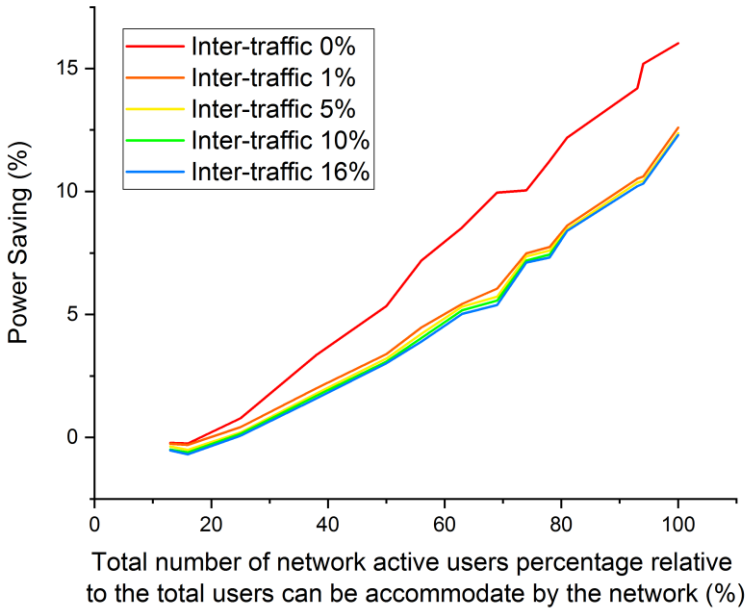


Figure 5.13 Power saving of virtualisation under different CNVMs intra-traffic versus total number of active users

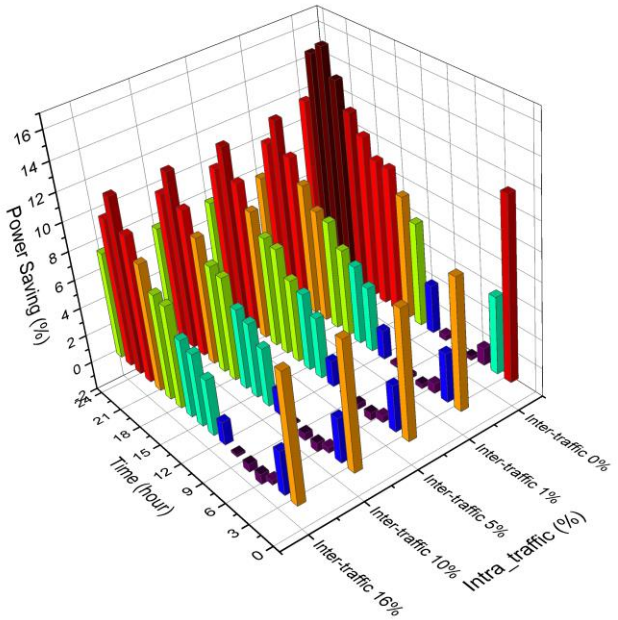


Figure 5.14 3-Dimensional presentation of the total power saving for virtualisation under different CNVMs intra-traffic in one day time

Although virtualisation has saved a maximum of 16% (with no intra-traffic) and 12.3% (with 16% intra-traffic) of the total power consumption, it cannot provide such level of power saving over all the entire day. As the number of active users varies with the time of day (as in Figure 5.9), the power saving achieved by virtualisation varies accordingly. The results in Figure 5.13 and Figure 5.14 show that a high-power saving is achieved at high number of active user (during the day rush hours), whilst a very low power saving (around zero) is recoded when the total number of active users is less than 20% (around 4 to 8 AM). At small number of active users, the MILP model tends to consolidate all the VMs in the IP over WDM network and this causes high traffic specially in the fronthaul which result in high power consumption.

Figure 5.15 and Figure 5.16 show the VMs consolidation and distribution over the network with low number of active users (13%) under CNVMs of 0% and 16% respectively. At low number of active users and 0% intra-traffic, the MILP model consolidates the VMs at the IP over WDM network. Since the total number of active users is low, the fronthaul traffic is relatively low and consequently the power consumption induced by the fronthaul traffic is low compared to the hosting power consumption (servers power). For this reason, the MILP model tends to pack BBUVMs in the IP over WDM network as much as possible to reduce the induced power due to hosting servers. Also, the MILP model tends to host CNVMs close to the BBUVMs as the intra-traffic between CNVMs is zero. Once the intra-traffic is greater than zero, the MILP model consolidates the CNVMs at one location as in Figure 5.16.

Figure 5.17 and Figure 5.18 show the VMs consolidation and distribution over the network with high number of active users (around 100%) under 0% and 16% CNVMs intra-traffic. When the number of active users is high, the amount of fronthaul traffic is high, for that reason. The MILP models tends to distribute the BBUVMs at the closest centralised location to the users which is OLTs, whilst the CNVMs intra-traffic has no effect on the distribution of BBUVMs. Hosting BBUVMs in OLTs when the number of users is high, ensures shorter paths for this traffic than hosting BBUVMs in the IP over WDM network and consequently, the power induced by this traffic is less. For CNVMs, the MILP models tends to distribute them close to the BBUVM when there is no intra-traffic between them, and this is clearly seen in Figure 5.17. In contrast, when the intra-traffic between CNVMs is greater than zero, the MILP model tends to centralise the location of CNVMs in the IP over WDM network to reduce the power consumption induced by the intra-traffic and the power of the hosting servers as shown in Figure 5.18.

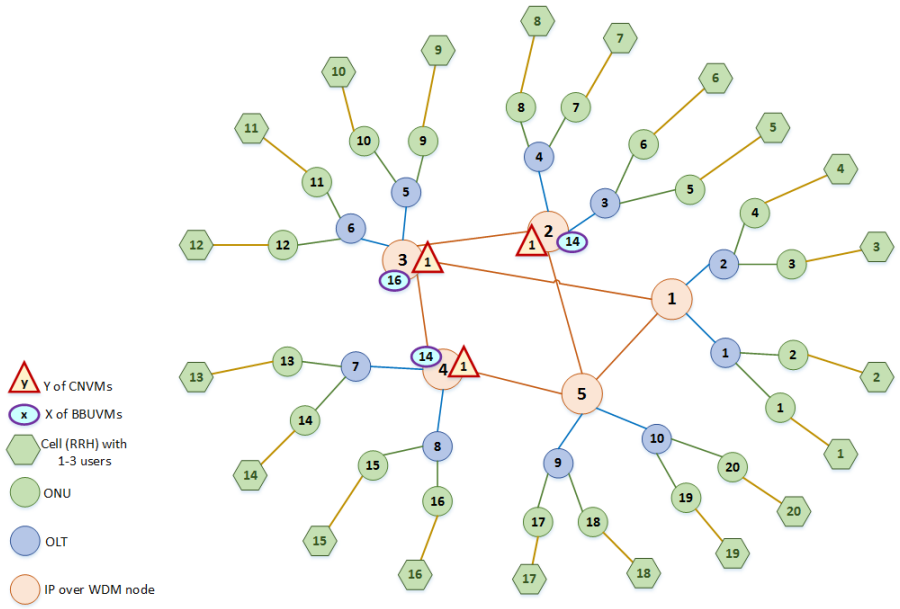


Figure 5.15 Virtual machines distribution over network under active users 13% of the total capacity and 0% CNVMs intra-traffic

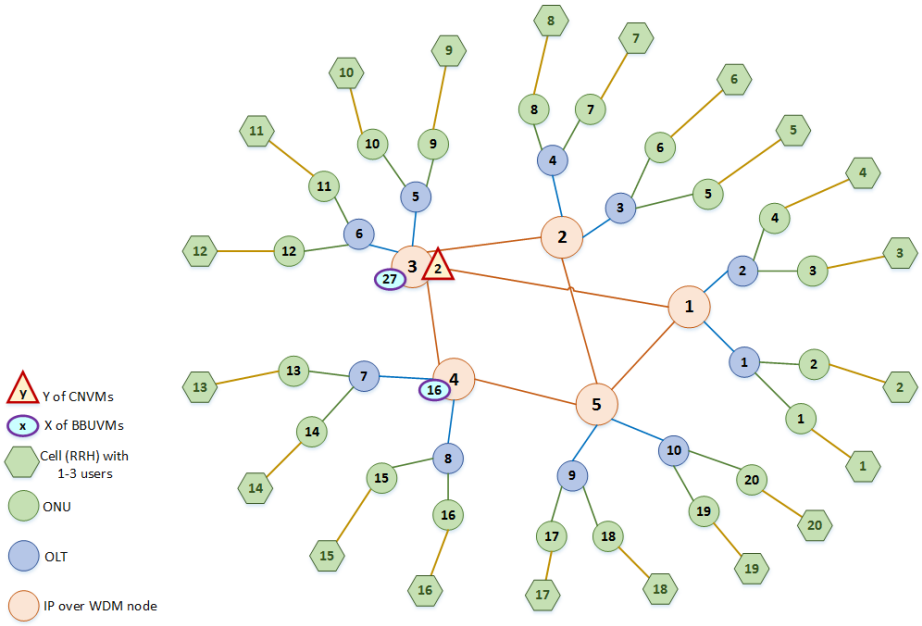


Figure 5.16 Virtual machines distribution over network under active users 13% of the total capacity and 16% CNVMs intra-traffic

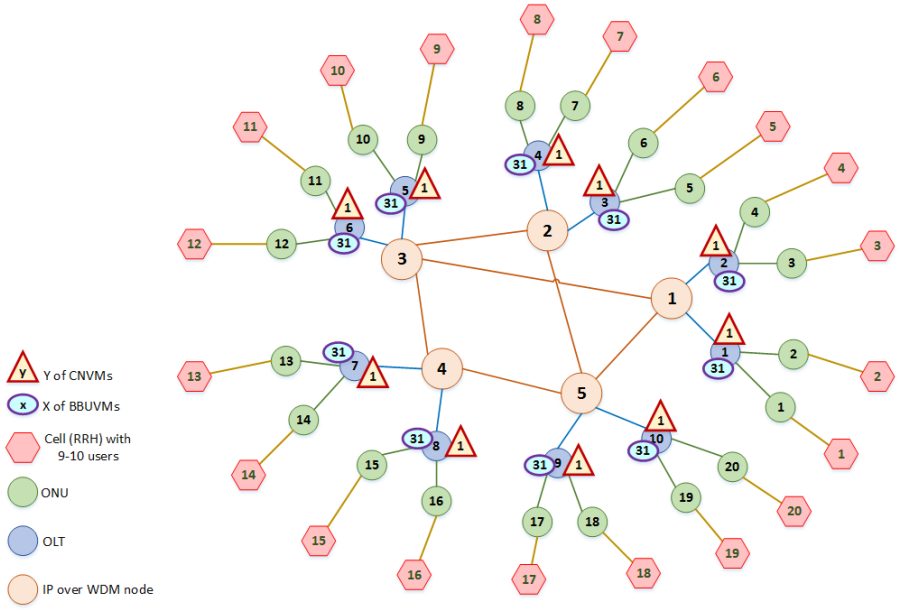


Figure 5.17 Virtual machines distribution over network under active users 100% of the total capacity and 0% CNVMs intra-traffic

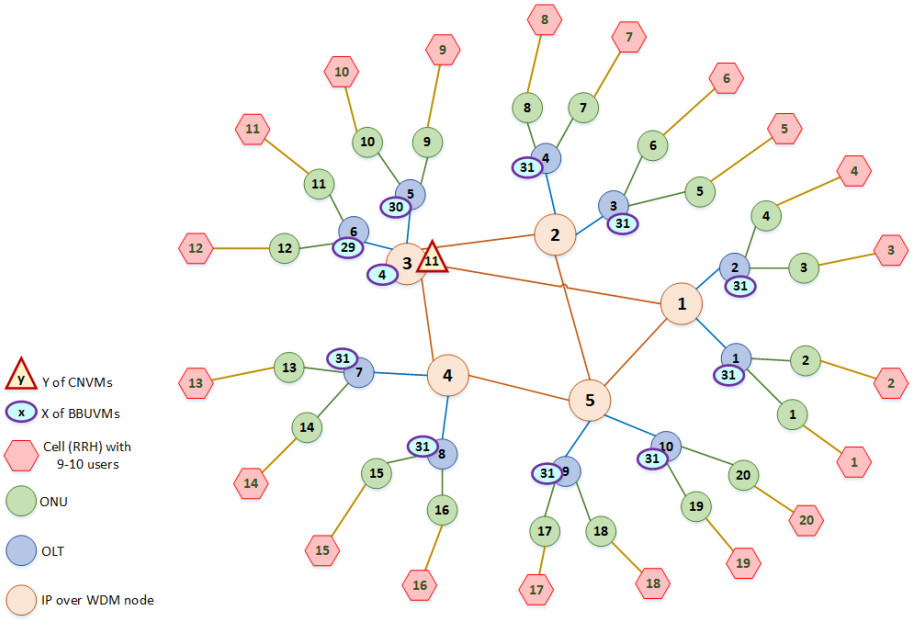


Figure 5.18 Virtual machines distribution over network under active users 100% of the total capacity and 16% CNVMs intra-traffic

5.5 Real-time heuristic models implementation

This section introduces two heuristic approaches for real-time implementation of the MILP model. The first heuristic approach considers the case where no inter-traffic flows between CNVMs whilst the second heuristic approach considers the inter-traffic between CNVMs.

5.5.1 Energy Efficient NFV with no CNVMs inter-traffic (EENFVnoITr) heuristic model

The EENFVnoITr heuristic provides real-time implementation of the MILP model without CNVMs inter-traffic. The pseudocode of the heuristic is shown in Algorithm A.1. The network is modelled by sets of network elements *NE*, and links *L*. The heuristic obtains the network topology $G = (NE, L)$ and the physical topology of the IP over WDM network $G_p = (N, L_p)$, where N is the set of IP over WDM nodes and L_p is the set of physical links. The total download request (fronthaul traffic) of each RRH node is calculated based on the total number of active users in each cell (RRH). The heuristic determines the amount of baseband workload needed to process each RRH download request. According to the baseband workload for each requested download traffic and the available capacity of the hosting VM server, the EENFVnoITr heuristic model chooses the closest place to accommodate BBUVM in such a way that it serves as much RRH requests as possible. The EENFVnoITr heuristic model may host a BBUVM in an OLT node if it has enough processing capacity to serve all the requests from the closest RRH nodes. In this way, the heuristic exploits bin packing techniques to reduce the processing power

consumption. The amount of fronthaul traffic delivered by each BBUVM determines the backhaul traffic flows from each CNVMs toward BBUVMs. The EENFVnoITr heuristic determines the total amount of backhaul traffic that may flow from each IP over WDM node and sorts them in a descending order. The nodes in the top of the sorted list of IP over WDM nodes represent highly recommended nodes to host CNVMs. In such a scenario, the EENFVnoITr heuristic ensures less of backhaul traffic flows in the IP over WDM network. The EENFVnoITr heuristic uses the sorted list to accommodate CNVMs. Once the VMs are distributed and the logical traffic is routed, the EENFVnoITr heuristic obtains the physical graph G_p = (N, L_p) and determines the traffic in each network segment. The IP over WDM network configuration such as the number of fibres, router ports, and the number of EDFA is determined and the total power consumption is evaluated. The heuristic reduces the number of CNVMs candidate locations by one, re-configures the IP over WDM network, and re-evaluates the power consumption to determine the best number and location of CNVMs for minimum power consumption.

5.5.2 Energy Efficient NFV with CNVMs inter-traffic (EENFVwithITr) heuristic

This section describes the energy efficient NFV with CNVMs inter-traffic heuristic (EENFVwithITr). The EENFVwithITr heuristic extends the EENFVnoITr heuristic to provide real-time implementation of the MILP model where the CNVMs are considered. The pseudocode of the heuristic is shown in Algorithm A.2. It uses

the same approach used by EENFVnoITr but it evaluates the CNVMs inter-traffic after the locations of CNVMs are determined.

5.5.3 EENFVnoITr and EENFVwithITr heuristic models results

In order to verify the results of the proposed MILP model, the network topology in Figure 5.8 used for the MILP model is also used to evaluate the heuristics. All the considerations and parameters considered in the MILP model such as the wireless bandwidth, number of resources blocks per user, and the parameters in Table 5.6 are considered in the developments of both EENFVnoITr and EENFVwithITr heuristics. The number of users allocated to each cell in the heuristics is the same as in the MILP model to ensure the requested traffic by each RRH node is the same in all models. Figure 5.19 compares the total power consumption of MILP without CNVMs inter-traffic and EENFVnoITr model at different times of the day when the CNVMs inter-traffic is not considered. It is clearly seen that there is a small difference in the total power consumption of the two models and it varies over the day according to the total number of active users. The total power consumption of the MILP model is less than the EENFVnoITr heuristic with a maximum of 9% (average 5%) drop in the total power consumption. This is mainly caused by the distribution of CNVMs in the EENFVnoITr heuristic. As there is no traffic flowing between CNVMs, the EENFVnoITr accommodates them close to the BBUVMs wherever the VM servers have enough capacity. To accommodate the CNVMs, the heuristic sequentially examines the capacity of the VM servers in the OLT nodes that are close to the BBUVMs before investigating other servers in the IP over

WDM networks. As the distance and capacity requirements of the VM servers are met, the heuristic accommodate a CNVM in the server. This case results in high EENFVnoITr VM server power consumption compared with MILP model. This is clearly seen in Figure 5.20 where the VM servers power consumption of MILP and the EENFVnoITr heuristic are compared. The total network power consumption of both EENFVnoITr heuristic and MILP model are the same for most of the time of the day. Figure 5.21 shows the network power consumption of MILP model compared with EENFVnoITr heuristic. It shows that there is a small difference in the network power consumption between the two models during the time of the day when the total number of active users is low. This is driven by the approach of the MILP model where it tends to accommodate the CNVMs at the IP over WDM nodes rather than OLT at the time of the day where the total number of users is low. In contrast, the heuristic tends to accommodate the CNVMs wherever the VM server is close to the BBUVMs and it has enough capacity.

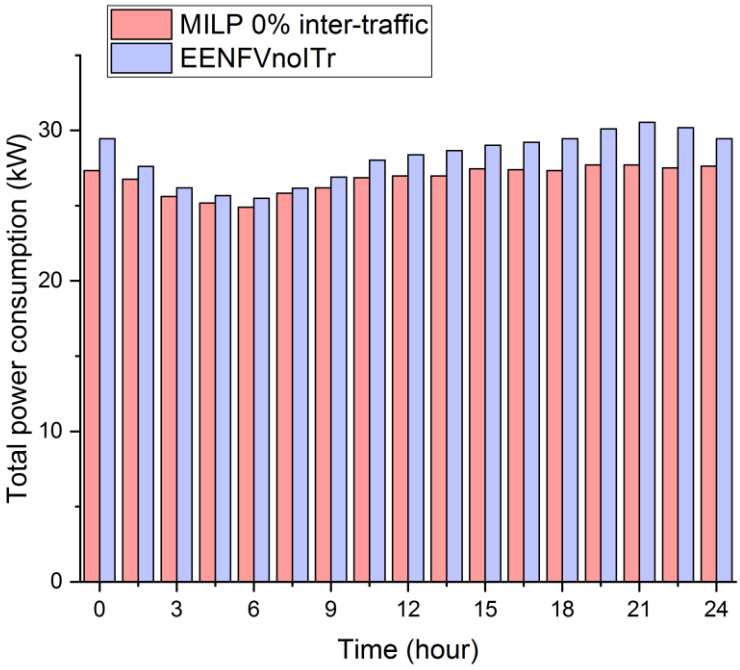


Figure 5.19 Total power consumption of MILP without CNVMS inter-traffic compared with EENFVnoITr heuristic model

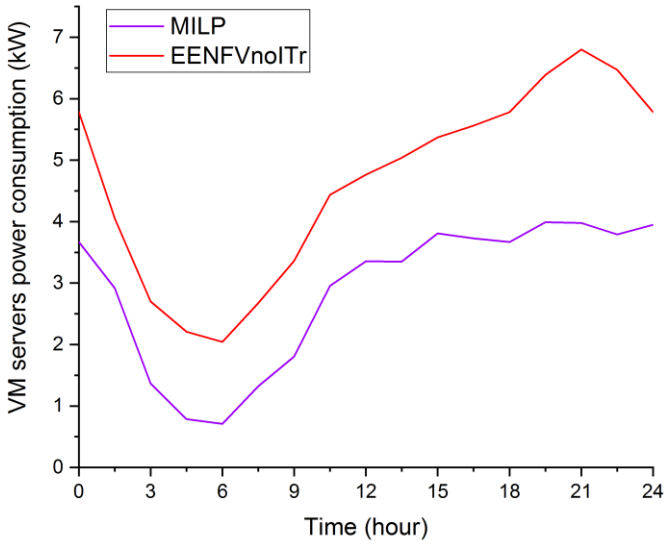


Figure 5.20 VM servers power consumption of MILP model compared with EENFVnoITr model

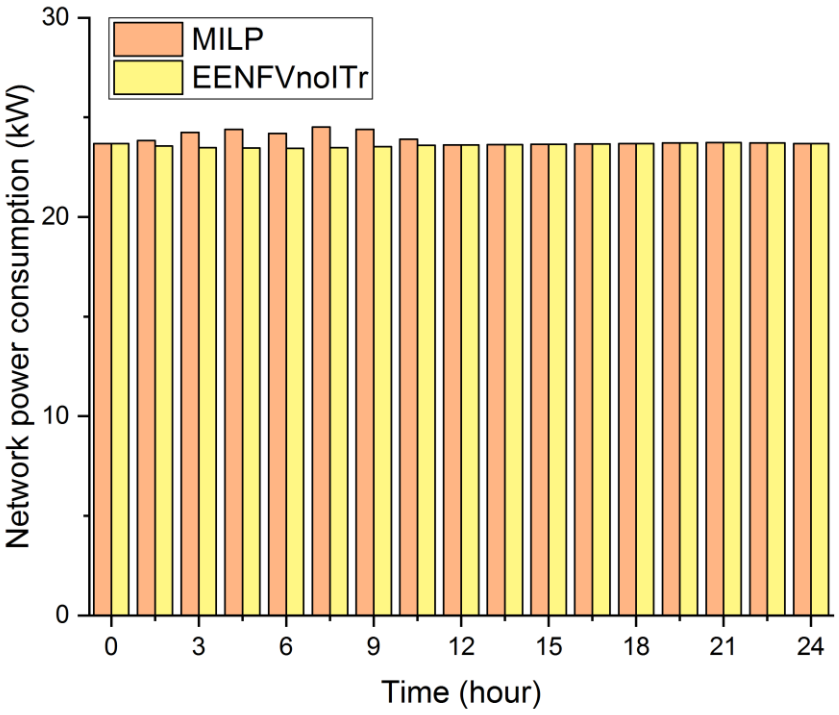


Figure 5.21 Network power consumption of MILP model compared with EENFVnoITr model

Figure 5.22 compares the total power consumption of EENFVwithITr with the MILP model when the CNVMs inter-traffic is 16% of the total backhaul traffic. It is clearly seen that there is a small difference in the total power consumption of the two models and this varies over the day according to the total number of active users. The total power consumption of the MILP model is less than the EENFVnoITr model with a maximum drop of 9.5% (average 5%) in the total power consumption. This is mainly driven by the distribution of both CNVMs and BBUVM over the network nodes. The MILP model tends to accommodate BBUVMs and CNVMs at the IP over WDM network during times of the day when there is a small number of active users. This causes more traffic from BBUVMs and

CNVMs flow in the IP over WDM network which eventually increases the IP over WDM network power consumption as shown in Figure 5.23 which compares the IP over WDM network power consumption of both MILP model and EENFVwithITr when CNVMs inter-traffic is considered 16% of the total backhaul traffic. In contrast, the IP over WDM network power consumption of EENFVwithITr varies according to the total number of active users during the day. The sequential examination by EENFVwithITr of VM servers, their location, and available capacity increases the processing distribution of VMs in the network which leads to a high VM servers power consumption compared with the MILP model as shown in Figure 5.24 which compares the VM servers power consumption of the MILP model with EENFVwithITr heuristic during different times of the day.

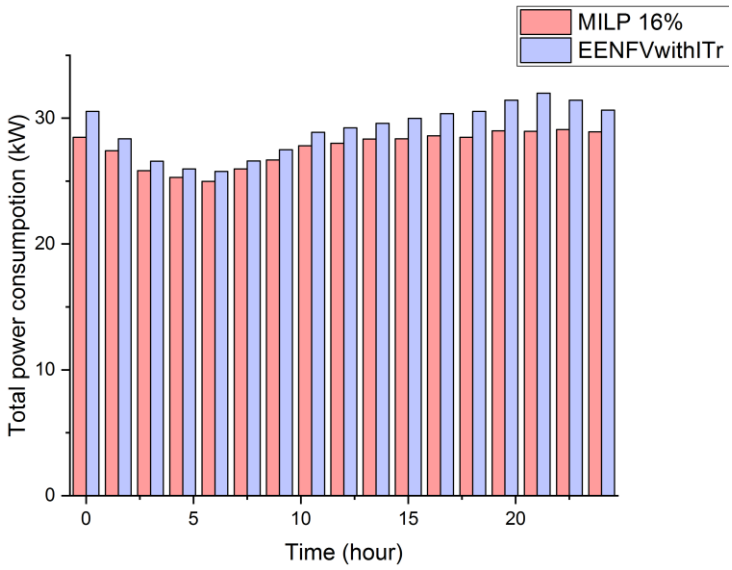


Figure 5.22 Total power consumption of MILP model compared with EENFVwithITr heuristic model at CNVMs inter-traffic 16% of the total backhaul traffic

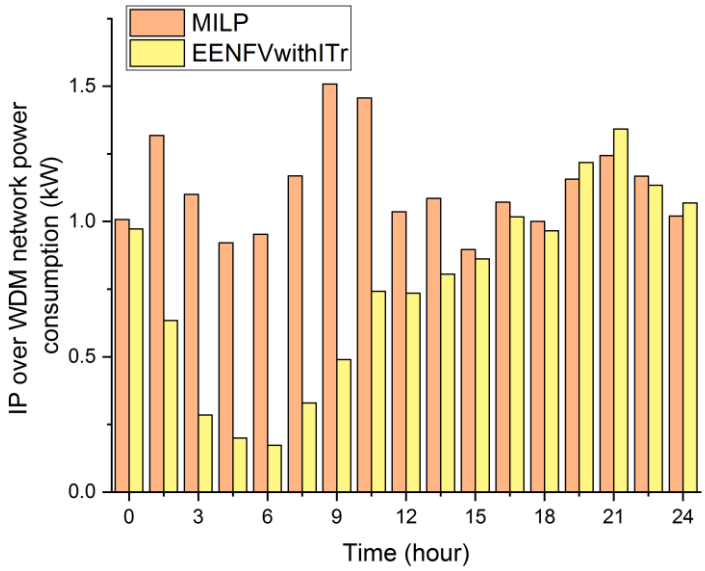


Figure 5.23 IP over WDM network power consumption of MILP model compared with EENFVwithITr heuristic model at CNVMs inter-traffic 16% of the total backhaul traffic

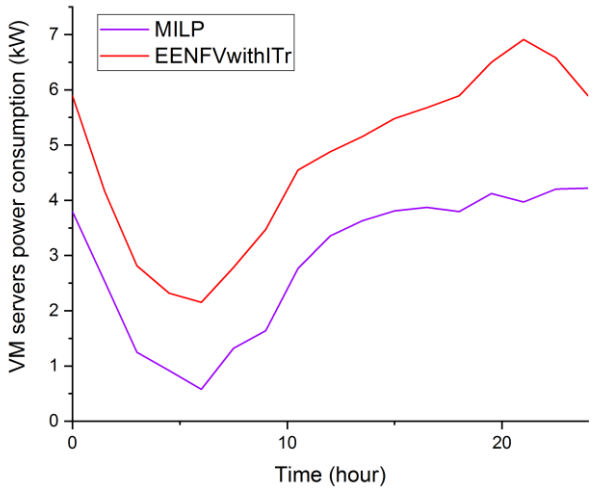


Figure 5.24 VM servers power consumption of MILP model compared with EENFVwithITr heuristic model at CNVMs inter-traffic 16% of the total backhaul traffic

5.6 Summary

This chapter extended the work in the previous chapter by including the total number of active users in the network during the day, the backhaul and fronthaul configuration and the required workload for baseband processing. A MILP optimisation model was developed with the objective of minimising the total power consumption by optimising the VMs locations and VM servers' utilisation. The MILP model results were investigated under the impact of CNVMs traffic and variation of total number of active users during different times of the day. The MILP model results show that virtualisation can save up to of 16% (average 8%) of the total power consumption during rush hours of the day whilst it is better to run the network without NFV in case a low number of users are active (around 4 am to 8 am). The results reveal how the total number of active users affects the BBUVMs distribution whilst CNVMs distribution is affected mainly by the intra-traffic between them.

For results validation and real-time implementation, this chapter introduced two heuristics. The first heuristic is the Energy Efficient NFV without CNVMs inter-traffic (EENFVnoITr) which was developed to mimic the behaviour of the MILP model when the CNVMs inter-traffic is not considered. The results of the EENFVnoITr heuristic were compared with the MILP model and they show that the total power consumption of the EENFVnoITr model is higher than the MILP model by a maximum of 9% (average 5%). The second heuristic is the Energy Efficient NFV with CNVMs inter-traffic (EENFVwithITr) which was developed to mimic the MILP behaviour when the CNVMs inter-traffic is considered. The results of the

EENFVwithITr heuristic were compared with the MILP model and they show that the total power consumption of the EENFVwithITr model is higher than the MILP model by a maximum of 9.5% (average 5%).

Chapter 6

Energy-Efficient Content Caching with Fixed and Variable Size Cache in 5G Networks

6.1 Introduction

The 5G era will witness an enormous number of connected IoT devices which will result in huge energy consumption [267]. In addition, Video-on-Demand (VoD), multimedia streaming, mobile Internet Protocol Television (IPTV) and other bulky multimedia services are the most power consuming applications in mobile smartphones and handsets [268].

The paradigm of storing the most popular content at the edge of the network is one of the effective paradigms to reduce the power consumption of video services. This chapter evaluates the power consumption of delivering video over optical-based architecture for 5G networks. It investigates the power saving introduced by optimising the cache nodes sizes and locations in the core and access network. A mixed integer linear programming (MILP) model was developed to optimise the location of a fixed-size cache at each node at different number of users to minimise the total power consumption. The MILP model was then extended to consider variable size caches. To achieve maximum power saving, the model finds the optimum cache size at each node at different times of the day for different number of users.

6.2 Energy-efficient cache contents in 5G networks architectures and MILP models

Caching contents close to the end users results in shorter paths to the content and lower traffic induced power consumption. However, this strategy requires more equipment to be added to the network which eventually increases the equipment induced power consumption. Therefore, the locations and sizes of the caches are functions of the two above drivers. This evaluation minimises the total power consumption including the video services by optimising the cache location and size at nodes.

6.2.1 Network architecture

Figure 6.1 illustrates a contents cache service for a video on demand (VoD) over optical-based network for 5G networks. The same architecture introduced in chapters 4 and 5 is used in this chapter, but with the deployment of cache-based service for a VoD and no virtualisation for any of the mobile function previously mentioned. As illustrated in Figure 6.1, the proposed network consists of three layers; IP over WDM network, wired optical access network represented by a passive optical network (PON), and mobile radio access network (RAN) represented by a group of eNodedBs. Both the video server and the mobile core node (ASR5000) locations are restricted to one of the IP over WDM nodes, whilst the content caches can be placed at any node (ONU, OLT, and IP over WDM node) in the network. Accordingly, the traffic to any eNodeB is composed of three traffic components: traffic from mobile core node (ASR5000), traffic from video streaming server, and traffic from cache nodes. According to [1] around 80% of the total

consumer Internet traffic will be IP video traffic by 2021. Therefore, the traffic from video streaming servers together with the traffic from cache nodes are considered as 80% of the total download traffic toward eNodeBs, whilst the traffic from the mobile core node (ASR5000) makes the remaining 20% of the total traffic. For clarity purposes and to avoid any clutter, the traffic components toward only three eNodeB nodes are shown in Figure 6.1.

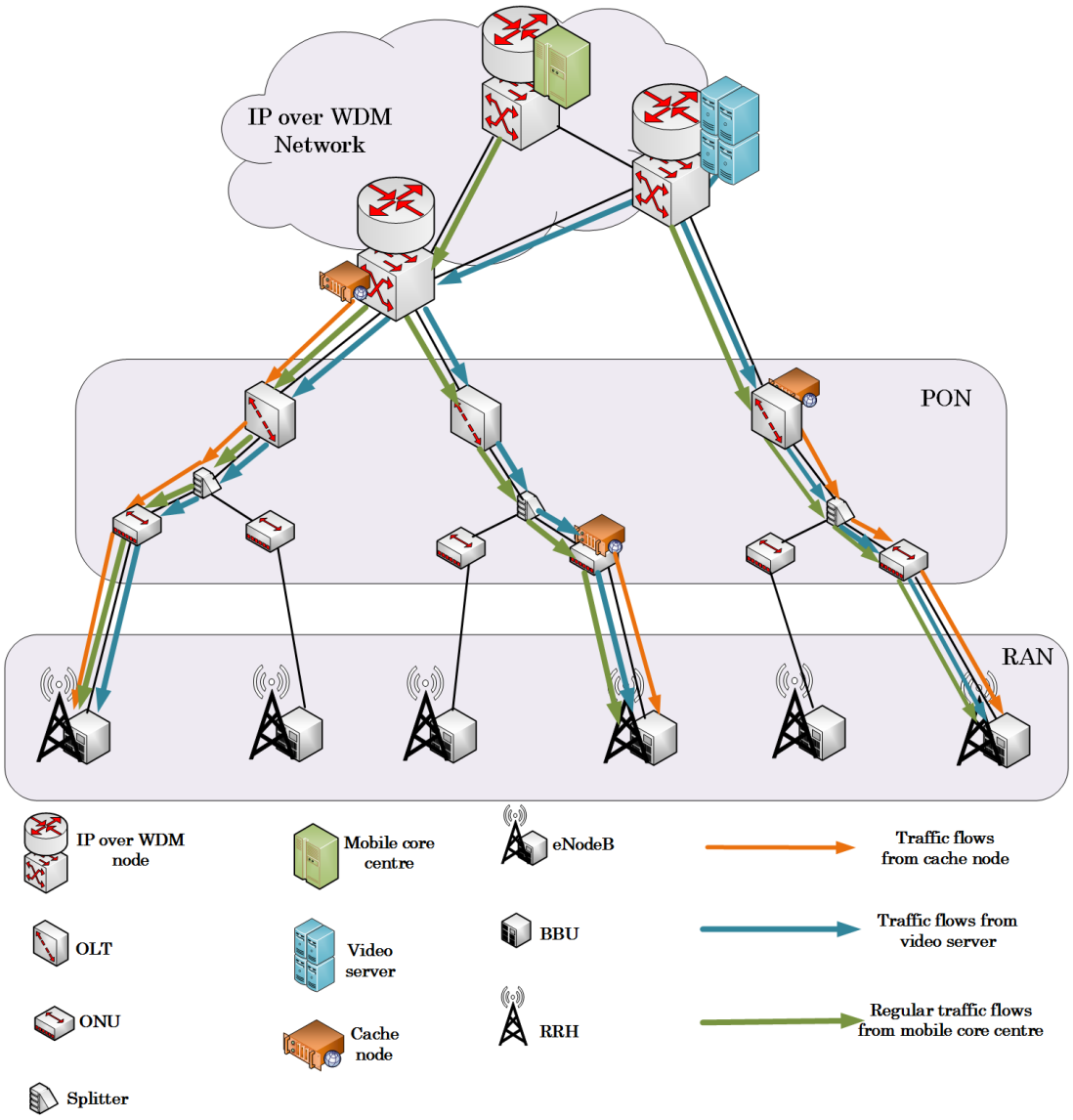


Figure 6.1 Contents caching in 5G network architecture

6.2.2 Energy-efficient fixed-size cache MILP model

This model considers cache nodes with fixed-size cache, whilst their locations are optimised to minimise the total power consumption. It evaluates the impact of a range of fixed-size caches and the locations of the cache nodes on the power consumption. The influence of each cache size in the range is evaluated separately for different number of users for a whole day. The model declares a number of indices, parameters, and variables. these are listed in Table 6.1,Table 6.2, and Table 6.3 respectively.

Table 6.1 Indices of the fixed-size cache MILP model

Indices	Comment
x, y	Indices of any two nodes in the developed model
m, n	Indices of any two nodes in the physical layer of the IP over WDM network
i, j	Indices of any two nodes in the IP layer of the IP over WDM network.
e	Index of eNodeB
h	Indices of the node where the cache node can be hosted

Table 6.2 Parameters of the fixed-size cache MILP model

Parameters	Comment
E	Set of eNodeB nodes
U	Set of ONU nodes
L	Set of OLT nodes
N	Set of IP over WDM nodes

T	Set of all nodes (eNodeB, ONU, OLT, and IP over WDM nodes)
NN_m	Set of neighbours of node m in the IP over WDM network, $\forall\ m\ \in N$
TN_x	Set of neighbours of node x, $\forall\ x\ \in T$
H	Set of hosting nodes (ONU, OLT, and IP over WDM nodes) where the cache node can be attached
K	Set of Linearization coefficients
λG_e	Regular traffic between RRH and BBU of eNodeB node e
λV_e	Video streaming traffic between RRH and BBU of eNB node e
cp	CPRI link data rate
ρ_e	Number of mobile users connected to the eNodeB node e
n	Maximum number of physical resources blocks for cell (e)
pb	Physical resources block per user
λR_e	Download traffic requested by eNodeB e calculated as: $[(pb/n) \cdot cp \cdot \rho_e]$, where $e\ \in E$
CZ_h	Cache size at node h (range)
a_k, b_k	Linearization coefficients (unitless)
β	Large number
α	The ratio of the backhaul to the fronthaul traffic (unitless)
B	Capacity of the WDM wavelength channel (Gbps)
w	Number of wavelengths per fiber
ΩT	Transponder power consumption
ΩRP	Router power consumption per port
ΩG	Regenerator power consumption
ΩE	EDFA power consumption

$NG_{m,n}$	Number of regenerators in the optical link (m, n)
S	Maximum span distance between EDFAs (km)
$D_{m,n}$	Distance between node pair (m, n) in the IP over WDM network (km)
$A_{m,n}$	Number of EDFAs between node pair (m, n) calculated as $A_{m,n} = \left((D_{mn}/S) - 1\right) + 2$
ΩU	ONU maximum power consumption
ΩL	OLT maximum power consumption
ΩLd	OLT idle power
CL	OLT maximum capacity
CU	ONU maximum capacity
ΩR_e	Power consumption of the Remote Radio Head (RRH) of the eNodeB node *e*
ΩBd_e	BBU idle power of the eNodeB *e*
ΩB_e	Maximum power consumption of the BBU of the eNodeB *e*
ΩC	Cache node maximum power consumption
CC	Cache node maximum storage capacity
εc	Cache node energy per stored gigabyte $\varepsilon c = \Omega C/CC$
εs	Video streaming server energy per bit

Table 6.3 Variables of the fixed-size cache MILP model

Variables	Comment
$\lambda G_{c,e}$	Regular download traffic from mobile core node at node *c* to the eNodeB *e*.
$\lambda S_{n,e}$	Video streaming traffic between RRH and BBU of eNodeB node *e* from a video server at node *n*
$\lambda C_{h,e}$	Traffic between RRH and BBU of eNodeB node *e* from the

	cache at node *h.*
$\sigma C_{h,e}$	Binary variable, set to 1 if the cache is located at node *h* to serve the eNodeB node *e*
σC_h	Binary variable, set to 1 if the cache is located at node *h*
δ_h	Hit ratio of the cache at node *h*
σA_c	Binary indicator, set to 1 if the mobile core node is located at node c, 0 otherwise
σS_n	Binary variable, set to 1 if a video server is located at node *n*
$\lambda R_{h,e}$	Total download traffic from node h to the eNodeB node e
$\lambda R_{x,y}^{h,r}$	Total download traffic from node h to the eNodeB e that traverses the link between the nodes (x, y) in the network.
$W_{i,j}$	Number of wavelength channels in the virtual link (i, j)
$W_{m,n}^{i,j}$	Number of wavelength channels in the virtual link (i, j) that traverse the physical link (m, n)
$f_{m,n}$	Number of fibres in the physical link (m, n)
$W_{m,n}$	Total number of wavelengths in the physical link (m, n)
Λ_m	Number of aggregation ports of the router at node m

The MILP model defines the total power consumption, composed of:

1. IP over WDM network power consumption which is composed of:
    a) Power consumption of routers aggregation ports calculated as the total number of port multiplied by the power consumption of a single port:

$$\Omega RP \cdot \left(\sum_{m \in N} \Lambda_m \right)$$

    b) Power consumption of high speed ports calculated as the total number of wavelength by the power consumption of a single port:

$$\Omega RP \cdot \left(\sum_{m \in N} \sum_{n \in NN_m} W_{m,n} \right)$$

c) Power consumption of transponders calculated as the total number of wavelengths multiplied by the power consumption of a single transponder:

$$\Omega T \cdot \sum_{m \in N} \sum_{n \in NN_m} W_{m,n}$$

d) Power consumption of EDFAs calculated as the total number of EDFAs multiplied by the total number of fibres multiplied by the power consumption of a single EDFA:

$$\Omega E \cdot \sum_{m \in N} \sum_{n \in NN_m} A_{m,n} \cdot f_{m,n}$$

e) Power consumption of regenerators calculated as the total number of wavelength multiplied by the number of regenerators and the power consumption of a single regenerator:

$$\Omega G \cdot \sum_{m \in N} \sum_{n \in NN_m} NG_{m,n} \cdot W_{m,n}$$

2. Power consumption of eNodeBs calculated as the addition of the eNodeBs idle power to the multiplication of eNodeB energy per bit by the total traffic delivered by eNodeBs

$$\sum_{e \in E} \left(\Omega R_e + \Omega Bd_e + \frac{\Omega B_e - \Omega Bd_e}{cp} \cdot \lambda R_e \right)$$

3. Power consumption of ONUs calculated as the ONU energy per bit multiplied by the total traffic delivered by ONUs:

$$\sum_{x\in \mathrm{U}} \frac{\Omega U}{CU} \cdot \sum_{h\in H} \sum_{e\in E} \sum_{y\in TN_x} \lambda R_{x,y}^{h,e}$$

4. Power consumption of OLTs calculated as the energy per bit of OLT multiplied by the total traffic delivered by the OLTs:

$$\sum_{x\in L} \left[\Omega Ld + \frac{\Omega L - \Omega Ld}{CL} \cdot \left(\sum_{h\in H} \sum_{e\in E} \sum_{y\in TN_x} \lambda R_{x,y}^{h,e} \right) \right]$$

5. Power consumption of video server:

$$\varepsilon s \cdot \sum_{n\in N} \sum_{e\in E} (\alpha \cdot \lambda S_{n,e})$$

6. Power consumption of content caching nodes:

$$\sum_{u\in H} \sigma C_h \cdot \left(\varepsilon c \cdot CC \cdot (C\mathrm{Z}_u/100) \right)$$

The MILP model objective is to minimise the total power consumption given by:

$$\Omega RP \cdot \left(\sum_{m\in N} \Lambda_m + \sum_{m\in N} \sum_{n\in NN_m} W_{m,n} \right) + \left(\Omega T \cdot \sum_{m\in N} \sum_{n\in NN_m} W_{m,n} \right)$$

$$+ \left(\Omega E \cdot \sum_{m\in N} \sum_{n\in NN_m} A_{m,n} \cdot f_{m,n} \right) + \left(\Omega G \cdot \sum_{m\in N} \sum_{n\in NN_m} NG_{m,n} \cdot W_{m,n} \right)$$

$$+ \sum_{e\in E} \left(\Omega R_e + \Omega Bd_e + \frac{\Omega B_e - \Omega Bd_e}{cp} \cdot \lambda R_e \right) + \sum_{x\in \mathrm{U}} \frac{\Omega U}{CU} \cdot \sum_{h\in H} \sum_{e\in E} \sum_{y\in TN_x} \lambda R_{x,y}^{h,e}$$

$$+ \sum_{x\in L} \left[\Omega Ld + \frac{\Omega L - \Omega Ld}{CL} \cdot \left(\sum_{h\in H} \sum_{e\in E} \sum_{y\in TN_x} \lambda R_{x,y}^{h,e} \right) \right]$$

$$+\left(\varepsilon s \cdot \sum_{n\in N}\sum_{e\in E}\left(\alpha \cdot \lambda S_{n,e}\right)\right)+\sum_{h\in H}\sigma C_h.\left(\varepsilon c \cdot CC \cdot (iC\mathrm{Z}_h/100)\right)$$

Subject to the following constraints:

1. Regular download traffic from mobile core node (ASR5000) to eNodeB:

$$\sum_{c\in N}\lambda G_{c,e} = \alpha \cdot \lambda G_e \qquad (6.1)$$
$$\forall e \in E$$

2. Video streaming traffic to eNodeB

$$\sum_{n\in N}\lambda S_{n,e} = \lambda \mathrm{V}_e - \sum_{h\in H}\lambda C_{h,e} \qquad (6.2)$$
$$\forall\, e \in E$$

Constraint (6.1) represents the regular traffic from the mobile core node represented by ASR5000 to eNB *e*, where α is a unitless quantity which represents the ratio of backhaul to fronthaul traffic. Constraint (6.2) determines the video streaming traffic sourced by both video (contents) server and content caching node.

3. Request to video streaming

$$\sum_{h\in H}\sigma C_{h,e} \leq 1 \qquad (6.3)$$
$$\forall\, e \in E$$

4. Traffic from cache to eNB nodes

$$\lambda \mathrm{C}_{h,e} = \delta_h \cdot \sigma C_{h,e} \cdot \lambda V_e \quad \forall\, h \in H, \forall e \in E \tag{6.4}$$

5. Cache node location

$$\beta \cdot \sum_{e \in E} \lambda \mathrm{C}_{h,e} \geq \sigma C_h \quad \forall\, h \in H \tag{6.5}$$

$$\beta \cdot \sigma C_h \geq \sum_{e \in E} \lambda C_{h,e} \quad \forall\, h \in H \tag{6.6}$$

6. Hit ratio of the content cache node

$$C\mathrm{Z}_h \geq (\delta_h \cdot a_k + b_k) \quad \forall\, h \in H, \forall k \in K \tag{6.7}$$

Constraint (6.3) determines whether the eNodeB *e* is served by cache at node *h*. Constraint (6.4) determines the amount of traffic flow from the cache node to the eNB based on the cache hit ratio. Constraints (6.5) and (6.6) determine the location of the cache node by setting the binary variable (σC_h) to zero or one. In constraints (6.5) and (6.6) the term ($\sum_{e \in E} \lambda C_{h,e}$) is the total traffic from the cache at node *h* to all eNodeBs. When a traffic from cache at node *h* flows to any eNodeB *e*, the value of ($\sum_{e \in E} \lambda C_{h,e}$) is greater than zero. In this case, any value of (σC_h) satisfies constraint (6.5), but the only value of (σC_h) that satisfies constraint (6.5) is 1. Consequently, the value of (σC_h) that satisfies both constrains at the same time is 1. Therefore, once the traffic flows from the cache at node *h* to any eNodeB node, the

value of (σC_h) is set to 1. Following the same methodology, the MILP model uses constraints (6.5) and (6.6) to set the value of (σC_h) to zero when there is no traffic flow from the cache at node *h*. Table 6.4 illustrates the operation of constraints (6.5) and (6.6). Constraint (6.7) calculates the hit ratio based on the cache size using a piecewise linear approximation.

Table 6.4 Illustration of constraints (6.5) and (6.6)

$\sum_{e\in E} \lambda C_{h,e}$	Constraint	Outcome	Possible values of σC_h	The value of σC_h that satisfies both constraints
$\sum_{e\in E} \lambda C_{h,e} > 0$	$\beta \cdot \sum_{e\in E} \lambda C_{h,e} \geq \sigma C_h$	$\beta \cdot \sum_{e\in E} \lambda C_{h,e} \gg 1$	0 or 1	1
	$\sum_{e\in E} \lambda C_{h,e} \leq \beta \cdot \sigma C_h$	$\beta \cdot \sigma B_h \gg 1$	1	
$\sum_{e\in E} \lambda C_{h,e} = 0$	$\beta \cdot \sum_{e\in E} \lambda C_{h,e} \geq \sigma C_h$	$\beta \cdot \sum_{e\in E} \lambda C_{h,e} = 0$	0	0
	$\sum_{e\in E} \lambda C_{h,e} \leq \beta \cdot \sigma C_h$	$\beta \cdot \sigma B_h = 0$	0 or 1	

7. Location of mobile core node (ASR5000)

$$\beta \cdot \sum_{e\in E} \lambda G_{c,e} \geq \sigma A_c \quad \forall c \in N \tag{6.8}$$

$$\sum_{e\in E} \lambda G_{c,e} \leq \beta \cdot \sigma A_c \quad \forall c \in N \tag{6.9}$$

$$\sum_{c \in N} \sigma A_c = 1 \tag{6.10}$$

Constraints (6.8) and (6.9) determine the location of the mobile core node by setting the binary variable σA_e to 1 if the mobile core node is attached to node *c*, otherwise it is set to zero. These constraints operate in the same way as constraints (6.5) and (6.6) and they are illustrated in Table 6.5. Constraints (6.10) ensures that there is only one mobile core node (ASR5000) in the network, and that it is located at one of the IP over WDM nodes.

Table 6.5 Illustration of constraints (6.8) and (6.9)

$\sum_{e \in E} \lambda G_{c,e}$	Constraint	Outcome	Possible values of σA_c	Value of σA_c that satisfies both constraints
$\sum_{e \in E} \lambda G_{c,e} > 0$	$\beta \cdot \sum_{e \in E} \lambda G_{c,e} \geq \sigma A_c$	$\beta \cdot \sum_{e \in E} \lambda G_{c,e} \gg 1$	0 or 1	1
	$\sum_{e \in E} \lambda G_{c,e} \leq \beta \cdot \sigma A_c$	$\beta \cdot \sigma A_c \gg 1$	1	
$\sum_{e \in E} \lambda G_{c,e} = 0$	$\beta \cdot \sum_{e \in E} \lambda G_{c,e} \geq \sigma A_c$	$\beta \cdot \sum_{e \in E} \lambda G_{c,e} = 0$	0	0
	$\sum_{e \in E} \lambda G_{c,e} \leq \beta \cdot \sigma A_c$	$\beta \cdot \sigma A_c = 0$	0 or 1	

8. Video server location

$$\beta \cdot \sigma S_n \geq \sum_{r \in R} \lambda S_{n,e} \tag{6.11}$$
$$\forall\, n \in N$$

$$\beta \cdot \sum_{e \in E} \lambda S_{n,r} \geq \sigma S_n \qquad \forall\, n \in N \tag{6.12}$$

$$\sum_{n \in N} \sigma S_n = 1 \tag{6.13}$$

Constraints (6.11) and (6.12) determine the location of the contents (video) server, whilst constraint (6.13) ensure it is hosted at one IP over WDM node.

The three traffic types that flow toward eNodeB and the MILP decision (binary) variables are illustrated in Figure 6.2.

9. Total download traffic to BBU node *e*:

$$\lambda R_{h,e} = \lambda G_{h,e} + \alpha \cdot \left(\lambda C_{h,e} + \lambda S_{h,e}\right) \qquad \forall h \in N, \forall e \in E \tag{6.14}$$

$$\lambda R_{h,e} = \alpha \cdot \lambda C_{h,e} \qquad \forall h \in (U \cup L), \forall e \in E \tag{6.15}$$

Constraint (6.14) and (6.15) calculate the total download traffic to eNodeB *e* that is sourced by three nodes. These are: video server ($\lambda S_{h,e}$), cache node ($\lambda C_{h,e}$), and the regular traffic from the mobile core node ($\lambda G_{h,e}$).

10. Flow conservation of total downlink traffic to eNB nodes:

$$\sum_{y \in TN_x} \lambda R_{x,y}^{h,e} - \sum_{y \in TN_x} \lambda R_{y,x}^{h,e} = \begin{cases} \lambda R_{hr} & if\ x = h \\ -\lambda R_{hr} & if\ x = r \\ 0 & otherwise \end{cases} \qquad (6.16)$$

$$\forall h \in H, \forall e \in E, \forall x \in TN_x$$

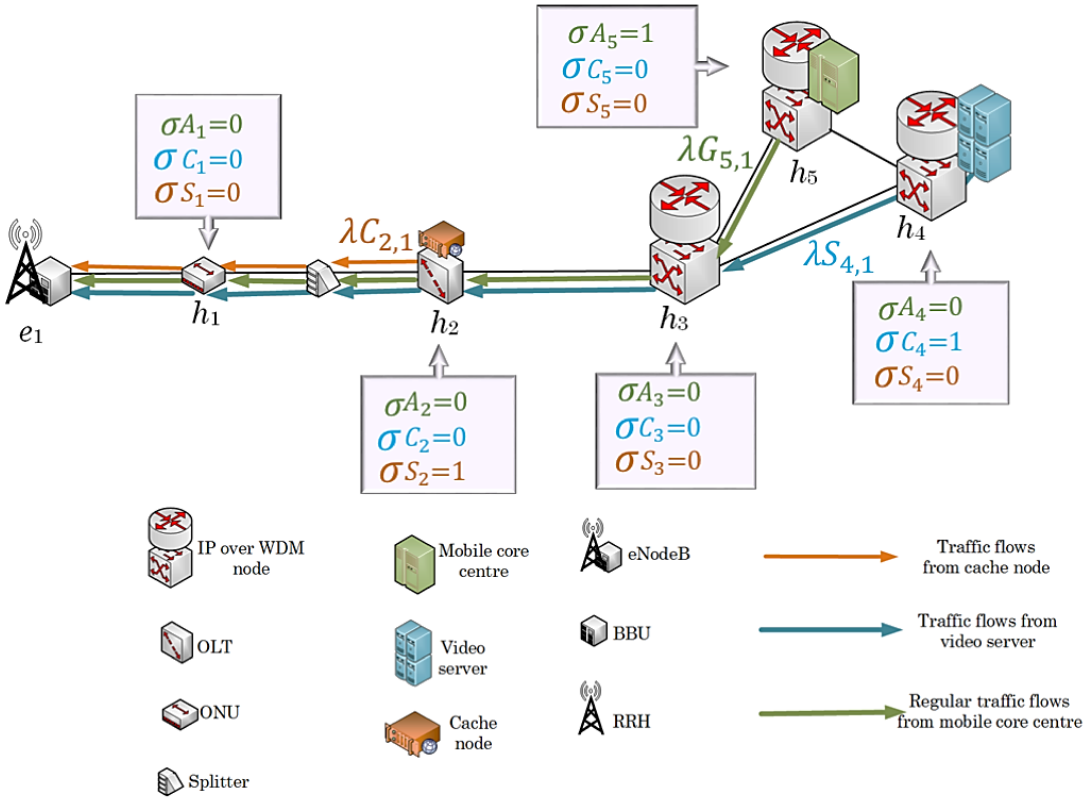


Figure 6.2 Illustration of MILP binary variables and traffic to eNodeB

11. GPON link constraints

$$\sum_{h \in H} \sum_{e \in E} \sum_{j \in TN_i \cap L} \lambda R_{i,j}^{h,e} \leq 0 \qquad (6.17)$$

$$\forall i \in U$$

$$\sum_{h\in H}\sum_{e\in E}\sum_{j\in \mathrm{TN_i}\cap \mathrm{N}} \lambda R_{i,j}^{h,e} \le 0 \qquad (6.18)$$
$$\forall i \in L$$

12. Constraint (6.16) represents the flow conservation of total download traffic to eNodeBs. Constraints (6.17) and (6.18) ensure that the download traffic of GPON does not flow in the opposite direction.
13. Virtual link capacity of IP over WDM network

$$\sum_{h\in H}\sum_{e\in E} \lambda R_{i,j}^{h,e} \le W_{i,j} \cdot B \qquad (6.19)$$
$$\forall i,j \in N, i \neq \mathrm{j}.$$

14. Flow conservation in the optical layer of IP over WDM network:

$$\sum_{n\in NN_m} W_{m,n}^{i,j} - \sum_{n\in NN_m} W_{n,m}^{i,j} = \begin{cases} W_{i,j} & if\ n = i \\ -W_{i,j} & if\ n = j \\ 0 & otherwise \end{cases} \qquad (6.20)$$
$$\forall i,j,m \in N, i \neq j$$

Constraint (6.19) ensures that the total traffic traversing the virtual link (*i*, *j*) does not exceed its capacity, in addition it determines the number of wavelength channels that carry the traffic burden of that link. Constraint (6.20) represents the flow conservation in the optical layer of the IP over WDM network. It ensures that the total expected number of incoming wavelengths for the IP over WDM nodes of the virtual link (*i*, *j*) is equal to the total number of outgoing wavelengths of that link.

15. Number of wavelength channels

$$\sum_{i \in N} \sum_{j \in N: i \neq j} W_{m,n}^{i,j} \leq w \cdot f_{m,n} \qquad (6.21)$$

$$\forall m \in N, \forall n \in NN_m$$

$$W_{m,n} = \sum_{i \in N} \sum_{j \in N: i \neq j} W_{m,n}^{i,j} \qquad (6.22)$$

$$\forall m \in N, \forall n \in NN_m$$

16. Number of aggregation ports

$$\Lambda_{\mathrm{i}} = \left(\sum_{\mathrm{j} \in \mathrm{L} \cap \mathrm{TN}_{\mathrm{i}}} \left(\sum_{\mathrm{h} \in \mathrm{H}} \sum_{\mathrm{e} \in \mathrm{E}} \lambda \mathrm{R}_{\mathrm{i,j}}^{\mathrm{h,e}} \right) \right) \Big/ \mathrm{B} \qquad (6.23)$$

$$\forall i \in N$$

Constraints (6.21) and (6.22) are the constraints of the physical link (*m*, *n*). Constraint (6.21) ensures that the total number of wavelength channels in the logical link (*i*, *j*) that traverses the physical link (*m*, *n*) does not exceed the fibre capacity. Constraint (6.22) determines the number of wavelength channels in the physical link and ensures it is equals to the total number of wavelength channels in the virtual link traversing that physical link. Constraint (6.23) determines the required number of aggregation ports in each IP over WDM router.

6.2.3 Energy-efficient variable size cache MILP model

The variable size cache MILP model optimises the cache size of each node for different total number of users over the time of the day. This model aims to determine the additional power savings that can be achieved compared to the fixed

cache sizes case. It also aims to analyse the impact of variable size caches on the total power consumption with the variation in the total number of network users over the time of day.

The variable size cache MILP model defines the same indices, parameters, and variables defined previously for the fixed cache MILP model. However, the cache sizes are variable for each node; therefore, they need to be set as variables, and additional variables are defined as follows:

$\Theta_{h,e}$: Floating variable equivalent to the multiplication of the binary variable σC_{he} by the cache hit ratio δ_h.

$iC\mathrm{Z}_h$: The integer part of the cache size at node h.
$fC\mathrm{Z}_h$: The floating part of the cache size at node h.

The same objective function defined for the fixed-size cache MILP model is defined for variable size cache MILP model except for the cache nodes power consumption, which is defined as follows:

$$\sum_{u \in H} \left(\varepsilon c \cdot CC \cdot (iC\mathrm{Z}_u / 100) \right)$$

In addition, the following constraints related to cache size and hit ratio are added:

$$\Theta_{h,e} \leq \sigma C_{h,e} \qquad \forall\, h \in H, \forall e \in E \tag{6.24}$$

$$\Theta_{h,e} \leq \delta_h \qquad \forall\, h \in H, \forall e \in E \tag{6.25}$$

$$\Theta_{h,e} \geq \delta_h - \left(1 - \sigma C_{h,e}\right) \quad \forall\, h \in H, \forall e \in E \tag{6.26}$$

$$\Theta_{h,e} \geq 0 \quad \forall\, h \in H, \forall e \in E \tag{6.27}$$

$$\delta_h \leq 1 \quad \forall\, h \in H, \forall e \in E \tag{6.28}$$

$$CZ_h = iCZ_h + fCZ_h \quad \forall\, h \in H \tag{6.29}$$

These constraints ((6.24) to (6.28)) determine the cache hit ratio for any cache at node h. Constraints ((6.24) to (6.26)) are equivalent to the multiplication of the hit ratio by the binary variable ($\sigma C_{h,e}$); whilst constraints (6.27) and (6.28) ensure that the hit ratio does not go beyond 1 or less than 0. Constraints (6.29) rounds down the cache size (flooring) to the nearest integer.

As a binary variable, the variable ($\sigma C_{h,e}$) can be either zero or 1. When the value of ($\sigma C_{h,e}$) equals to 1, the value of ($\Theta_{h,e}$) in constraint (6.24) takes any value between 0 and 1 including value of the hit ratio (δ_h), whilst its value is less or equal to the hit ratio in constraint (6.25). In constraint (6.26) the value of ($\Theta_{h,e}$) is greater or equal to the hit ratio (δ_h). For the three constraints ((6.24) to (6.26)), the only value of ($\Theta_{h,e}$) that satisfies all the three constraints is the value of the hit ratio (δ_h). Using the same approach, the MILP model uses constraints ((6.24) to (6.26)) to set

the value of ($\Theta_{h,e}$) to zero when the binary variable ($\sigma C_{h,e}$) is zero. The operation of constraints ((6.24) to (6.26)) is summarised in Table 6.6.

Table 6.6 Illustration of constraints ((6.24) to (6.26))

$\sigma C_{h,e}$	Constraint	Possible values of $\Theta_{h,e}$	The value of $\Theta_{h,e}$ that satisfies all constraints
$\sigma C_{h,e} = 1$	$\Theta_{h,e} \leq \sigma C_{h,e}$	any value between 0 and 1	δ_h
	$\Theta_{h,e} \leq \delta_h$	$\Theta_{h,e} = \delta_h$ or $\Theta_{h,e} < \delta_h$	
	$\Theta_{h,e} \geq \delta_h - (1 - \sigma C_{h,e})$	$\Theta_{h,e} = \delta_h$ or $\Theta_{h,e} > \delta_h$	
$\sigma C_{h,e} = 0$	$\Theta_{h,e} \leq \sigma C_{h,e}$	$\Theta_{h,e} = 0$	0
	$\Theta_{h,e} \leq \delta_h$	$\Theta_{h,e} = \delta_h$ or $any\ value < \delta_h$	
	$\Theta_{h,e} \geq \delta_h - (1 - \sigma C_{h,e})$	$\delta_h = 0$ or any positive number	

In addition, the constraint of the traffic from the cache node to eNodeB, defined in equation (6.4), is amended as follows:

$$\lambda C_{h,e} = \Theta_{he} \cdot \lambda V_e \qquad (6.30)$$
$$\forall\, h \in H, \forall e \in E$$

6.3 MILP model setup and results

This section describes the network topology considered and specifies the input parameters used in the developed MILP model. In addition, the MILP model results are discussed in detail.

6.3.1 Network topology and input parameters

The tested network topology in chapter 5 is considered in this chapter with all the RRH nodes replaced by eNodeBs as shown in Figure 6.3. The topology consists of 5 IP over WDM nodes and 10 GPON networks connected in pairs to the IP over WDM nodes; two GPON networks for each IP over WDM node. Each GPON network consists of one OLT and two ONUs and each ONU is connected to one eNodeB. The topology has one video server and one mobile core node (ASR5000) where the locations of both in the IP over WDM network are optimised by the developed MILP models in such a way as to save the total power consumption. Each eNodeB in the network represents a small cell with a maximum number of users equal to 10. Each user in the small cell is allocated 5 physical resource blocks (PRB) as the users are assumed to request the same task from the network. Additionally, the average number of users in the network is varied over the time of the day according to the network user profile shown in Figure 5.9 in chapter 5. Accordingly, the amount of downlink traffic to each eNodeB depends on the total number of active users in the small cell and its maximum value is considered to be less than10 Gbps. The total video streaming traffic from both cache nodes and video streaming server toward eNodeBs is considered 80% of the total download traffic. The input parameters to the

developed MILP model are listed in Table 6.7. It is worth mentioning that all the caches are considered to have the same maximum capacity. The developed MILP model considers 17 time slots over the day from 0:00 to 24:00 hours in steps of 1.5 hours and a different number of network users at each time slot is considered.

It should be noted that the cache is specified in terms of three quantities mainly (i) storage capacity, (ii) the data rate it can support and (iii) the cache power consumption. The cache used,[269], has power consumption of 550W, storage capacity of 14.4TB, and has up to 40Gb/s data rate. The RRH/BBU works at 10Gb/s maximum in 5G. The ONU supports 2 RRHs/BBUs. Therefore, this content cache is enough for ONU placement in terms of data rate. The OLT serves in this case (chosen topology of Figure 6.3), 2 ONUs, ie needs 40Gb/s. Therefore, this cache is also suitable for placement at the OLT. The WDM node serves 2 OLTs, therefore the cache needs 80Gb/s data rate in this case. The cache line card in [269] can however be upgraded to one 100Gb/s line card interface which consumes comparable amount of power as four 10 Gb/s line cards [270].

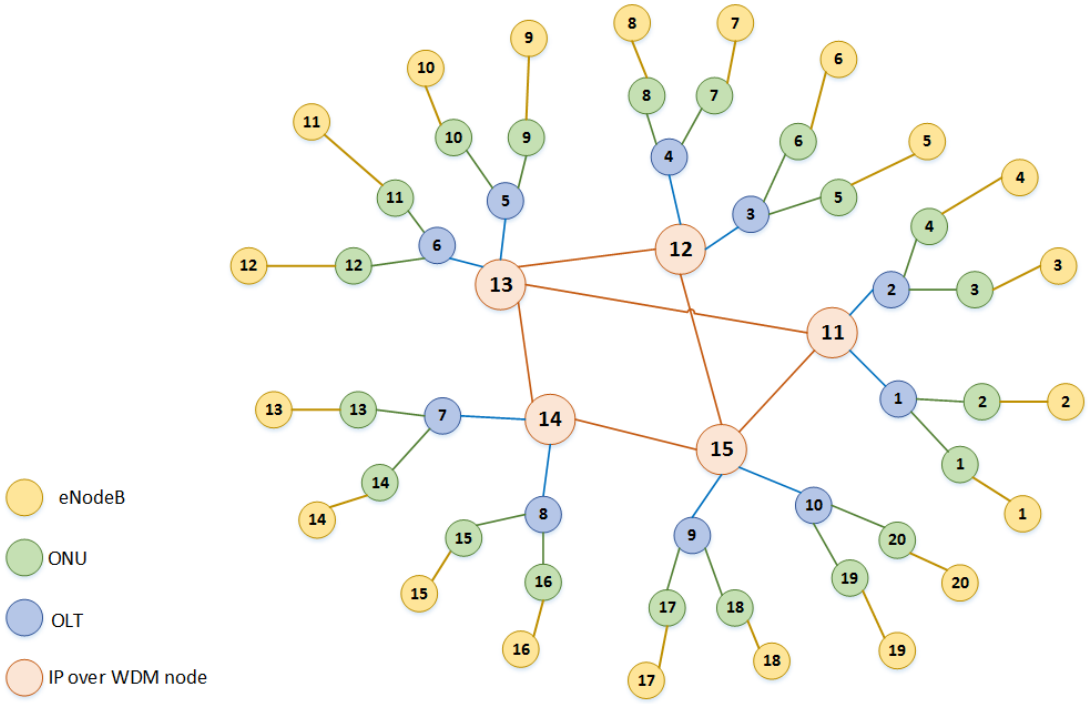


Figure 6.3 Tested network topology

Table 6.7 MILP model input parameters

Maximum fronthaul (CPRI) data rate for CPRI line rate option 7 (cp)	9.8304 (Gbps) [256]
Number of active users in a small cell (ρ_r)	Uniformly distributed (1-10 users)
Maximum number of users per cell (n)	10 (users)
Number of physical resources blocks per user (pb)	5 (PRB)
Hit ratio δ_h (for fixed-size cache MILP model only)	Range (0.1, 0.3, 0.5, 0.7, 1) of the maximum cache size
The ratio of the backhaul to the front haul traffic (α)	0.1344 (unitless)
ONU maximum power consumption (ΩU)	15 (W) [233]
OLT maximum power consumption (ΩL)	1940 (W) [231]
OLT idle power (ΩLd)	60 (W) [231]
OLT maximum capacity (CL)	8600 (Gbps) [231]
ONU maximum capacity (CU)	10 (Gbps) [233]
RRH node power consumption (ΩR_e)	1140 (W) [234]

BBU idle power (ΩBd_e)	51 (W) [263]
BBU maximum power consumption (ΩB_e)	531 (W) [263]
Cache node maximum power consumption (ΩC)	550 (W) [269]
Cache node maximum storage capacity (CC)	14.4 (TB) [269]
Video streaming server energy per bit (εs)	211.1 (Joul/Gb) [260]
Capacity IP over WDM wavelength channel (B)	40 (Gbps) [236]
Number of wavelengths per fibre in IP over WDM (w)	32 [236]
Transponder power consumption (ΩT)	167 (W) [237]
Router port power consumption (ΩRP)	825 (W) [172]
Regenerator power consumption (ΩG)	334 (W) [172]
EDFA power consumption (ΩE)	55 (W) [172]
Maximum span distance between EDFAs (S)	80 (km) [236]

6.3.2 Fixed-size cache MILP model results

A MILP model was utilised to evaluate the power consumption of five fixed-size cache approaches which are 10%, 30%, 50%, 70%, and 100% of the maximum cache node size. The MILP model was also used to compare the power consumption in this case to the power consumption when no caches are deployed at any node. In the no caches approach, all the video streaming traffic is delivered by streaming from the central video server only. Figure 6.4 shows the total power consumption of the no cache scenario and the power consumption under different fixed-size cache approaches (10%, 30%, 50%, 70%, and 100% of the maximum cache size). It is clearly seen that the no cache approach has a higher power consumption compared to the other approaches. In addition, the power consumption of the fixed-size cache MILP model increases as the cache size decreases. The MILP results show that all approaches consume very close amounts of total power for the time interval from 3

am to 9 am, where the total number of active users in the network is small during this period. This is clearly shown in Figure 6.5 where the total power consumption of the no cache and fixed-size cache approaches are compared for different number of active users in the network.

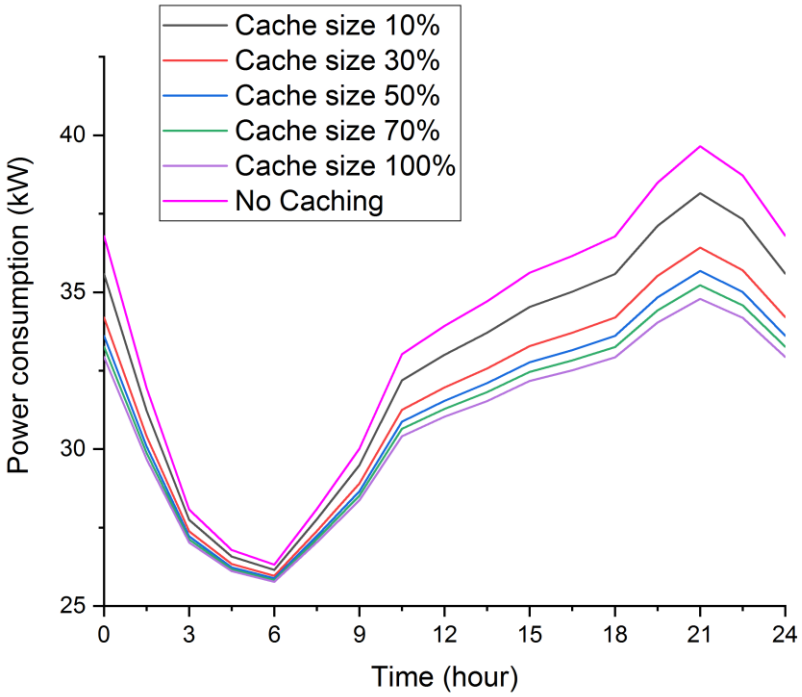


Figure 6.4 Total power consumption of no cache and fixed-size cache approaches at different times of the day

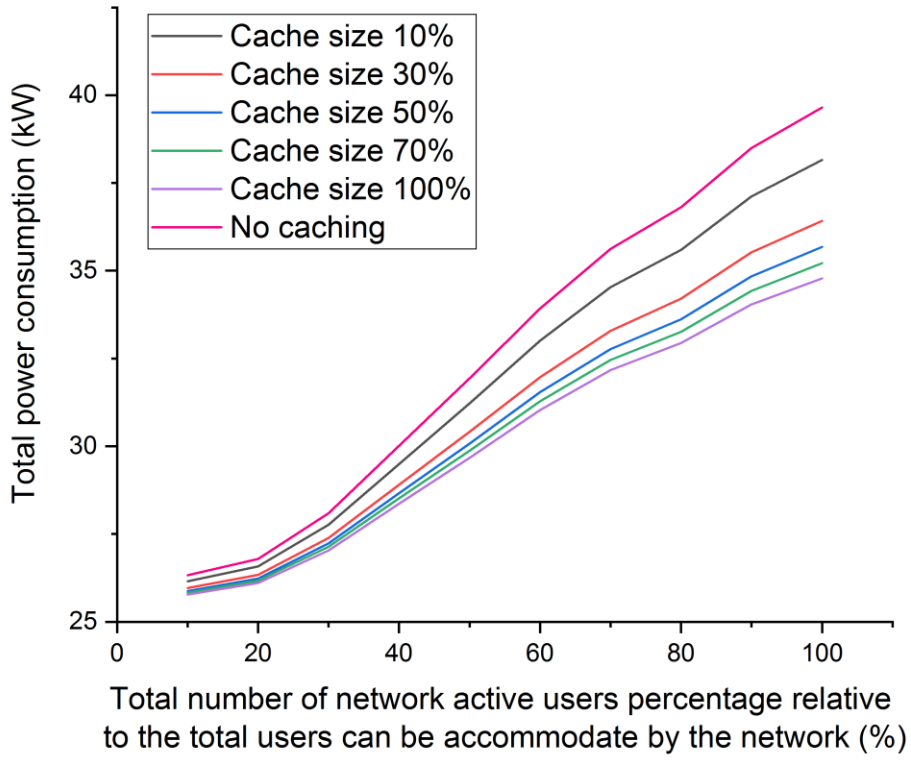


Figure 6.5 Total power consumption of no cache and fixed-size cache approach at different number of users

Figure 6.6 compares the total power saving for different fixed-size cache approaches with the no cache approach when the total numbers of active users in the network is varied. It is clearly seen that the power saving is affected by the cache size and the total number of users. The highest power saving (12.3%) is recorded when the highest cache size was deployed to serve a fully loaded network where all the users are active during the peak time of the day. This percentage drops during the early morning where only few users are active and it reaches its lowest value (around 1%) when the lowest cache size is deployed during this period of the day. As the total number of active users is low; the content caches are distributed mostly

at the IP over WDM nodes. This ensures that the active users are served by a small number of cache nodes.

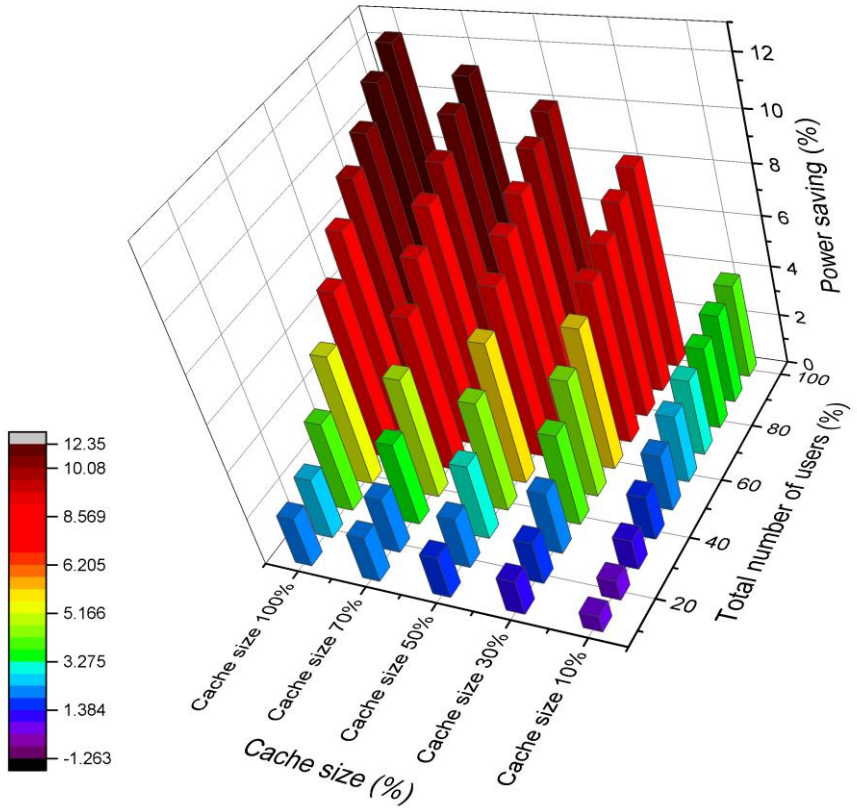
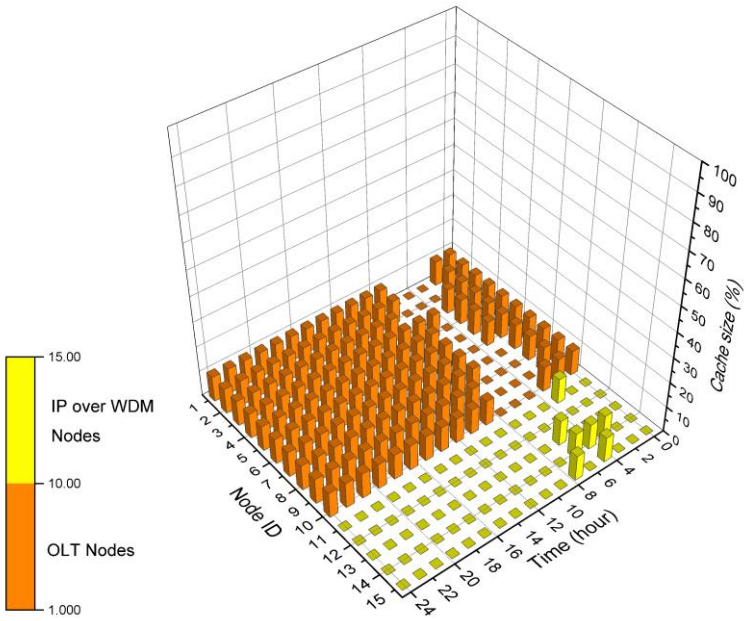


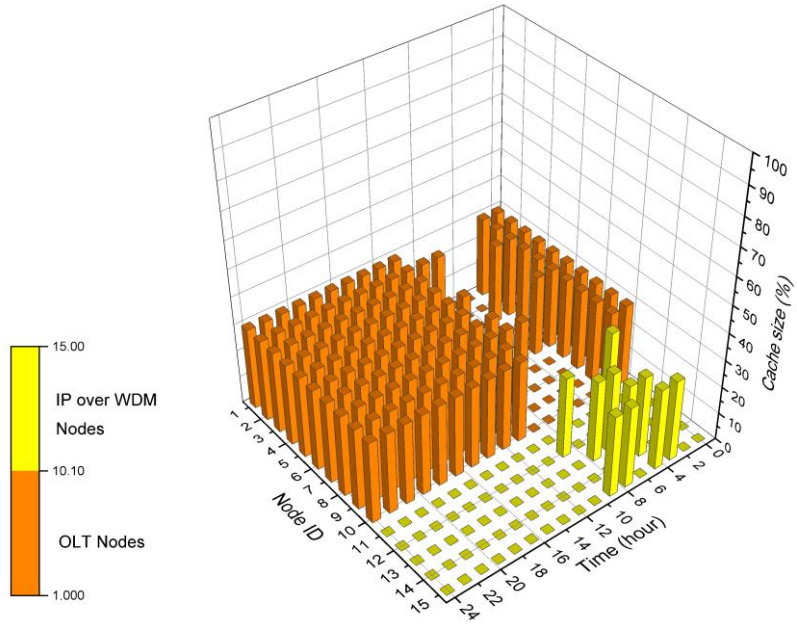
Figure 6.6 Total power saving of different fixed-size cache approaches compared with no cache approach

As the number of users increases, the requests for video streaming increase. Therefore, the MILP model tends to distribute more cache nodes close to the users to reduce the traffic induced power consumption. On the other hand, the distribution of cache nodes close to the user increases the number of deployed cache node which results in increase in the total power consumption of caching the content. Accordingly, the MILP model distributed the content caching nodes at OLTs to serve as many users as possible and maintain the total cache nodes power consumption as illustrated in Figure 6.7. Figure 6.7 depicts the optimum location of

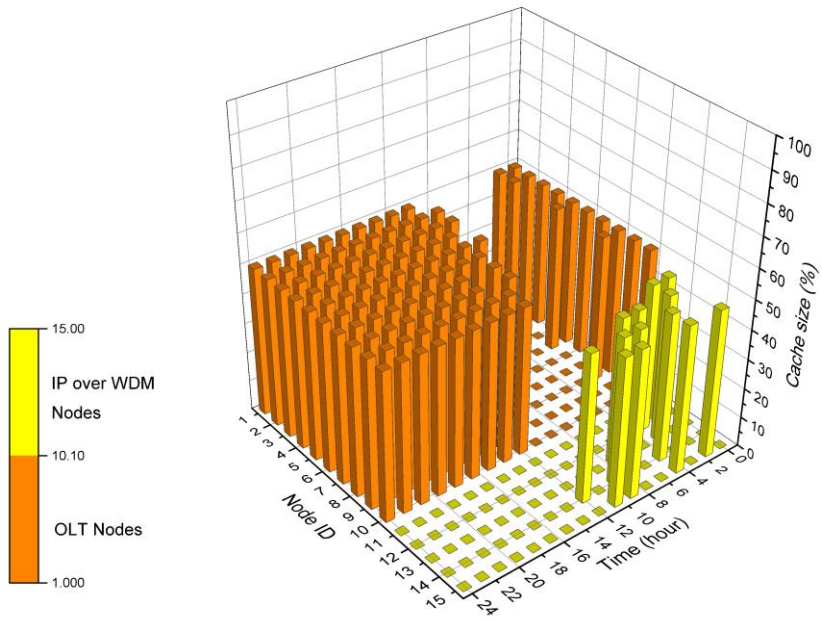
each cache node in the network at different times of the day for different fixed-size caches. For each node, the optimum cache location varies with the total number of active users at a certain time of the day. Therefore, when considering one row along the (Time) axis, the optimum cache node location for a certain node over 24 hours follows the trend of average number of users shown in Figure 5.9 in chapter 5. Considering a column, shows the optimum location in all the nodes at a certain time of the day. According to the MILP model results, there is no cache at any of the ONUs at any time of the day; therefore, Figure 6.7 does not list any ONU.



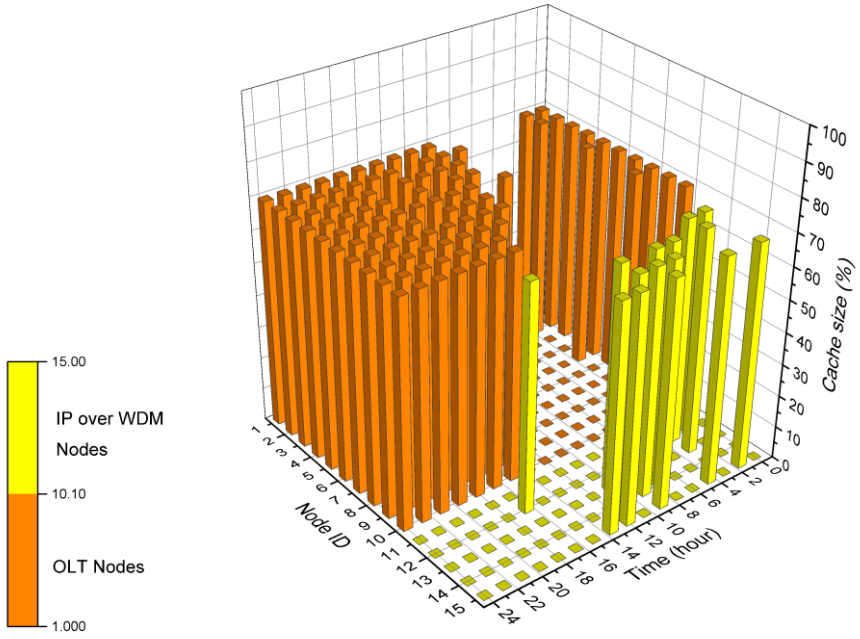
a) Cache size 10%



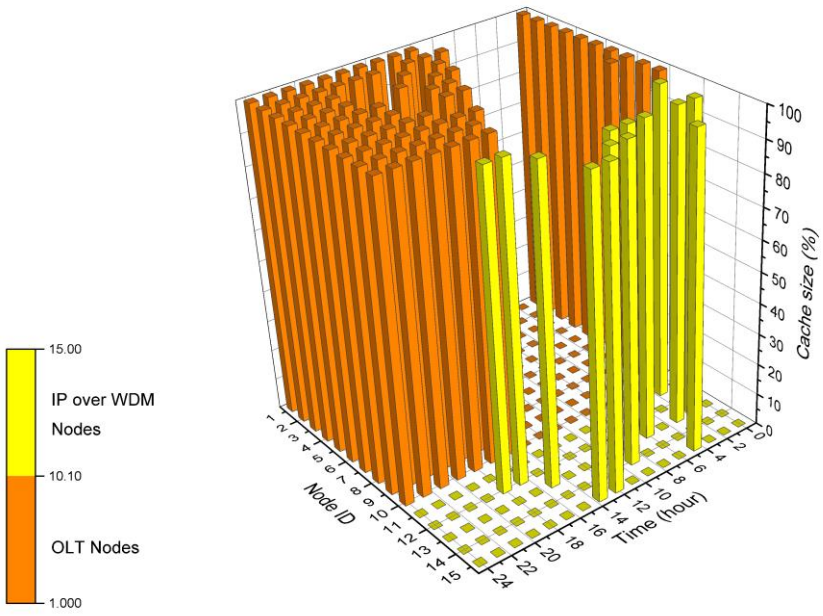
b) Cache size 30%



c) Cache size 50%



d) Cache size 70%



e) Cache size 100%

Figure 6.7 Cache distribution over the network for different cache size at different time of the day

6.3.3 Variable size cache MILP model

An MILP model was developed to evaluate the power consumption of a variable size cache approach and compare it with the power consumption of the five fixed-size cache approaches that were discussed earlier. Figure 6.8 illustrates the total power consumption of the variable and fixed-size cache approaches for different times of the day. It is clearly seen that the variable size cache approach has the lowest power consumption compared to the other approaches as the cache size and its location are optimised. Figure 6.9 compares the video server power consumption of the variable size cache with the fixed-size cache approaches at different times of the day. When the cache size increases, the amount of traffic that is offloaded from the video server decreases. This is because more users are served by cache nodes

which results in low video server power consumption. The video streaming power consumption of the variable size cache has almost the same trend as the fixed-size cache of 100% of the maximum cache node capacity.

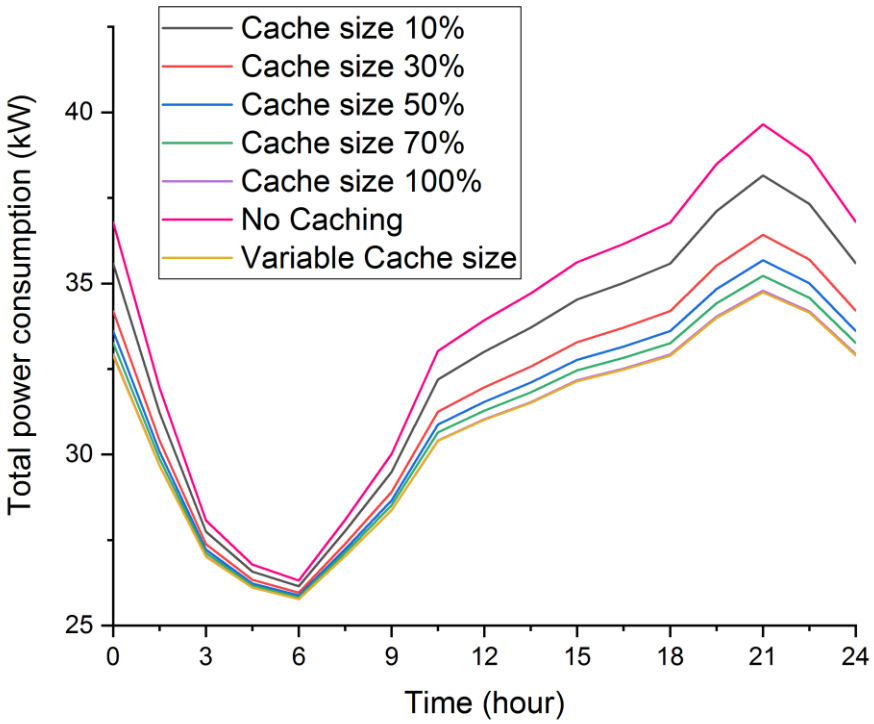


Figure 6.8 Power consumption of fixed and variable size cache approaches

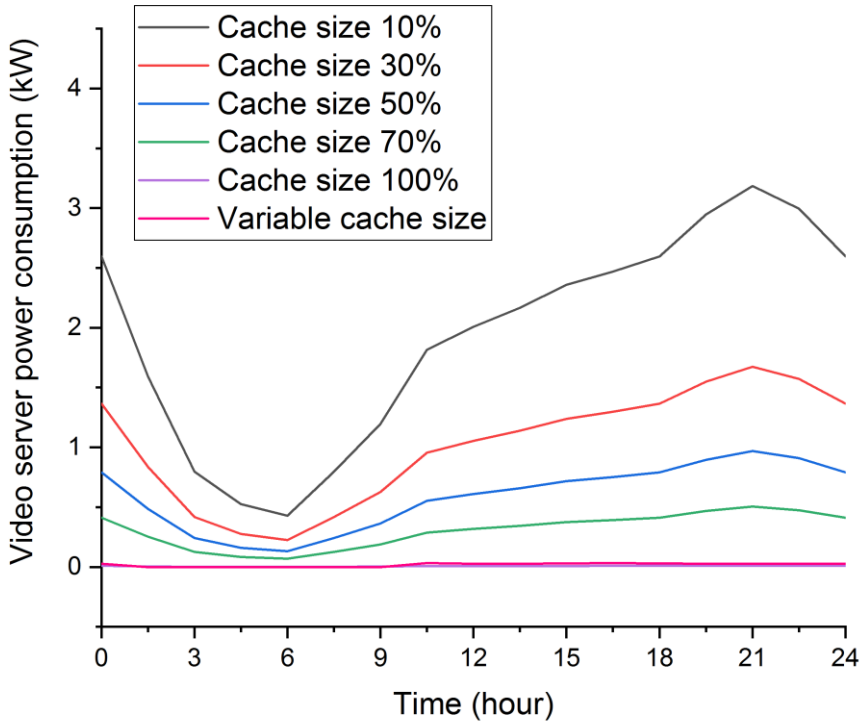


Figure 6.9 Video server power consumption of fixed and variable size cache approaches

Figure 6.10 compares the saving in total power consumption of the variable size cache with the fixed-size cache approaches for different number of active users. For each number of active user in time of the day, the optimum cache size was determined and the total power consumption was evaluated and compared with the power consumption of different cache size approaches. It is clear the total power saving is affected by the cache size and the total number of users. The highest power saving (9%) is recorded when the variable size cache was deployed to serve a fully loaded network where all the users are active during the peak time of the day. This percentage drops down during early morning where only few users are active and it reaches its lowest value (almost zero (0.006%)) when the variable size cache is

deployed during this period of the day. Figure 6.10 shows that the variable size cache approach has a very small power saving compared with the fixed-size cache of 100% of the maximum capacity of the cache node. The incentive in this case is that in both approaches, the MILP model offloads the traffic from the video server and distributes the cache contents at the same nodes in both approaches as illustrated in Figure 6.11. Figure 6.11 depicts the optimum cache size and its location in the network at different times of the day. For each node, the cache size varies with the total number of active user at a certain time. Therefore, considering a row in the (node) axis shows the optimum cache size for each node at a specific time of the day, whilst considering one row in the (time) axis shows the optimum cache size for one node over the day. Figure 6.11 shows that OLTs are the most optimum location for content caching during the day when the total number of users is high.

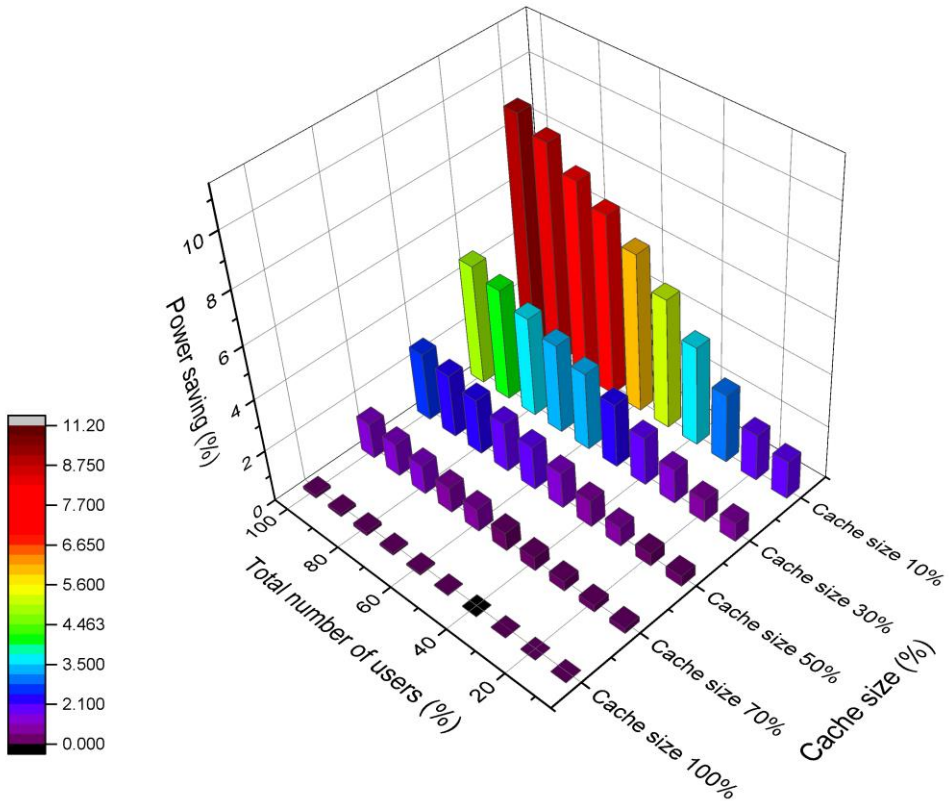


Figure 6.10 Total power saving of the variable size cache compared with the fixed-size cache approaches

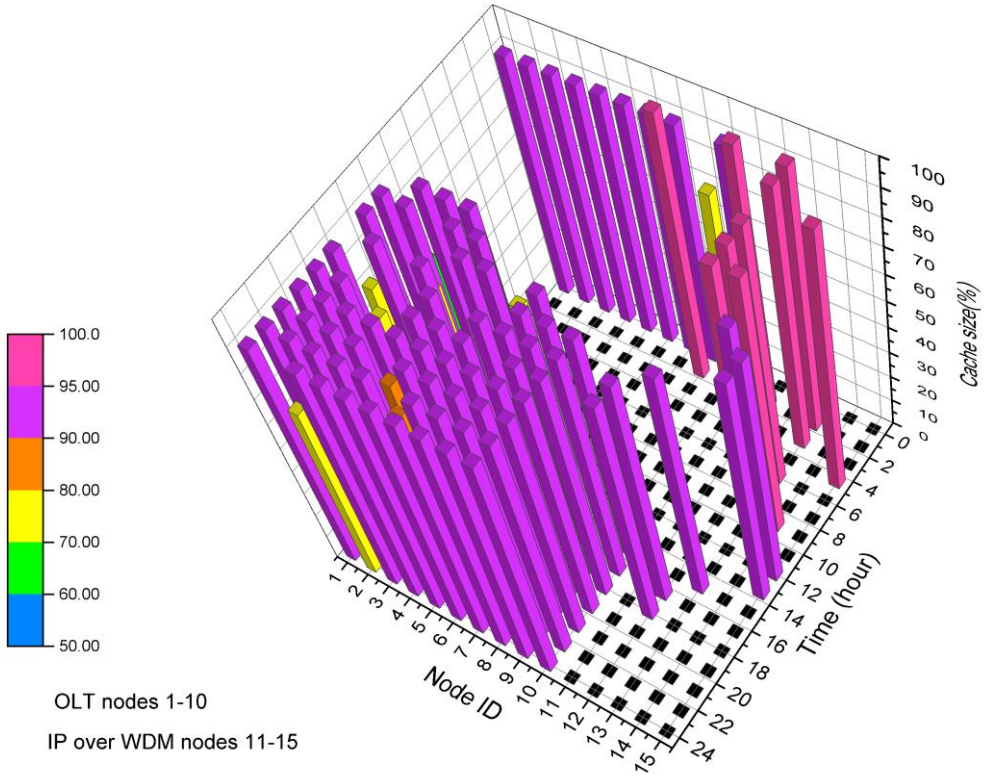


Figure 6.11 Cache distribution over the network for the variable size cache approach at different time of the day

6.4 Real-time heuristic models implementation

This section introduces two heuristic approaches for real-time implementation of the developed MILP model. The first heuristic approach considers the case where the cache node has a fixed size whilst the second approach considers the case where the cache nodes have a variable cache size.

6.4.1 Energy Efficient Fixed Cache Size (EEFCZ) heuristic model

The EEFCZ heuristic was developed to provide real-time implementation of the fixed cache size MILP model. The pseudocode of the heuristic is shown in Algorithm A.3. The network is modelled by sets of network elements *NE*, and links

L. The heuristic obtains the network topology $G = (NE, L)$ and the physical topology of the IP over WDM network $G_p = (N, L_p)$, where N is the set of IP over WDM nodes and L_p is the set of physical links. The total download request (fronthaul traffic) of each eNodeB node is calculated based on the total number of active users in each cell (eNodeB). The heuristic determines the hit ratio at each cache node according to the cache size. The EEFCZ heuristic chooses the closest place to the eNodeB node in such a way that the cache node serves as many eNodeB requests as possible. The heuristic examines the OLT nodes to determine the closest OLT for each eNodeB to accommodate the cache node. For each IP over WDM node, the heuristic determines the closest eNodBs and calculates the total number of users connected to the IP over WDM node. It then sorts the IP over WDM nodes in descending order according to the total number of user (node load). The first node at the top of the sorted IP over WDM nodes list is the recommended node to accommodate the mobile core node (ASR5000) and the video streaming servers. In such an approach, the EEFCZ heuristic ensures that a lower volume of backhaul traffic flows in the IP over WDM network. After accommodating the cache nodes, mobile core node, and the video streaming servers in the network, the EEFCZ heuristic obtains the physical graph $G_p = (N, L_p)$ and determines the traffic in each network segment. The IP over WDM network configuration such as the number of fibres, router ports, and the number of EDFA is determined and the total power consumption is calculated.

6.4.2 Energy Efficient Variable Cache Size (EEVarCZ) heuristic model

This section discusses the Energy Efficient Variable Cache Size (EEVarCZ) heuristic. It extends the EEFCZ heuristic which was discussed in the previous

section and it provides a real-time implementation of the MILP model where variable cache sizes are considered instead of fixed cache sizes. The pseudocode of EEVarCZ is shown in Algorithm A.4. To mimic the MILP model behaviour, the EEVarCZ heuristic calculates the total number of users in the network at a specific time of the day and compares the results to the maximum network capacity. If the total number of users is greater than half the network capacity, then the heuristic examines the OLT nodes to accommodate the cache nodes. Otherwise, it examines the IP over WDM network to accommodate the cache nodes. With such an approach, the heuristic ensures that a few cache nodes at the IP over WDM network server as many users as possible when the total number of users is small. The total number of candidate locations to host the cache nodes is limited to a specific number which is reduced by one every time the total power consumption is evaluated until the number of candidate locations that produces the least power consumption are found.

6.4.3 EEFCZ and EEVarCZ heuristic models results

In order to validate the results of the MILP model, the network topology shown in Figure 6.3 used for the MILP model, is used for both heuristic models. All the parameters considered in the fixed and variable cache size MILP models such as the wireless bandwidth, number of resources blocks per user, and the parameters listed in Table 6.2 are considered in both EEFCZ and EEVarCZ heuristics. The number of users allocated to each cell in the heuristic model are considered the same as in the MILP model to ensure that the traffic requested by each eNodeB node is the same in all models. Figure 6.12, Figure 6.13, and Figure 6.14 show the total power consumption of the MILP model compared to EEFCZ heuristic model at different

times of the day for cache sizes of 10%, 30%, and 70% of the maximum node cache size respectively. The EEFCZ heuristic successfully mimics the behaviour of the MILP model to the point that the difference in the total power consumption of the two is barely noticed. Figure 6.15 compares the total power consumption of the MILP model and the EEVarCZ heuristic for variable cache size at different times of the day. It is clearly seen that the total power consumption of the MILP model and the EEVarCZ heuristic are almost the same. This situation is mainly driven by the way that the EEVarCZ examines the candidate nodes to accommodate the cache nodes. It uses the total number of active users in the network to determine the place in which the cache nodes are accommodated. As alluded earlier, the IP over WDM and OLT nodes are the highly recommended locations to accommodate cache nodes, but the precedence is determined according to the total demanded traffic attributed to the number of users in each cell.

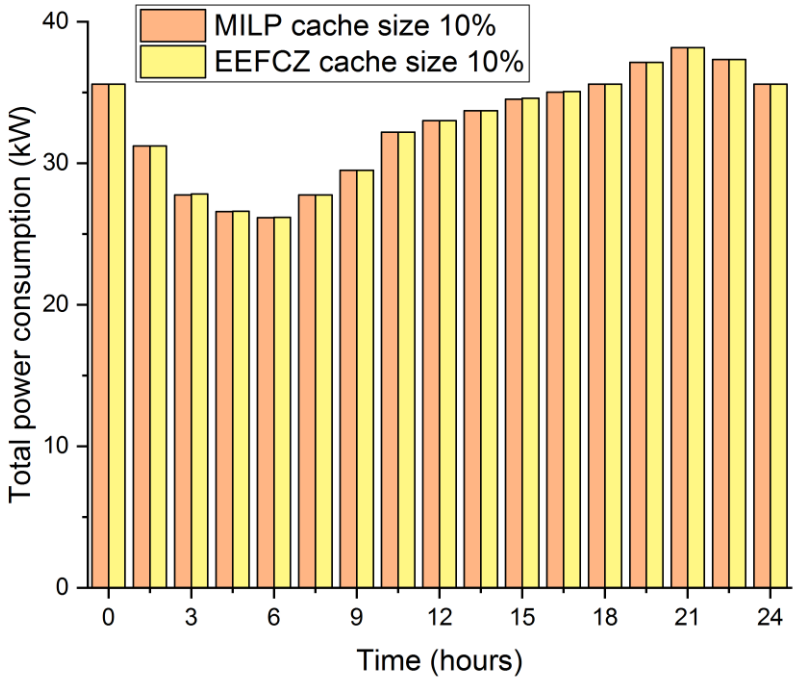


Figure 6.12 Total power consumption of MILP and EEFCZ models at cache size 10% of the cache node maximum capacity and at different times of the day

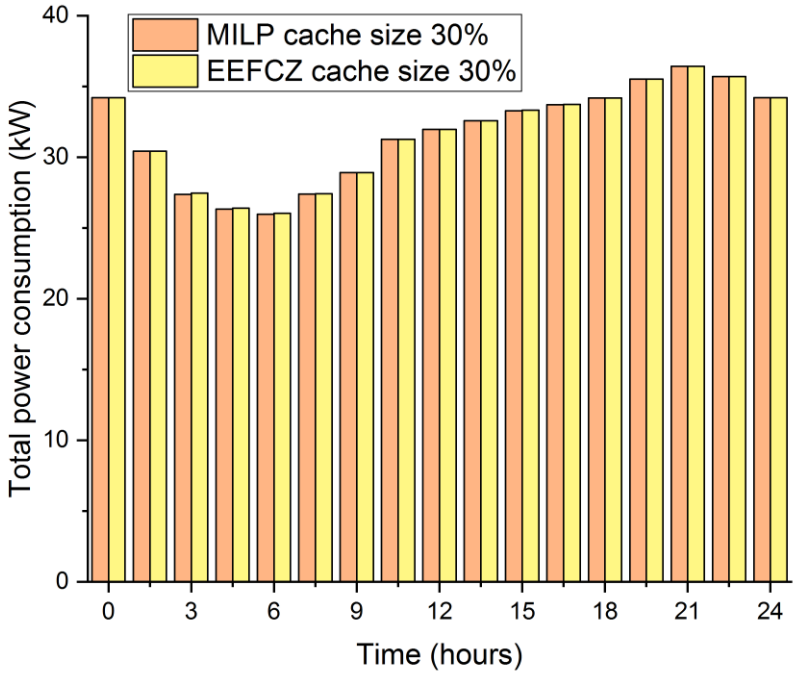


Figure 6.13 Total power consumption of MILP and EEFCZ models at cache size 30% of the cache node maximum capacity and at different times of the day

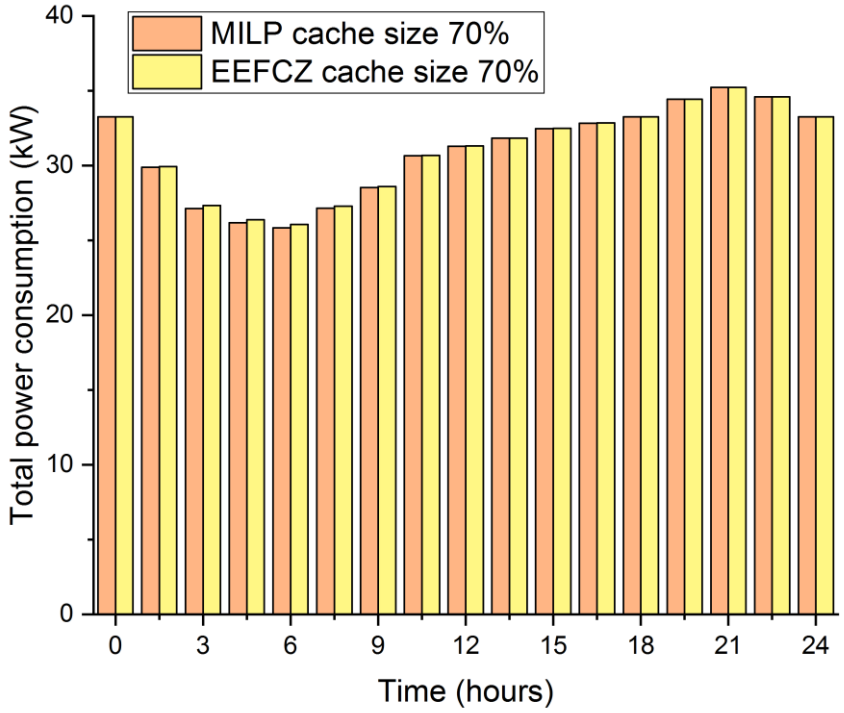


Figure 6.14 Total power consumption of MILP and EEFCZ models at cache size 70% of the cache node maximum capacity and at different times of the day

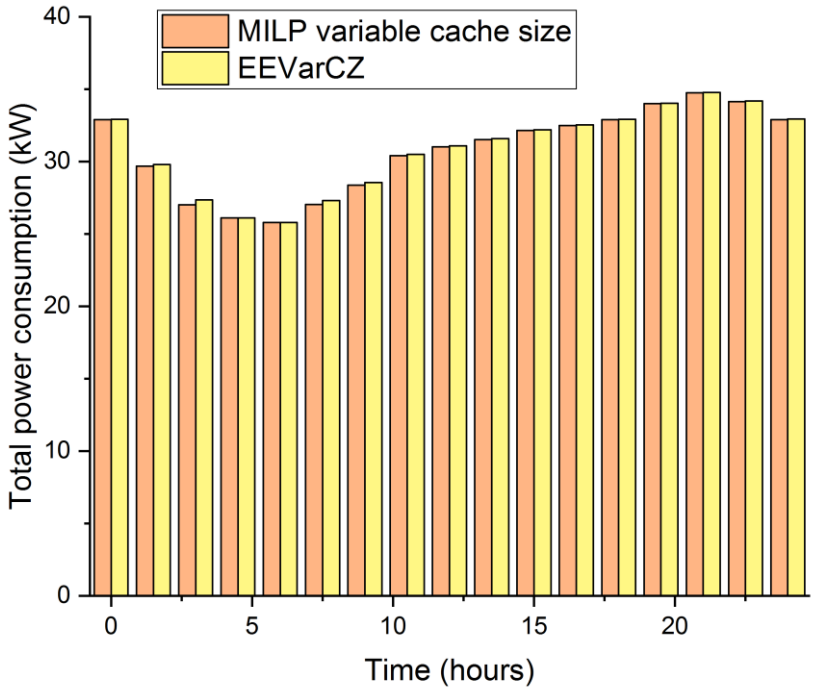


Figure 6.15 Total power consumption of MILP and EEVarCZ models for a variable cache size at different times of the day

6.5 Summary

Content caching has become a key technology in the design and implementation of communication systems especially with the growth in video streaming. This chapter has introduced optimised content caching in next generation of mobile networks, 5G, and has evaluated the associated energy efficiency. A MILP model was developed to minimise the total power consumption by optimising the location of fixed-size caches. Five scenarios were used in the developed MILP model with different fixed-size caches representing 10%, 30%, 50%, 70%, and 100% of the total cache node storage capacity and the results were compared to the approach where no cache was deployed.

The results show that OLTs are the optimum location to cache the contents when the cache size is small and the total number of active users is high, whilst the IP over WDM nodes are the optimum place to cache the contents when the number of users is low. As the cache size increases, more IP over WDM nodes are utilised to cache the contents resulting in a lower number of caching participant OLTs. In addition, in a fully loaded network where all the users are active during the peak time of the day, the optimum is the highest cache size (for the set of parameters and equipment power consumption values considered). This provides a power saving of (12.3%) compared to the no cache approach. This percentage drops during the early morning where only few users are active and it reaches its lowest value (0.65%) when the lowest cache size is deployed during this period of the day

The developed MILP model was extended to consider variable size caches by optimising the cache size and the location for each node over the time of the day. The results were compared with the five fixed size caching approaches. The results reveal that the variable size caching approach achieves a power saving of (9%) compared to the fixed size caching approach when the network is fully loaded, whilst this percentage drops during the early morning where only few users are active and it reaches its lowest value (almost zero (0.006%)).

For MILP models validation and also for real-time implementation of our approaches, this chapter has introduced two heuristics. The first heuristic is the Energy Efficient Fixed Cache Size (EEFCZ) heuristic and the second heuristic is EEVarCZ and deals with the variable size caching approach. The EEFCZ heuristic was developed to validate the results of the fixed cache size MILP model whilst the EEVarCZ was developed to validate the results of the variable cache size MILP

model. The heuristics' results were compared with their counterparts MILP models' results and are found to be in close agreement.

Chapter 7

Synergy of Virtualisation and Caching the Contents for an Energy-Efficient 5G Networks

7.1 Introduction

In recent years, the appetite for multimedia services has witnessed a tremendous increase. Therefore, the current mobile network may not be able to tackle the growth in traffic volume and the user needs without consuming unnecessarily large energy resources [271]. Meanwhile, caching the contents and NFV are considered promising technologies for the design and implementation of 5G networks [91]. The efficient deployment of content caching in mobile networks can improve the quality of service (QoS) and reduce the backhaul and core network traffic congestion. On the other hand, NFV enables better network resource utilisation, reduces the operation cost and provides seamless development of new services. Although good work has been done on NFV and content caching, these two technologies have been investigated separately in the literature. However, it is beneficial to jointly investigate these two technologies to improve the energy efficiency in 5G networks. Therefore, this chapter jointly considers content caching and NFV for reduced energy consumption in 5G networks. It evaluates the energy consumption of delivering video over optical-based architectures for 5G networks with virtualised mobile functions and resources. A mixed integer linear programming (MILP) model was developed to jointly optimise the cache size and the location of both caches and

VMs. To achieve maximum power saving, the MILP model finds the optimum cache size and the VMs workload at each node at different times of the day for different number of users.

7.2 Energy-efficient content caching and NFV in 5G networks: Architecture and MILP model

NFV and contents caching are promising technologies for 5G networks where content caching and VM processing of user data close to the end users results in shorter paths to the contents and low traffic induced power consumption. However, this approach requires more equipment for caching the contents and hosting VMs which eventually increases the equipment induced power consumption. Therefore, the cache size, VM utilisation of nodes, and the location of both cache nodes and VMs are considered. This evaluation minimises the total power consumption by optimising the cache size, VM utilisation of nodes and the location of caches and VMs.

7.2.1 Network architecture

Figure 7.1 illustrates a content caching service for video on demand over optical-based networks for 5G with virtualised mobile functions. The architectures proposed in Chapter 5 and 6 are integrated in this chapter. The integrated network consists of three layers; IP over WDM network, wired optical access network represented by a passive optical network (PON), and mobile radio access network (RAN) represented by a set of RRH nodes. As in the previous chapters, two types of VMs are proposed; the first type carries out the mobile core network functions and is dubbed (CNVM),

whilst the second is in charge of the BBU function and is dubbed (BBUVM). The video server location is restricted to one of the IP over WDM nodes, whilst the content cache and the two types of VMs (CNVM and BBUVM) can be anywhere (at ONU, OLT, and IP over WDM node) in the network. Accordingly, four types of traffic flow in the network: traffic from video server, traffic from content cache, traffic from CNVM and traffic from BBUVM. As shown in Figure 7.1, the BBUVM aggregates the traffic from CNVM, video server, and cache to perform baseband processing and transmits the processed traffic to the RRH node. It is worth mentioning that the traffic from the video streaming server and the cache nodes is considered 80% of the total download traffic toward RRH nodes, whilst the traffic from the mobile core VMs (CNVM) is considered 20% of the total traffic as discussed earlier in Chapter 6. For clarity and to avoid any clutter of lines, the traffic toward only three eNodeB nodes are shown in Figure 7.1.

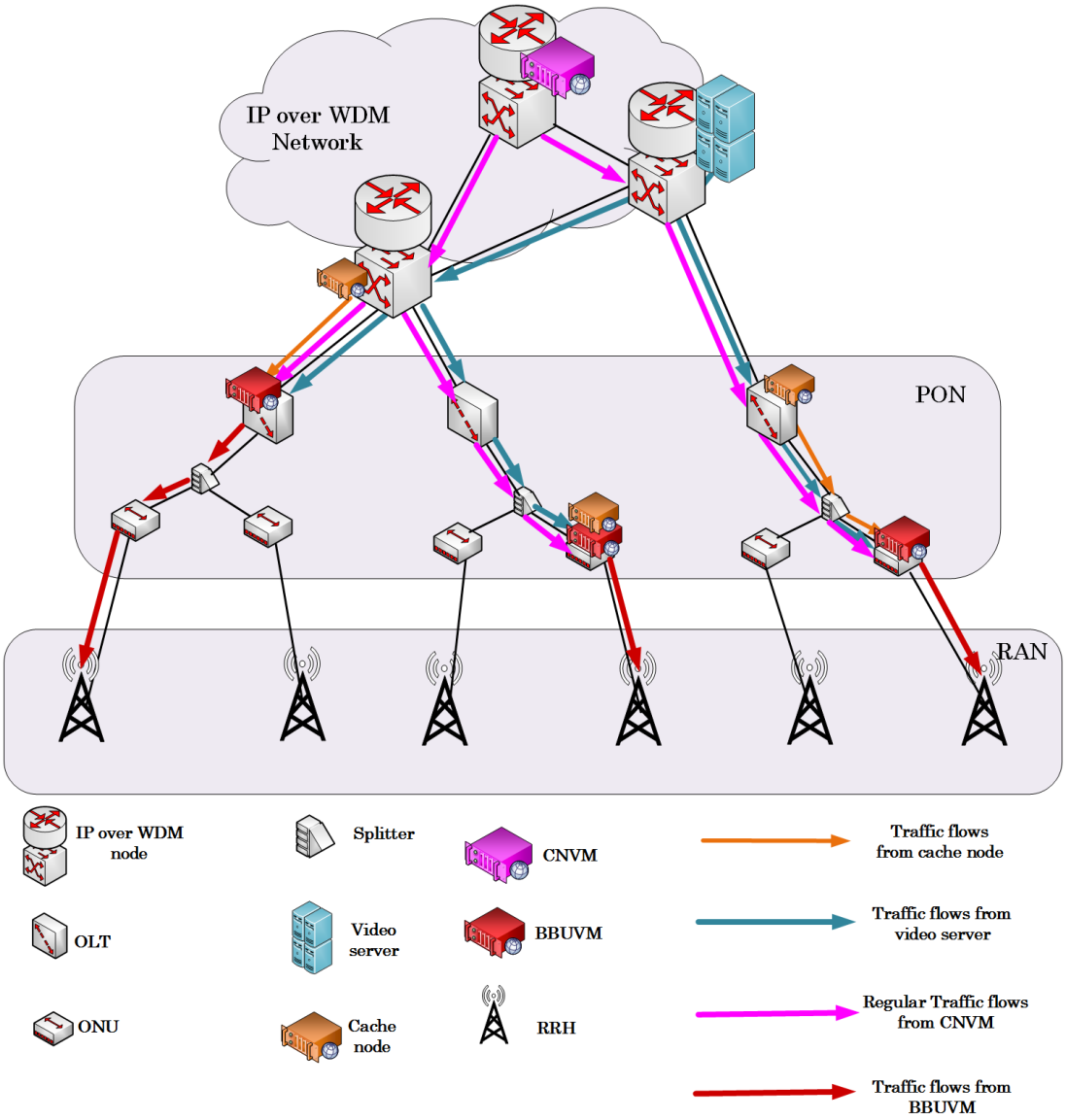


Figure 7.1 Contents caching with NFV in 5G networks

7.2.2 Model for energy-efficient content caching and NFV in 5G

The MILP model in this section is developed by utilising the features of both MILP models developed in Chapters 5 and 6. It jointly optimises the cache size, CNVM, and BBUVM utilisation at each node for different number of users over the time of the day. This model aims to investigate the additional power savings that can be achieved through combined caching and NFV in 5G compared to the virtualisation only and caching only approaches of Chapters 5 and 6. The model also

analyses the impact of integration of NFV and content caching on the total power consumption with variation in the total number of network users over the time of the day.

The model declares a number of indices, parameters, and variables. These are listed in Table 7.1, Table 7.2, and Table 7.3 respectively

Table 7.1 Energy-efficient caching and virtualisation MILP model indices

Indices	Comment
x, y	Indices of any two nodes in the network
m, n	Indices of any two nodes in the physical layer of the IP over WDM network
i, j	Indices of any two nodes in the IP layer of the IP over WDM network.
r	Index of RRH node
h, u, p, q	Indices of the nodes where the VM or cache could be hosted

Table 7.2 Energy-efficient caching and virtualisation MILP model parameters

Parameters	Comment
R	Set of RRH nodes
U	Set of ONU nodes
L	Set of OLT nodes
N	Set of IP over WDM nodes
T	Set of all nodes (RRH, ONU, OLT, and IP over WDM nodes)
NN_m	Set of neighbours of node m in the IP over WDM network, $\forall\ m\ \in N$
TN_x	Set of neighbours of node x, $\forall\ x\ \in T$
H	Set of nodes where the VM or cache can be placed (ONU, OLT,

	and IP over WDM nodes)
K	Set of Linearization coefficients
l	Line coding rate (bits per sample)
y	Number of MIMO layers
q	Number of bits used in QAM modulation
a	Number of antennas in a cell
cp	CPRI link data rate
ΨX	Maximum BBU workload needed for fully loaded RRH (GOPS); calculated as: $30 \cdot a + 10 \cdot a^2 + 20 \cdot q \cdot l \cdot y$
ΨS	Server CPU maximum workload (GOPS)
ΨC	Workload needed for hosting one CNVM (GOPS)
ρ_r	Number of mobile users connected to RRH node r
n	Maximum number of resources blocks per cell (per RRH node)
pb	Physical resource block per user
λR_r	RRH node r traffic demand (Gbps); calculated as: $[(pb/n) \cdot cp \cdot \rho_r]$, where $r \in R$
λV_r	Video streaming traffic to RRH node r
λG_r	Regular traffic to the RRH node r
μ	Large number
$\nabla_{p,q}$	Intra-traffic between core network VMs (CNVM) at nodes p, and q (Gbps)
α	The ratio of the backhaul to the fronthaul traffic (unitless)
ΩU	ONU maximum power consumption (W)
ΩL	OLT maximum power consumption (W)
ΩLd	OLT idle power (W)
CL	OLT maximum capacity (Gbps)

CU	ONU maximum capacity (Gbps)
ΩR_x	Power consumption of the Remote Radio Head (RRH) connected to ONU node *x* (W)
ΩS	Server maximum power consumption (W)
ΩSd	Server idle power consumption (W)
ΩH_h	Maximum power consumption of hosting VMs at node *h* (W)
ΩC	Cache node maximum power consumption (W)
CC	Cache node maximum storage capacity (GB)
εc	Cache node energy per stored gigabyte $\varepsilon c = \Omega C / CC$ (W/GB)
εs	Video streaming server energy per bit (Joule/Gb)
a_k, b_k	Linearisation coefficients (unitless)
β	Large number
η	Very small number (unitless)
B	Capacity of the wavelength channel (Gbps)
w	Number of wavelengths per fiber
ΩT	Transponder power consumption (W)
ΩRP	Router power consumption per port (W)
ΩG	Regenerator power consumption (W)
ΩE	EDFA power consumption (W)
$NG_{m,n}$	Number of regenerators in the optical link (m, n)
S	Maximum span distance between EDFAs (km)
$D_{m,n}$	Distance between node pair (m, n) in the IP over WDM network (km)
$A_{m,n}$	Number of EDFAs between node pair (m, n) calculated as $A_{m,n} = \left((D_{mn}/S) - 1\right) + 2$

Table 7.3 Energy-efficient caching and virtualisation MILP model variables

Variables	Comment
$\lambda B_{p,h}$	Traffic from CNVMs in node p to the BBUVMs in node h (Gbps)
$\lambda R_{h,r}$	Total download traffic from BBUVMs in node h to the RRH node r (Gbps)
$\sigma B_{h,r}$	Binary indicator, set to 1 if the node h hosts BBUVMs to serve the RRH node r, 0 otherwise
σB_{h}	Binary indicator, set to 1 if the node h hosts a BBUVM, 0 otherwise
$\sigma E_{p,h}$	Binary indicator, set to 1 if the node h hosts CNVMs to serve the BBUVMs at hosting node h, 0 otherwise
σE_{p}	Binary indicator, set to 1 if the hosting node p hosts CNVMs is, 0 otherwise
$\psi_{p,q}$	Binary indicator, set to 1 if two different hosting nodes p and q host CNVMs, 0 otherwise. It is equivalent to the ANDing of the two binary variables (σE_p, σE_q).
$\sigma \chi_h$	Binary indicator, set to 1 if the hosting node h hosts any virtual machine of any type, 0 otherwise. It is equivalent to the ORing of the two binary variables (σB_h, σE_h).
$\lambda E_{p,q}$	Traffic between hosting nodes due to CNVMs communication (Gbps)
$\lambda S_{s,r}$	Video streaming traffic between BBUVM and RRH node *r* from a video server at node *s*
$\lambda C_{u,r}$	Traffic between BBUVM and RRH node *r* from the cache at node *u*
$\sigma C_{u,r}$	Binary variable, set to 1 if the cache is located at node *u* to serve the RRH node *r*
$\lambda G_{h,r}$	Regular traffic from BBUVM in node *h* to the RRH node *r*
$\lambda C_{u,h,r}$	Traffic from the cache at node *u* to the RRH node *r* passing through the BBUVM in node *h*
$\lambda S_{s,h,r}$	Video streaming traffic from video server at node *s* to the node RRH
$\lambda S_{s,h}$	Video streaming traffic from video server at node *s* to a BBUVM at node *h*

σS_s	Binary variable, set to 1 if a video server is attached to the node s
$\lambda T_{p,q}$	Total download traffic from node p to node q (Gbps)
$\lambda R_{x,y}^{h,r}$	Traffic from hosting node h to RRH node r that traverses the link between the nodes (x, y) in the network in Gb/s
$\lambda T_{x,y}^{p,q}$	Total traffic from node p to node q that traverses the link between the nodes (x, y) in the network (Gbps)
ΨB_h	Total baseband workload at node h (GOPS)
Ψi_h	The integer part of the total normalised workload at node h.
Ψf_h	The fractional part of the total normalised workload at node h.
Ψi_h	The integer part of the total normalised workload at node h.
δ_u	Hit ratio of the cache at node u
$\Theta_{u,r}$	Floating variable equivalent to the multiplication of the binary variable σC_{ur} by the cache hit ratio
$C\mathrm{Z}_u$	Cache size at node u
$iC\mathrm{Z}_u$	The integer part of the cache size at node u
$fC\mathrm{Z}_u$	The fractional part of the cache size at node u
$W_{i,j}$	Number of wavelength channels in the virtual link (i, j)
$W_{m,n}^{i,j}$	Number of wavelength channels in the virtual link (i, j) that traverse the physical link (m, n)
$f_{m,n}$	Number of fibres in the physical link (m, n)
$W_{m,n}$	Total number of wavelengths in the physical link (m, n)
Λ_m	Number of aggregation ports of the router at node m

The total power consumption is composed of:

7. IP over WDM network power consumption which in turn is composed of:
   f) Power consumption of routers ports:

$$\Omega RP \cdot \left(\sum_{m \in N} \Lambda_m + \sum_{m \in N} \sum_{n \in NN_m} W_{m,n} \right)$$

   g) Power consumption of transponders:

$$\Omega T \cdot \sum_{m \in N} \sum_{n \in NN_m} W_{m,n}$$

h) Power consumption of EDFAs:

$$\Omega E \cdot \sum_{m \in N} \sum_{n \in NN_m} A_{m,n} \cdot f_{m,n}$$

i) Power consumption of regenerators:

$$\Omega G \cdot \sum_{m \in N} \sum_{n \in NN_m} NG_{m,n} \cdot W_{m,n}$$

8. Power consumption of RRHs and ONUs:

$$\sum_{x \in U} \left[\Omega R_x + \frac{\Omega U}{CU} \cdot \left(\sum_{h \in H} \sum_{r \in R} \sum_{y \in TN_x} \lambda R_{x,y}^{h,r} + \sum_{p \in H} \sum_{q \in H: p \neq q} \sum_{y \in TN_x \cap H} \lambda T_{x,y}^{p,q} \right) \right]$$

9. Power consumption of OLTs:

$$\sum_{x \in L} \left[\Omega Ld + \frac{\Omega L - \Omega Ld}{CL} \cdot \left(\sum_{h \in H} \sum_{r \in R} \sum_{y \in TN_x} \lambda R_{x,y}^{h,r} + \sum_{p \in H} \sum_{q \in H: p \neq q} \sum_{y \in TN_x \cap H} \lambda T_{x,y}^{p,q} \right) \right]$$

10. Power consumption of VMs servers

$$\sum_{h \in H} \left(\Omega Sd \cdot (\Psi i_h + \sigma \chi_h) + \Psi f_h \cdot (\Omega S - \Omega Sd) \right)$$

11. Power consumption of video server:

$$\varepsilon s \cdot \sum_{s \in N} \sum_{h \in H} \sum_{r \in RRH} (\alpha \cdot \lambda S_{shr})$$

12. Power consumption of content caching nodes:

$$\sum_{u \in H} \left(\varepsilon c \cdot CC \cdot (iCZ_u / 100) \right)$$

The objective is to minimise the total power consumption given by:

$$\Omega RP \cdot \left(\sum_{m \in N} \Lambda_m + \sum_{m \in N} \sum_{n \in NN_m} W_{m,n} \right) + \left(\Omega T \cdot \sum_{m \in N} \sum_{n \in NN_m} W_{m,n} \right)$$

$$+ \left(\Omega E \cdot \sum_{m \in N} \sum_{n \in NN_m} A_{m,n} \cdot f_{m,n} \right) + \left(\Omega G \cdot \sum_{m \in N} \sum_{n \in NN_m} NG_{m,n} \cdot W_{m,n} \right) +$$

$$\sum_{x \in U} \left[\Omega R_x + \frac{\Omega U}{CU} \cdot \left(\sum_{h \in H} \sum_{r \in R} \sum_{y \in TN_x} \lambda R_{x,y}^{h,r} + \sum_{p \in H} \sum_{q \in H: p \neq q} \sum_{y \in TN_x \cap H} \lambda T_{x,y}^{p,q} \right) \right] +$$

$$\sum_{x \in L} \left[\Omega Ld + \frac{\Omega L - \Omega Ld}{CL} \cdot \left(\sum_{h \in H} \sum_{r \in R} \sum_{y \in TN_x} \lambda R_{x,y}^{h,r} + \sum_{p \in H} \sum_{q \in H: p \neq q} \sum_{y \in TN_x \cap H} \lambda T_{x,y}^{p,q} \right) \right] +$$

$$\left(\sum_{h \in H} \left(\Omega Sd \cdot (\Psi i_h + \sigma \chi_h) + \Psi f_h \cdot (\Omega S - \Omega Sd) \right) \right)$$

$$+ \left(\varepsilon s \cdot \sum_{s \in N} \sum_{h \in H} \sum_{r \in RRH} (\alpha \cdot \lambda S_{shr}) \right) + \left(\sum_{u \in H} \left(\varepsilon c \cdot CC \cdot (iCZ_u / 100) \right) \right)$$

Subjected to the following constraints:

17. Traffic from CNVMs to BBUVMs:

$$\sum_{p \in H} \lambda B_{p,h} = \alpha \cdot \sum_{r \in R} \lambda R_{h,r} \qquad (7.1)$$
$$\forall h \in H$$

18. Regular traffic to RRH node:

$$\lambda G_r = \sum_{h \in H} \lambda G_{hr} \quad \forall h \in H \qquad (7.2)$$
$$\forall r \in RRH$$

Constraint (6.1) represents the regular traffic from CNVMs to the BBUVMs in node h where α is a unitless quantity which represents the ratio of backhaul to fronthaul traffic. Note that this enables BBUVMs to receive traffic from more than a single CNVM, which may occur for example in network slicing.

Constraints (7.2) represents the regular traffic to RRH nodes from all BBUVMs at different nodes. This enables a RRH to receive traffic from more than a single BBUVM.

19. Served RRH nodes and the location of BBUVMs

$$\beta \cdot \lambda G_{h,r} \geq \boldsymbol{\sigma} B_{h,r} \quad \forall r \in R, \forall h \in H \tag{7.3}$$

$$\lambda G_{h,r} \leq \beta \cdot \boldsymbol{\sigma} B_{h,r} \quad \forall r \in R, \forall h \in H \tag{7.4}$$

$$\beta \cdot \sum_{h \in H} \lambda G_{hr} \geq \boldsymbol{\sigma} B_h \quad \forall r \in R \tag{7.5}$$

$$\sum_{h \in H} \lambda G_{hr} \leq \beta \cdot \boldsymbol{\sigma} B_h \quad \forall r \in R \tag{7.6}$$

Constraints (7.3) and (7.4) ensure that the RRH node *r* is served by the BBUVM located at *h*. constraints (7.5) and (7.6) determine the location of BBUVM; β is a large enough number to ensure that $\boldsymbol{\sigma} B_{h,r}$ and $\boldsymbol{\sigma} B_h$ are equal to 1 when $\sum_{h \in H} \lambda G_{hr} > 0$. In constraint (7.5) there are two possibilities for the value of $(\sum_{h \in H} \lambda G_{hr})$ which are either zero (no traffic from *h* to *r*) or greater than zero (there is traffic from *h* to *r*). When the value is zero, the left-hand side of the inequality $(\beta \cdot \sum_{h \in H} \lambda G_{hr})$ should be zero and this sets the value of $\boldsymbol{\sigma} B_h$ to zero. In the second case when the value of $(\sum_{h \in H} \lambda G_{hr})$ is greater than zero, the left-hand side of the inequality $(\beta \cdot \sum_{h \in H} \lambda G_{hr})$ will be much greater than 1 because of the large value of β. In this case; the value of $\boldsymbol{\sigma} B_h$ may be set to 1 or zero. In the same way constraint (7.6) sets the value of $\boldsymbol{\sigma} B_h$. Table 7.4 illustrates the operation of constrains (7.5) and (7.6).

Table 7.4 BBUVMs constraints illustration

The value of traffic $\sum_{\forall r \in R} \lambda G_{h,r}$	Constraint	Outcome	σB_h	Value of σB_h that satisfies both constraints
$\sum_{\forall r \in R} \lambda G_{h,r} > 0$	$\beta \cdot \sum_{\forall r \in R} \lambda G_{h,r} \geq \sigma B_h$	$\beta \cdot \sum_{\forall r \in R} \lambda G_{h,r} \gg 1$	0 or 1	1
	$\sum_{\forall r \in R} \lambda G_{h,r} \leq \beta \cdot \sigma B_h$	$\beta \cdot \sigma B_h \gg 1$	1	
$\sum_{\forall r \in R} \lambda G_{h,r} = 0$	$\beta \cdot \sum_{\forall r \in R} \lambda G_{h,r} \geq \sigma B_h$	$\beta \cdot \sum_{\forall r \in R} \lambda G_{h,r} = 0$	0	0
	$\sum_{\forall r \in R} \lambda G_{h,r} \leq \beta \cdot \sigma B_h$	$\beta \cdot \sigma B_h = 0$	0 or 1	

20. Location of CNVMs

$$\beta \cdot \lambda B_{p,h} \geq \sigma E_{p,h} \quad \forall p, q \in H, p \neq q \tag{7.7}$$

$$\lambda B_{p,h} \leq \beta \cdot \sigma E_{p,h} \quad \forall p, q \in H, p \neq q \tag{7.8}$$

$$\beta \cdot \sum_{h \in H} \lambda B_{p,h} \geq \sigma E_p \quad \forall p \in H \tag{7.9}$$

$$\sum_{h \in H} \lambda B_{p,h} \leq \beta \cdot \sigma E_p \quad \forall p \in H \tag{7.10}$$

$$\psi_{pq} \leq \sigma E_p \quad \forall p, q \in H, p \neq q \tag{7.11}$$

$$\psi_{p,q} \leq \sigma E_q \quad \forall p,q \in H, p \neq q \tag{7.12}$$

$$\psi_{p,q} \geq \sigma E_p + \sigma E_q - 1 \quad \forall p,q \in H, p \neq q \tag{7.13}$$

Constraints (7.7) and (7.8) ensure that the BBUVMs at node *h* are served by CNVMs that are located at the node *p*. Constraints (6.8) and (6.9) determine the location of CNVMs by setting the binary variable σE_p to 1 if there is a CNVM at node *p*. Constraints (7.11), (7.12) and (7.13) ensure that the CNVMs communicate with each other if they are located at different nodes *p* and *q*. This is equivalent to the logical operation $\psi_{p,q} = \sigma E_p \; AND \; \sigma E_q$.

21. CNVMs intra-traffic

$$\lambda E_{p,q} = \nabla_{p,q} \cdot \psi_{p,q} \quad \forall p,q \in H: p \neq q \tag{7.14}$$

22. Hosting VM of any type (BBUVM or CNVM):

$$\sigma \chi_h \leq \sigma B_h + \sigma E_h \quad \forall\, h \in H \tag{7.15}$$

$$\sigma \chi_h \geq \sigma B_h \quad \forall\, h \in H \tag{7.16}$$

$$\sigma \chi_h \geq \sigma E_h \quad \forall\, h \in H \tag{7.17}$$

Constraint (7.14) represents the traffic between CNVMs at nodes *p* and *q*. Constraints (7.15), (7.16) and (7.17) determine if the node *h* hosts any VM of any type (BBUVM or CNVM). It is equivalent to the logical operation ($\sigma \chi_h = \sigma E_p \; OR \; \sigma E_q$).

23. Video streaming traffic to RRH nodes

$$\sum_{n \in N} \lambda S_{n,r} = \lambda V_r - \sum_{u \in H} \lambda C_{u,r} \quad \forall\, r \in R \tag{7.18}$$

24. Request for video caching

$$\sum_{u \in H} \sigma C_{u,r} \leq 1 \quad \forall\, r \in R \tag{7.19}$$

25. Video (contents) server location

$$\beta \cdot \sum_{r \in R} \lambda S_{n,r} \geq \sigma S_n \quad \forall\, n \in N \tag{7.20}$$

$$\beta \cdot \sum_{r \in R} \lambda S_{n,r} \geq \sigma S_n \quad \forall\, n \in N \tag{7.21}$$

$$\sum_{n \in N} \sigma S_n = 1 \quad \forall\, n \in N \tag{7.22}$$

Constraint (6.2) determines the video streaming traffic sourced by both video (contents) server and cache nodes. Constraint (6.3) determines whether the RRH node *r* is served by a cache at node *h*. Constraints (6.11) and (6.12) determine the location of the video (contents) server, whilst constraint (6.13) ensures it is located at one IP over WDM node.

26. Cache node hit ratio:

$$\Theta_{u,r} \leq \sigma C_{u,r} \quad \forall\, u \in H, \forall r \in R \tag{7.23}$$

$$\Theta_{u,r} \leq \delta_u \quad \forall\, u \in H, \forall r \in R \tag{7.24}$$

$$\Theta_{u,r} \geq \delta_u - (1 - \sigma C_{u,r}) \quad \forall\, u \in H, \forall r \in R \tag{7.25}$$

$$\Theta_{u,r} \geq 0 \quad \forall\, u \in H, \forall r \in R \tag{7.26}$$

$$\delta_u \leq 1 \quad \forall\, u \in H, \forall r \in R \tag{7.27}$$

Constraints (6.24), (6.25), (6.26), (6.27) and (6.28)) determine the cache hit ratio at any node *h*, where constraints (6.24) and (6.25) are equivalent to the multiplication of the hit ratio by the binary variable $\sigma C_{u,r}$; whilst constraints (6.26) and (6.27) ensure that the hit ratio does not go beyond 1 and is not less than 0. For instance, when the value of $\sigma C_{u,r}$ is 1, the value of $\Theta_{u,r}$ in constraint (6.24) takes any value between 0 and 1 including the value of the hit ratio δ_u, whilst in constraint (6.25) the value of $\Theta_{u,r}$ will be equal or less than the hit ratio δ_u. In constraint (6.26) the value of $\Theta_{u,r}$ is equal or greater than the hit ratio δ_u. In all three cases, the value of the hit ratio δ_u is the only value for $\Theta_{u,r}$ which satisfies the three constraints. In the same way, when the value of $\sigma C_{u,r}$ equals to zero, the value of $\Theta_{u,r}$ is zero. Table 7.5 illustrates the operation of the constraints (6.24) to (6.26).

Table 7.5 illustration of constraints (6.24) to (6.26)

The value of the binary variable $\sigma C_{u,r}$	constraint	outcome	The value of $\Theta_{u,r}$ that satisfies all constraints
$\sigma C_{u,r} = 1$	$\Theta_{u,r} \leq \sigma C_{u,r}$	*any value between* 0 *and* 1	δ_u
	$\Theta_{u,r} \leq \delta_u$	$\Theta_{u,r} = \delta_u$ or $\Theta_{u,r} < \delta_u$	
	$\Theta_{u,r} \geq \delta_u - (1 - \sigma C_{u,r})$	$\Theta_{u,r} = \delta_u$	

		or $\Theta_{u,r} > \delta_u$	
$\sigma C_{u,r} = 0$	$\Theta_{u,r} \leq \sigma C_{u,r}$	$\Theta_{u,r} = 0$	0
	$\Theta_{u,r} \leq \delta_u$	$\Theta_{u,r} = \delta_u$ or $any\ value < \delta_u$	
	$\Theta_{u,r} \geq \delta_u - (1 - \sigma C_{u,r})$	$\delta_u = 0$ or *any positive number*	

27. Traffic from content cache to the RRH nodes

$$\lambda C_{u,r} = \Theta_{ur} \cdot \lambda V_r \quad \forall\, u \in H, \forall r \in R \tag{7.28}$$

28. Size of content cache node

$$CZ_u \geq \delta_u \cdot a_k + b_k \quad \forall\, u \in H, \forall k \in K \tag{7.29}$$

$$CZ_u = iCZ_u + fCZ_u \quad \forall\, u \in H \tag{7.30}$$

Constraint (6.4) determines the amount of traffic from the cache node to the RRH based on the cache hit ratio. Constraint (6.7) calculates the cache size based on the hit ratio using a piecewise linear approximation. Constraints (6.29) rounds down the cache size (ceiling) to the nearest integer.

29. Cache traffic that passes through BBUVM:

$$\lambda C_{u,h,r} \leq \mu \cdot \sigma B_{h,r} \quad \forall\, u, h \in H, \forall r \in R \tag{7.31}$$

$$\lambda C_{u,h,r} \leq \lambda C_{u,r} \quad \forall\, u, h \in H, \forall r \in R \tag{7.32}$$

$$\lambda C_{u,h,r} \geq \lambda C_{u,r} - \mu \cdot (1 - \sigma B_{h,r}) \quad \forall\, u, h \in H, \forall r \in R \tag{7.33}$$

$$\lambda C_{uhr} \geq 0 \quad \forall\, u, h \in H, \forall r \in R \tag{7.34}$$

Constraints (7.31), (7.32), (7.33) and (7.34) ensure that the traffic from the cache node *u* to the RRH node *r* passes through the BBUVMs that serve the RRH node *r* where μ is a large number. The operation of constraints (7.31), (7.32) and (7.33) is illustrated in Table 7.6.

Table 7.6 illustration of constraints (7.31)(7.32)(7.33)

The value of the binary variable $\sigma B_{h,r}$	constraint	outcome	The value of $\lambda C_{u,h,r}$ that satisfies all constraints
$\sigma B_{h,r} = 1$	$\lambda C_{u,h,r} \leq \mu \cdot \sigma B_{h,r}$	*any value between 0 and* μ	$\lambda C_{u,r}$
	$\lambda C_{u,h,r} \leq \lambda C_{u,r}$	$\lambda C_{u,h,r} = \lambda C_{u,r}$ or $\lambda C_{u,h,r} < \lambda C_{u,r}$	
	$\lambda C_{u,h,r} \geq \lambda C_{u,r} - \mu \cdot (1 - \sigma B_{h,r})$	$\lambda C_{u,h,r} = \lambda C_{u,r}$ or $\lambda C_{u,h,r} > \lambda C_{u,r}$	
$\sigma B_{h,r} = 0$	$\lambda C_{u,h,r} \leq \mu \cdot \sigma B_{h,r}$	$\lambda C_{u,h,r} = 0$	0
	$\lambda C_{u,h,r} \leq \lambda C_{u,r}$	$\lambda C_{u,h,r} = \lambda C_{u,r}$ or *any value* $< \lambda C_{u,r}$	
	$\lambda C_{u,h,r} \geq \lambda C_{u,r} - \mu \cdot (1 - \sigma B_{h,r})$	$\lambda C_{u,h,r} = 0$	

		or *any positive number*	

30. Video (contents) server traffic which passes through BBUVMs:

$$\lambda S_{n,h,r} \leq \mu \cdot \boldsymbol{\sigma} B_{h,r} \quad \forall s \in N, \forall h \in H, \forall r \in R \tag{7.35}$$

$$\lambda S_{n,h,r} \leq \lambda S_{n,r} \quad \forall s \in N, \forall h \in H, \forall r \in R \tag{7.36}$$

$$\lambda S_{n,h,r} \geq \lambda S_{n,r} - \mu \cdot \left(1 - \boldsymbol{\sigma} B_{h,r}\right) \quad \forall s \in N, \forall h \in H, \forall r \in R \tag{7.37}$$

$$\lambda S_{n,h,r} \geq 0 \quad \forall s \in N, \forall h \in H, \forall r \in R \tag{7.38}$$

Constraints (7.35), (7.36), (7.37) and (7.38) ensure that the traffic from contents (video) server *n* to the RRH node *r* passes through the BBUVMs that serve the RRH node *r* where μ is a large number. The operation of these constraints is the same as the constraints (7.31), (7.32), (7.33) and (7.34) explained earlier and summarised in Table 7.6.

31. Total download traffic from BBUVMs to RRH nodes:

$$\lambda R_{h,r} = \lambda G_{h,r} + \sum_{\forall u \in H} \lambda C_{u,h,r} + \sum_{\forall n \in N} \lambda S_{n,h,r} \quad \forall h \in H, \forall r \in R \tag{7.39}$$

32. Total Traffic between two hosting nodes:

$$\lambda T_{p,q} = \lambda E_{p,q} + \lambda B_{p,q} + \alpha \cdot \sum_{r \in R} \lambda C_{p,q,r} \tag{7.40}$$

$$\forall p \in U \cup L, \forall q \in H, p \neq q$$

$$\lambda T_{p,q} = \lambda E_{p,q} + \lambda B_{p,q} + \alpha \cdot \left(\sum_{r \in RRH} \lambda C_{p,q,r} + \sum_{r \in RRH} \lambda S_{p,q,r} \right) \quad (7.41)$$
$$\forall p \in N, \forall q \in H, p \neq q$$

Constraint (6.14) calculates the total download traffic from BBUVMs at node h to the RRH node r which is sourced by three nodes. These are: video server ($\lambda S_{n,h,r}$), cache node ($\lambda C_{u,h,r}$), and the regular traffic from CNVMs ($\lambda G_{h,r}$). Constraints (7.40), (7.41) calculate the total download traffic between any two hosting nodes.

33. Total BBU workload at hosting node h:

$$\Psi B_h = \left(\left(\sum_{\forall r \in R} \lambda R_{h,r} \right) \Big/ cp \right) \cdot \Psi \mathrm{X} \quad (7.42)$$
$$\forall h \in H$$

34. Total normalised workload at hosting node h:

$$\Psi i_h + \Psi f_h = (\Psi B_h + \Psi C)/\Psi S \quad (7.43)$$
$$\forall h \in H$$

35. Hosting node capacity

$$\left(\Omega Sd \cdot (\Psi i_h + \sigma \chi_h) + \Psi f_h \cdot (\Omega S - \Omega Sd) \right) \leq \Omega H_h \quad (7.44)$$
$$\forall h \in H$$

Constraint (7.42) calculates the total BBU workload needed for the total download traffic toward RRH nodes, whilst constraint (7.43) determines the fractional and integer parts of the normalised workload. Constraint (7.44) ensures that the total workload of a VM hosting server does not exceed the total capacity of the hosting node.

36. Flow conservation of total downlink traffic from BBUVMs to RRH nodes:

$$\sum_{y \in TN_x} \lambda R_{hr,xy} - \sum_{y \in TN_x} \lambda R_{hr,yx} = \begin{cases} \lambda R_{hr} & if\ x = h \\ -\lambda R_{hr} & if\ x = r \\ 0 & otherwise \end{cases} \quad (7.45)$$

$$\forall h \in H, \forall r \in RRH, \forall x \in TN_x$$

37. Flow conservation of total downlink traffic between two hosting nodes

$$\sum_{y \in TN_x \cap H} \lambda T_{x,y}^{p,q} - \sum_{y \in TN_x \cap H} \lambda T_{y,x}^{p,q} = \begin{cases} \lambda T_{p,q} & if\ x = p \\ -\lambda T_{p,q} & if\ x = q \\ 0 & otherwise \end{cases} \quad (7.46)$$
$$\forall p, q, x \in H : p \neq q$$

38. GPON link constraints

$$\sum_{h \in H} \sum_{r \in R} \sum_{j \in TN_i \cap L} \lambda R_{i,j}^{h,r} + \sum_{p \in H} \sum_{q \in H, q \neq p} \sum_{j \in TN_i \cap L} \lambda T_{i,j}^{p,q} \leq 0 \quad (7.47)$$
$$\forall i \in U$$

$$\sum_{h \in H} \sum_{r \in R} \sum_{j \in \mathrm{TN_i} \cap \mathrm{N}} \lambda R_{i,j}^{h,r} + \sum_{p \in H} \sum_{q \in H, q \neq p} \sum_{j \in \mathrm{TN_i} \cap \mathrm{N}} \lambda T_{i,j}^{p,q} \leq 0 \quad (7.48)$$
$$\forall i \in L$$

Constraint (6.16) represents the flow conservation of total download traffic toward RRH nodes, whilst constraint (7.46) represent the flow conservation of total download traffic between two hosting nodes. Constraints (6.17) and (6.18) ensure that the download traffic of GPONs does not flow in the opposite direction.

39. Virtual link capacity of IP over WDM network

$$\sum_{p \in H} \sum_{q \in H, q \neq p} \lambda T_{i,j}^{p,q} + \sum_{h \in H} \sum_{r \in R} \lambda R_{i,j}^{h,r} \leq W_{i,j} \cdot B \quad (7.49)$$
$$\forall i, j \in N, i \neq \mathrm{j}.$$

40. Flow conservation in the optical layer of IP over WDM network:

$$\sum_{n \in NN_m} W_{m,n}^{i,j} - \sum_{n \in NN_m} W_{n,m}^{i,j} = \begin{cases} W_{i,j} & if\ n = i \\ -W_{i,j} & if\ n = j \\ 0 & otherwise \end{cases} \quad (7.50)$$
$$\forall i, j, m \in N, i \neq j$$

Constraint (6.19) ensures that the total traffic that travers the virtual link (i, j) does not exceed its capacity, in addition it determines the number of

wavelength channels that carry the traffic burden of that link. Constraint (6.20) represents the flow conservation in the optical layer of the IP over WDM network. It ensures that the total expected number of incoming wavelengths for the IP over WDM nodes of the virtual link (*i*, *j*) is equal to the total number of outgoing wavelengths of that link.

41. Number of wavelength channels

$$\sum_{i \in N} \sum_{j \in N: i \neq j} W_{m,n}^{i,j} \leq w \cdot f_{m,n} \qquad \forall m \in N, \forall n \in NN_m \tag{7.51}$$

$$W_{m,n} = \sum_{i \in N} \sum_{j \in N: i \neq j} W_{m,n}^{i,j} \qquad \forall m \in N, \forall n \in NN_m \tag{7.52}$$

42. Number of aggregation ports

$$\Lambda_i = \left(\sum_{j \in L \cap TN_i} \left(\sum_{p \in H} \sum_{q \in H, q \neq p} \lambda T_{i,j}^{p,q} + \sum_{h \in H} \sum_{r \in R} \lambda R_{i,j}^{h,r} \right) \right) \Big/ B \qquad \forall i \in N \tag{7.53}$$

Constraints (6.21) and (6.22) are the constraints of the physical link (*m*, *n*). Constraint (6.21) ensures that the total number of wavelength channels in the logical link (*i*, *j*) that traverses the physical link (*m*, *n*) do not exceed the fibre capacity. Constraint (6.23) determines the number of wavelength channels in the physical link and ensures that it is equal to the total number of wavelength channels in the virtual link traversing that physical link. Constraint (6.23) determines the required number of aggregation ports in each IP over WDM router.

7.3 MILP model setup and results

This section describes the network topology considered and provides the input parameters used in the developed MILP model. In addition, the MILP model results are discussed in detail.

7.3.1 Network topology and input parameters

The network topology considered is shown in Figure 7.2. This topology consists of 3 IP over WDM nodes and 6 GPON networks connected in pairs to the IP over WDM nodes; two GPON networks for each IP over WDM node. Each GPON network consists of one OLT and three ONUs and each ONU is connected to one RRH node. The topology has one video server whose location in the IP over WDM network is optimised by the developed MILP model to minimise the total power consumption. Each RRH node in the network supports a small cell with maximum number of users equals to 10. Each user in the small cell is allocated 5 physical resource blocks (PRB) as the users are assumed to request the same task from the network. The average number of users in the network varies over the time of day according to the network user profile shown in previous chapter. Therefore, the amount of downlink traffic to each RRH node is influenced by the total number of active users in the small cell where its maximum value is considered less than 10 Gbps. The total download traffic from both video server and cache nodes to the RRH nodes is considered 80% of the total download traffic as explained in the previous chapter. Two virtualisation approaches were considered; with and without CNVMs inter-traffic. In addition, the content caching approach was considered with variable cache size where the developed MILP model optimised the size of the

cache at each node. The input parameters to the developed MILP model are listed in Table 7.7.

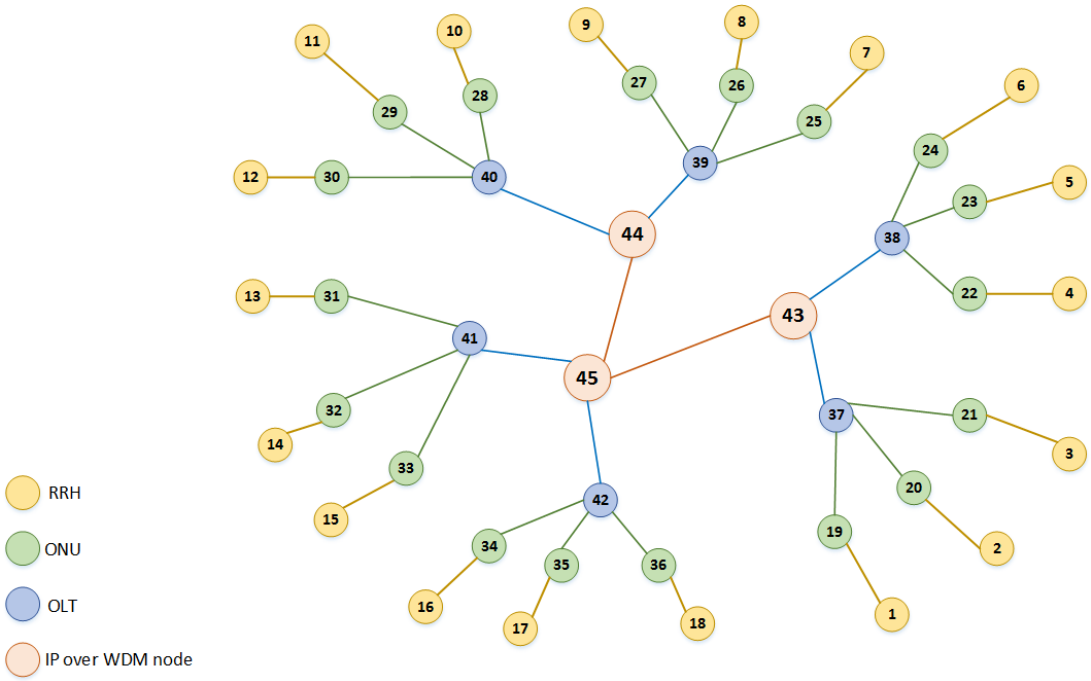


Figure 7.2 Tested network topology

Table 7.7 MILP model input parameters

Line coding rate for 8B/10B line coding (l)	10/8 (bit / sample)
Number of MIMO layers (y)	2
Number of bits used in QAM modulation for 64 QAM modulation (q)	6 (bits)
Number of antennas in a cell (a)	2
Maximum fronthaul (CPRI) data rate for CPRI line rate option 7 (cp)	9.8304 (Gbps) [256]
Maximum baseband processing workload deeded for full loaded RRH (ΨX) given by: $30 \cdot a + 10 \cdot a^2 + 20 \cdot q \cdot l \cdot y$	400 (GOPS)
Server CPU maximum workload (ΨS)	368 (GOPS) [265]
Workload needed for hosting one CNVM (ΨC)	26.17 (GOPS)
Number of active users in a small cell (ρ_r)	Uniformly distributed (1-10 users)

Maximum number of users per cell (n)	10 (users)
Number of physical resources blocks per user (pb)	5 (PRB)
The ratio of the backhaul to the front haul traffic (α)	0.1344 (unitless)
ONU maximum power consumption (ΩU)	15 (W) [233]
OLT maximum power consumption (ΩL)	1940 (W) [231]
OLT idle power (ΩLd)	60 (W) [231]
OLT maximum capacity (CL)	8600 (Gbps) [231]
ONU maximum capacity (CU)	10 (Gbps) [233]
RRH node power consumption (ΩR_x)	1140 (W) [234]
Hosting server maximum power consumption (ΩS)	365 (W) [266]
Hosting server idle power (ΩSd)	112 (W) [266]
Cache node maximum power consumption (ΩC)	550 (W) [269]
Cache node maximum storage capacity (CC)	14.4 (TB) [269]
Video streaming server energy per bit (εs)	211.1 (Joul/Gb) [260]
Capacity IP over WDM wavelength channel (B)	40 (Gbps) [236]
Number of wavelength per fibre in IP over WDM (w)	32 [236]
Transponder power consumption (ΩT)	167 (W) [237]
Router port power consumption (ΩRP)	825 (W) [172]
Regenerator power consumption (ΩG)	334 (W) [172]
EDFA power consumption (ΩE)	55 (W) [172]
Maximum span distance between EDFAs (S)	80 (km) [236]

7.3.2 MILP model results

A MILP model was developed to minimise the total power consumption by optimising the caches sizes, VM servers' utilisation, and the location of both VMs and caches at each node. The results compare the utilisation of virtualisation and content caching in 5G individually and investigate the total power consumption of each approach. In addition, the impact of integrating content caching and

virtualisation is compared with the impact of the deployment of each technology individually. Figure 7.3 illustrates the total power consumption of caching-only, virtualisation-only (with and without CNVMs inter-traffic), and the integrated (integrated caching and virtualisation) approaches for different times of the day, whilst Figure 7.4 illustrates the total power consumption for the same approaches for different number of users.

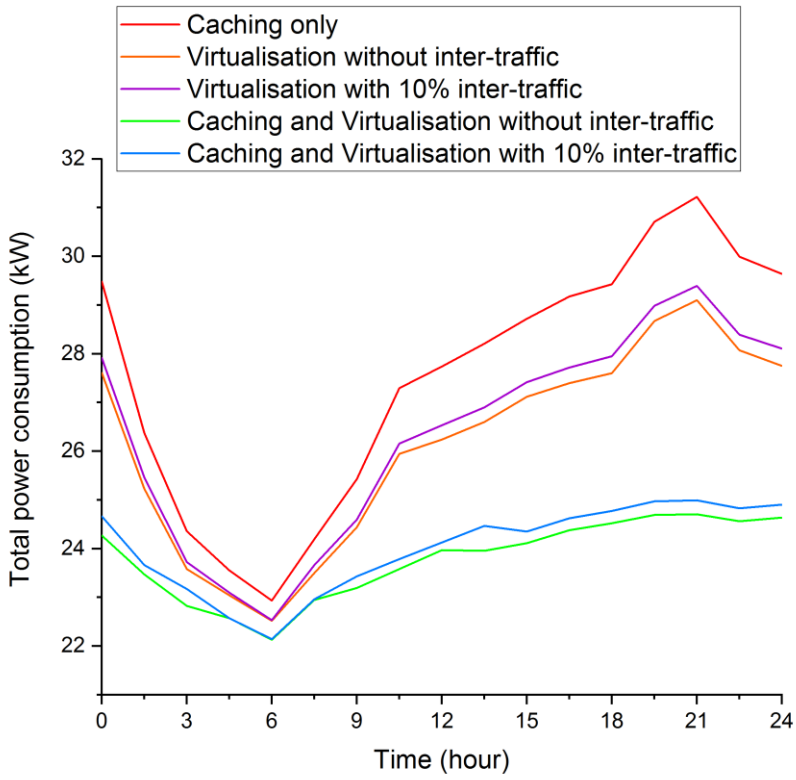


Figure 7.3 Total power consumption of different approaches at different times of the day

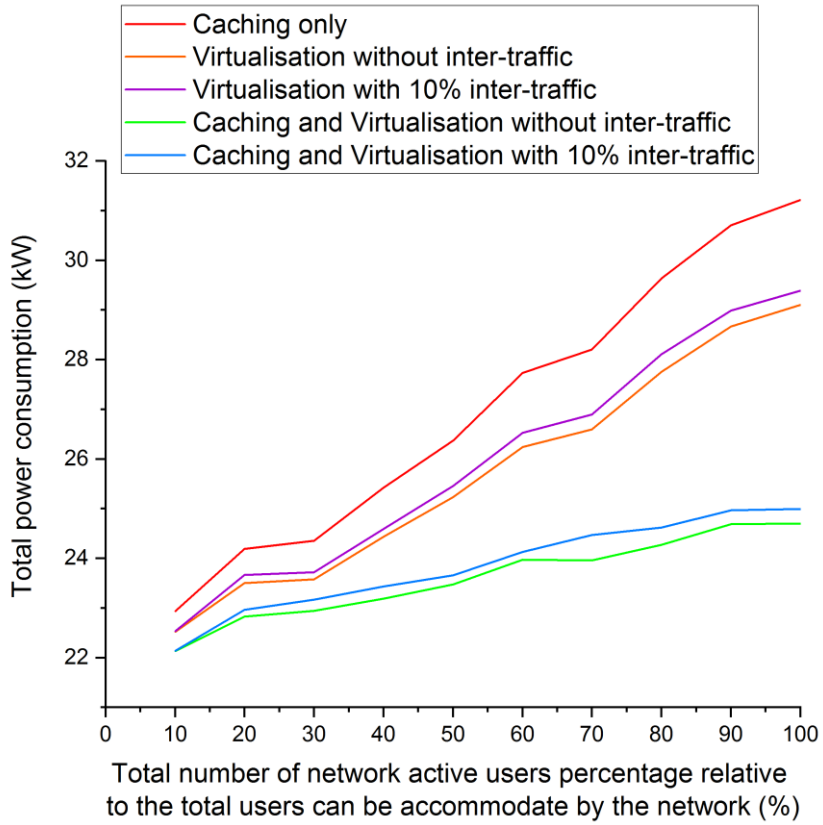


Figure 7.4 Total power consumption of different approaches for different number of users

The caching-only approach has higher power consumption compared to the other approaches whilst the integrated caching and virtualisation approaches without CNVMs inter-traffic have the lowest power consumption among all the approaches.

Although the virtualisation only approaches (with and without CNVMs inter-traffic) have high power consumption compared to the approach where the caching and virtualisation are integrated (integrated approach), they have less power consumption compared to the caching-only approach. This is attributed to the fact that the virtualisation-only approach achieves much lower mobile functions power consumption compared to the caching-only approach as shown in Figure 7.5. Therefore, the total power consumption of caching-only approach is higher than the

virtualisation-only approach in spite of its lower video streaming service power consumption compared to the virtualisation-only approach as shown in Figure 7.6.

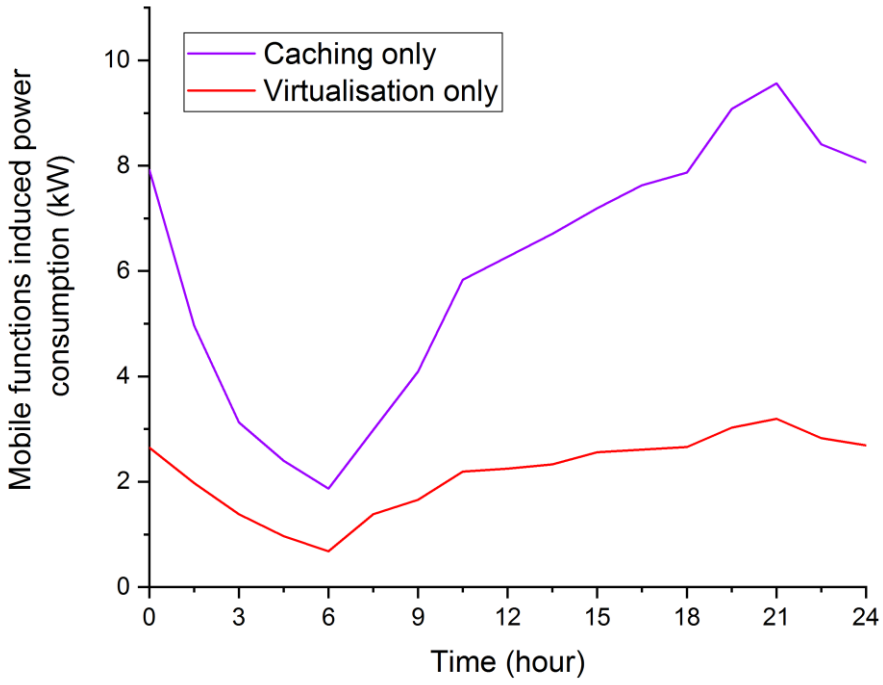


Figure 7.5 Mobile function induced power consumption of caching and virtualisation only approaches

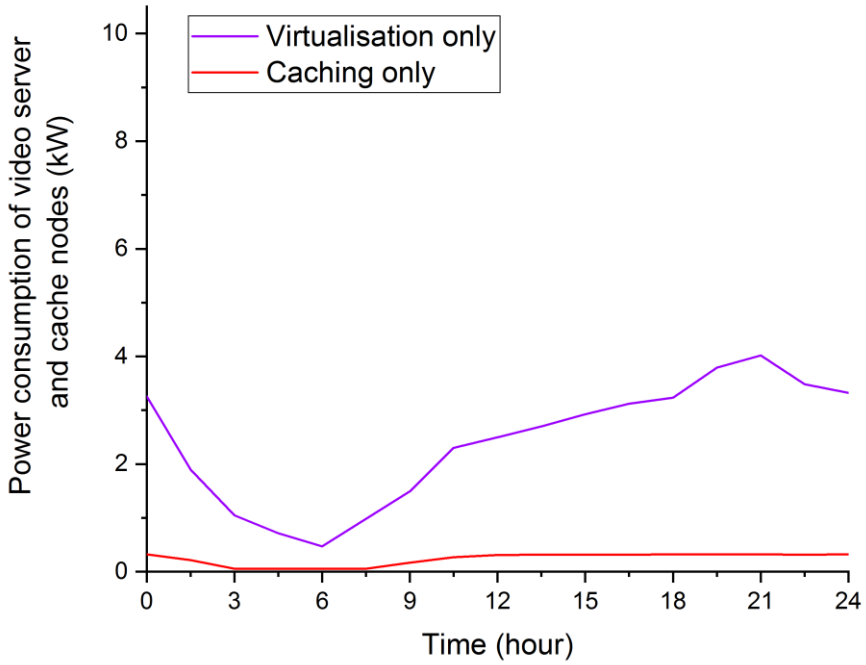
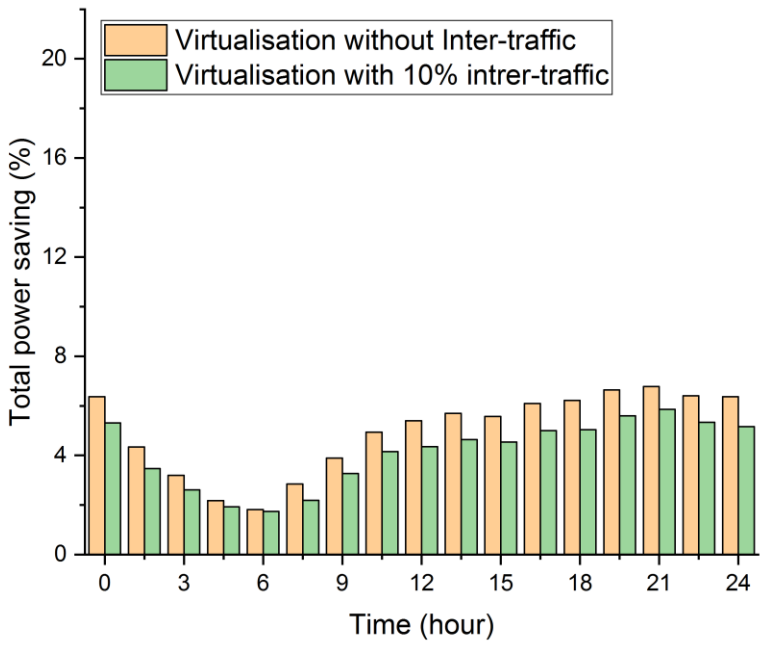
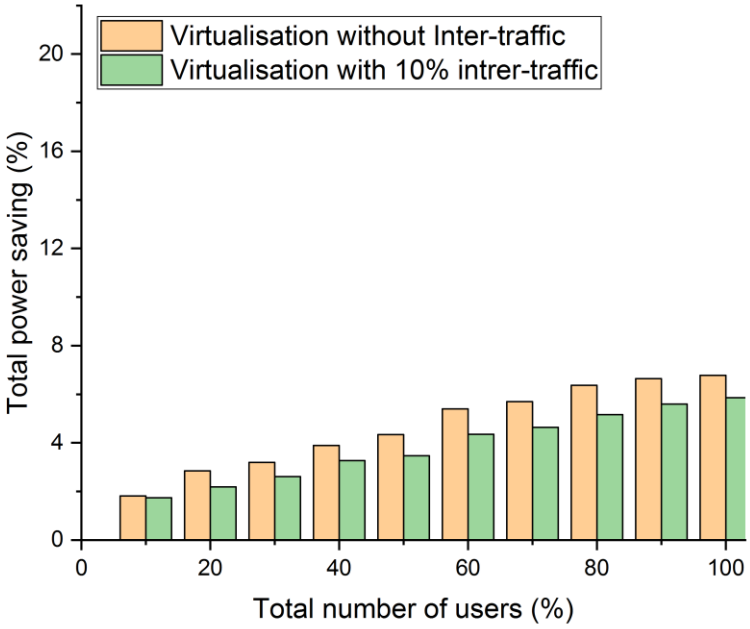


Figure 7.6 Video streaming service power consumption in caching-only and virtualisation-only approaches

Figure 7.7 illustrates the saving in total power consumption of virtualisation-only approach compared to the caching-only approach for different times of the day and different number of users. Compared to the caching-only approach, the virtualisation-only approach achieves maximum total power saving of 7% (average 5%) when no CNVMs inter-traffic is considered and 6% (average 4%) with CNVMs inter-traffic of 10% of the total backhaul traffic. According to these findings, the utilisation of virtualisation is better than content caching in 5G mobile networks in term of energy saving and power consumption.



a) Power saving at different times of the day

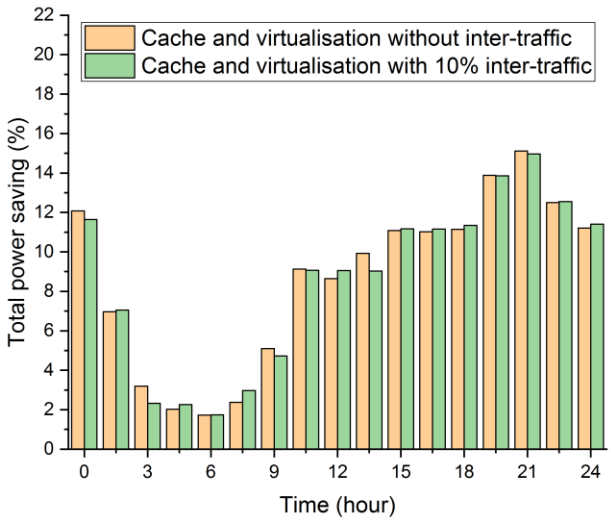


b) Power saving at different number of users

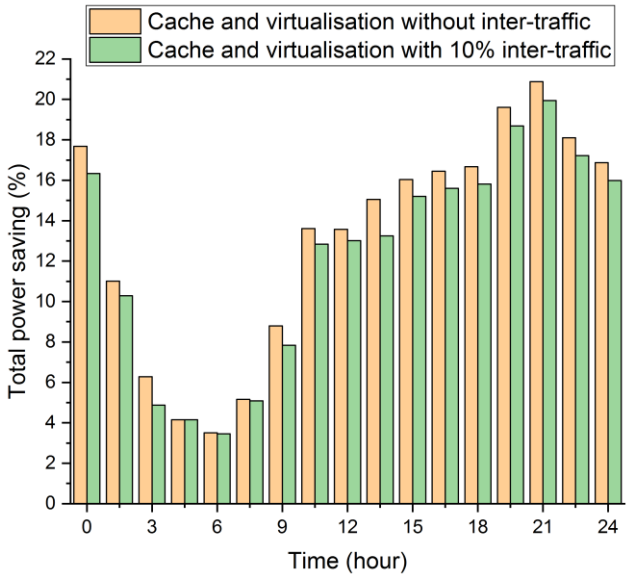
Figure 7.7 Power saving of virtualisation-only compared with caching-only approach

The deployment of virtualisation and content caching in one integrated architecture has a great impact on the total power consumption. Figure 7.8 illustrates the total power saving of the integrated approach compared with the virtualisation-only and caching-only approaches. Compared with the virtualisation-only approach, the integrated approach has a maximum total power saving of 15% (average 9%) with and without CNVMs inter-traffic. This is due to the lower video streaming service power consumption of the integrated approach compared to the virtualisation only approach (as shown in Figure 7.9), where a maximum video streaming service power saving of 95% (average 90%) is achieved by the integrated approach compared with virtualisation-only approach.

Figure 7.10 depicts the integrated approach optimum cache utilisation of each node at different times of the day when there is no CNVMs inter-traffic and also for CNVMs inter-traffic of 10% of the total backhaul traffic. The optimum cache size at each node varies with the delivered traffic from the node over time. The optimum cache size is relatively high when the total number of users is high during the busy hours of the day and the caches are distributed close to the users. The OLT nodes are highly utilised by caches during the busy times of the day whilst the IP over WDM nodes are utilised when few users are active. During the busy time of the day, there is a large number of active users and therefore the demand for video streaming is high. In this case the integrated approach accommodates a large amount of the video files (objects) at OLT nodes to serve as much as possible users and offload the traffic away from the core network whilst in the virtualisation-only approach, the users are served by the video server in the core network.



a) Total power saving compared to virtualisation-only approach



b) Total power saving compared to caching-only approach

Figure 7.8 Power saving of integrated virtualisation and caching approach compared with other approaches at different time of the day

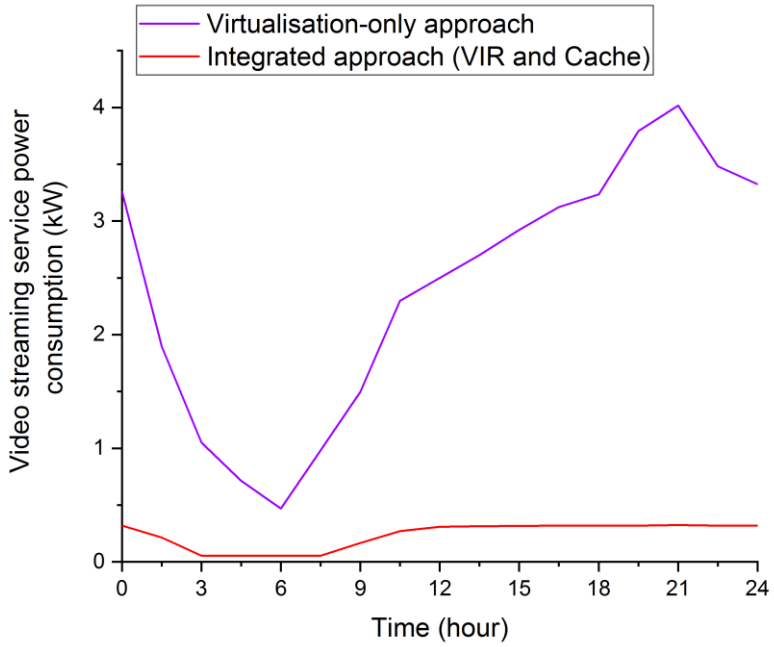
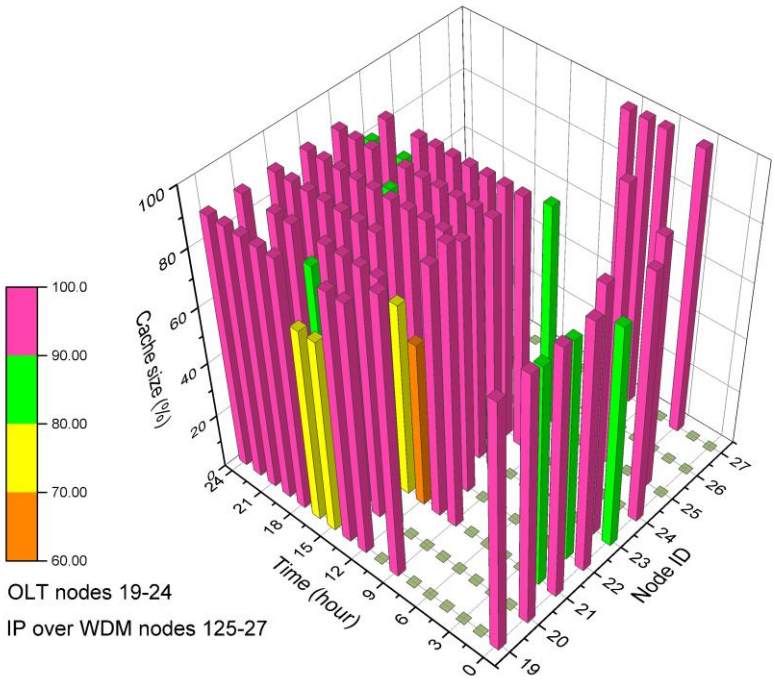
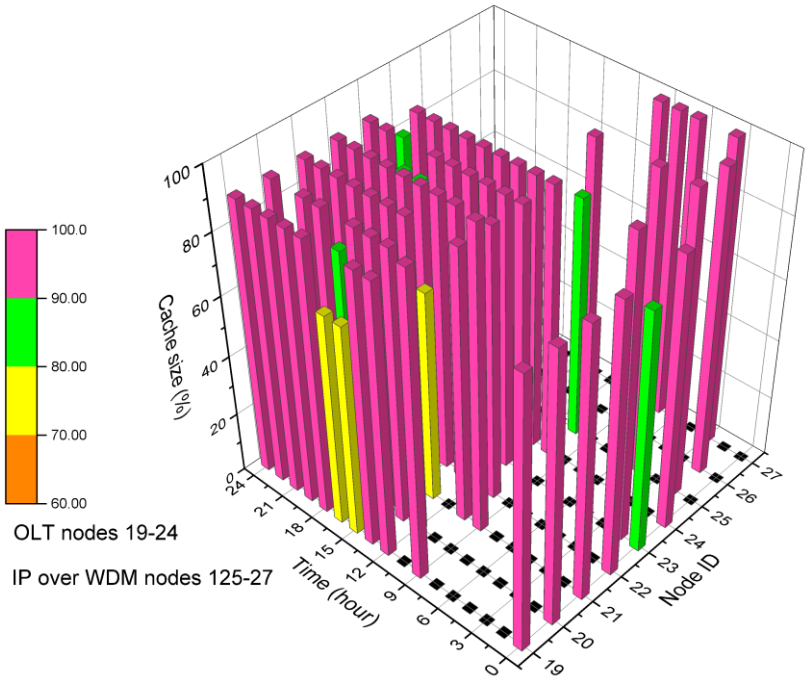


Figure 7.9 Video streaming service power consumption for virtualisation-only and integrated approaches at different times of the day



a) Without CNVMs inter-traffic



b) With CNVMs inter-traffic 10% of the total backhaul traffic

Figure 7.10 Cache utilisation of each nodes at different times of the day

The integrated approach achieves a maximum total power saving of 21% (average 13%) without CNVMs inter-traffic and 20% (average 12%) when CNVMs inter-traffic is considered compared with the caching-only approach. Additionally, the integrated approach has a maximum mobile function power saving of 65% (average 57%) without CNVMs inter-traffic and 70% (average 58%) when CNVMs inter-traffic is considered 10% of the total backhaul traffic as compared with the caching-only approach. This is driven by the higher mobile functions power consumption in the caching-only approach which is much higher than the integrated approach as shown in Figure 7.11. It is worth mentioning that compared to Figure 7.5, the integrated approach (Caching and virtualisation) in Figure 7.11 does not

provide improvement over the results of virtualisation only (Figure 7.5) as the traffic in Figure 7.11 does not include video traffic.

Figure 7.12 depicts the integrated approach optimum VM (BBUVM and CNVM) utilisation of each node at different times of the day when the CNVMs inter-traffic is 0% and 10% of the total backhaul traffic. The optimum utilisation of each node varies with the traffic delivered from the node at a certain time. The optimum VM (BBUVM and CNVM) workload at each node is relatively high when the total number of users served by the VM is high during the busy hours of the day where the VMs are distributed close to the small cells. The optimum VM workload and location are mainly driven by: the total number of active users, the amount of offloaded backhaul traffic, and the inter-traffic between CNVMs. It is clearly seen in Figure 7.12 that in the integrated approach, the OLT nodes are highly utilised by VMs during the busy time of the day to maximise the number of severed users by each VM and minimise the backhaul traffic carried over the core network. When the total number of users is low during off peak times of the day, IP over WDM nodes are utilised by VMs. In this case the traffic induced power consumption is low compared to the VM power consumption therefore, the total number of active users is served by a small number of VMs located far from the cells specifically in the IP over WDM network nodes.

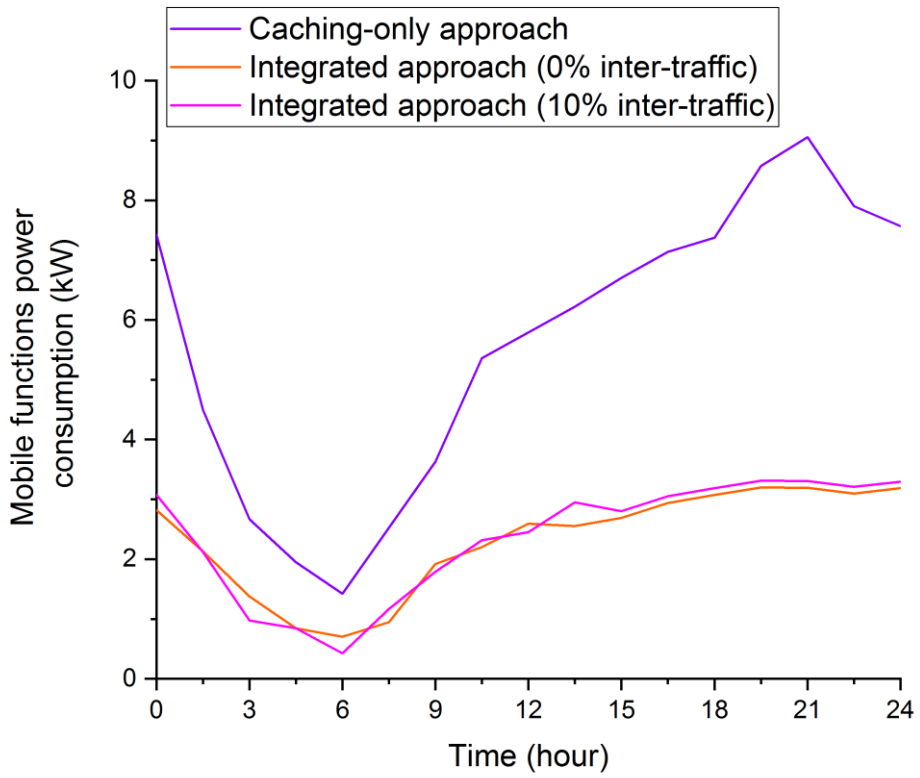
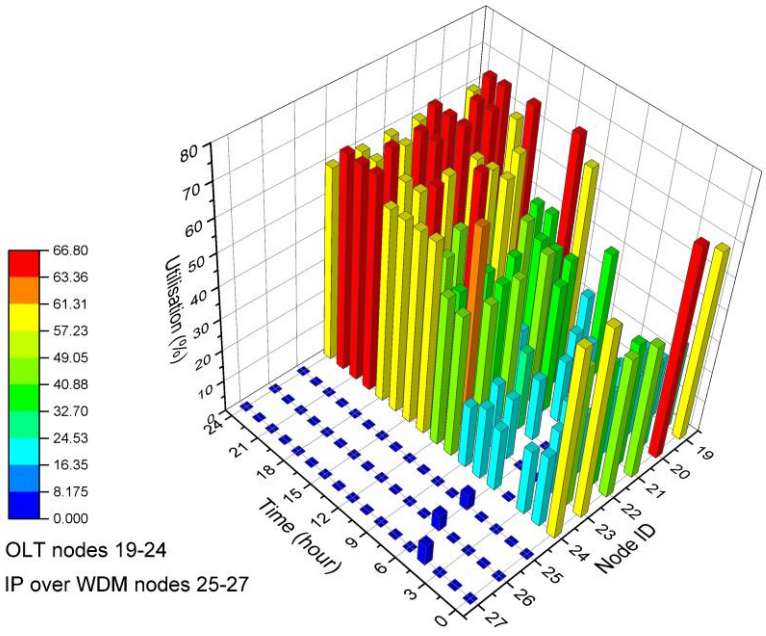
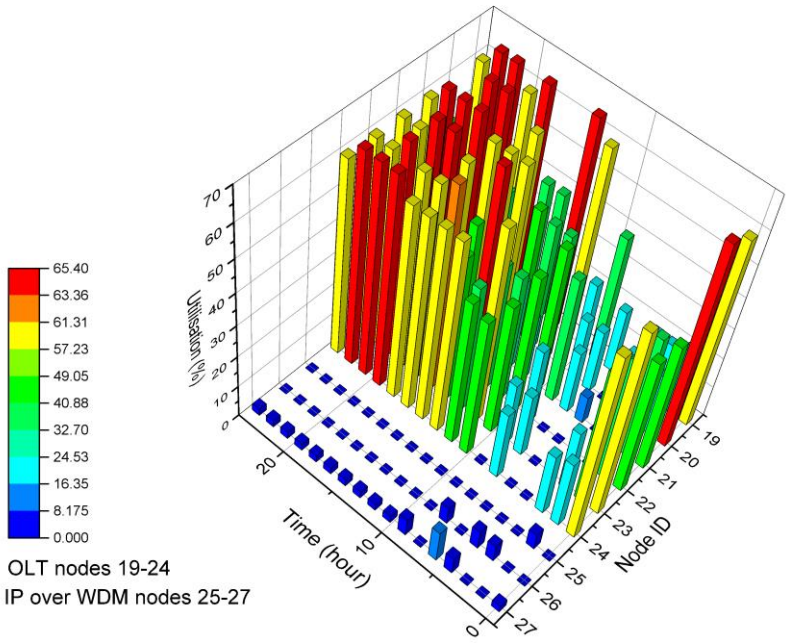


Figure 7.11 Mobile functions power consumption of caching-only and integrated approaches at different times of the day



a) Without CNVMs inter-traffic



b) With CNVMs inter-traffic 10% of backhaul traffic

Figure 7.12 VM utilisation of each nodes at different times of the day

7.4 Real-time Energy-Efficient Virtualisation and content Caching (EEVIRandCa) heuristic

This section introduces an Energy-Efficient Virtualisation and content caching (EEVIRandCa) heuristic approach for real-time implementation of the developed MILP model. The pseudocode of EEVIRandCa heuristic is shown in Algorithm A.5. The network is modelled by sets of network elements *NE*, and links *L*. The heuristic obtains the network topology $G = (NE, L)$ and the physical topology of the IP over WDM network $G_p = (N, L_p)$, where N is the set of IP over WDM nodes and L_p is the set of physical links. Network elements *NE* in the model are defined by group of sets. These sets are *RRH*, *ONU*, *OLT*, and IP over WDM nodes *N*. In addition to these sets, there are three network elements that contribute to the power consumption which are video streaming servers, cache nodes, and VM servers. The total download request (fronthaul traffic) by each eNodeB node is calculated based on the total number of active users in each cell (RRH node). Both video streaming and regular traffic are determined according to Cisco VNI data, whilst the cache size of each node is determined based on the optimised hit ratio of each cache node. The EEVIRandCa heuristic starts accommodating BBUVMs in such a way that they serve as many RRH requests as possible. The model examines the OLT nodes to determine the closest OLT for each RRH node to accommodate a BBUVM.

To accommodate CNVMs, the EEVIRandCa heuristic sets an initial value for the number of CNVMs in the IP over WDM network. The initial number of CNVMs in the IP over WDM network is reduced by one every time the heuristic assesses the IP over WDM network power consumption to determine the optimum number of

CNVMs. To determine the highly recommended nodes for accommodating CNVMs, the heuristic builds a sorted list of IP over WDM nodes based on the total number of users connected directly to each node. The first node at the top of the sorted list of IP over WDM nodes is assigned by the heuristic to accommodate the video streaming servers.

Using the same methodology, the EEVIRandCa heuristic determines the location of the cache. It set an initial value for the maximum number of cache locations in the IP over WDM network and decreases this value every time it evaluates the IP over WDM network power consumption to obtain the best value for the number of caches in the IP over WDM network. The heuristic calculates the total number of active users and compares it to the maximum network capacity. If the total number of active users is less than 50% of the total network capacity, the EEVIRandCa heuristic examines the IP over WDM network otherwise it examines the OLT nodes to cache the contents.

Once the CNVMs, BBUVM, video servers, and cache nodes location are determined, the traffic from these entities is determined and routed toward the RRH nodes. The traffic from CNVMs, video servers and the cache nodes should pass through BBUVMs for BBU processing. In addition, the inter-traffic between CNVMs and total traffic flows in the IP over WDM network are determined.

The EEVIRandCa heuristic obtains the physical graph $G_p = (N, L_p)$ and determines the traffic in each network segment. The IP over WDM network configuration such as the number of fibres, router ports, and the number of EDFAs is determined and the total power consumption is calculated.

7.4.1 EEVIRandCa heuristic model results

For fair validation of the integrated approach MILP model, the network topology in Figure 7.2 is used by the EEVIRandCa heuristic. In addition, the parameters considered in the integrated approach MILP models such as the wireless bandwidth, number of resources blocks per user, and the parameters listed in Table 7.7 are considered in the heuristic. The number of users allocated to each cell in the heuristic is the same as in the MILP model to ensure the requested traffic by each RRH node is the same in both approaches.

Figure 7.13 compares the total power consumption of the EEVIRandCa heuristic with the integrated approach MILP model at different times of the day when there is no CNVMs inter-traffic flows between the VMs. It is clearly seen that the EEVIRandCa heuristic has a higher power consumption compared to the integrated approach MILP model. The difference in power consumption between the two models varies according to the total number of users during the day and is 3.3% max (2% average).

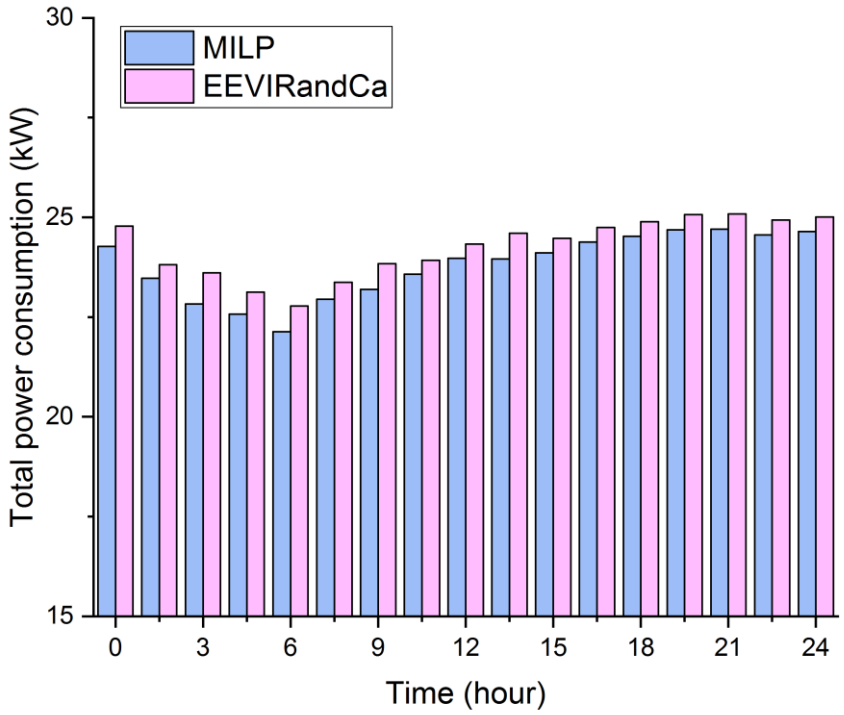


Figure 7.13 Total power consumption of EEVIRandCa heuristic model compared with the integrated approach MILP model when no CNVMs is considered

Figure 7.14 compares the total power consumption of the EEVIRandCa heuristic with the integrated approach MILP model when the CNVMs inter-traffic is 10% of the total backhaul traffic recorded at different times of the day. The EEVIRandCa heuristic successfully mimics the integrated approach MILP model when the CNVMs inter-traffic is 10%. The difference in power consumption between the two models varies during the day and is 2% max (average 1%).

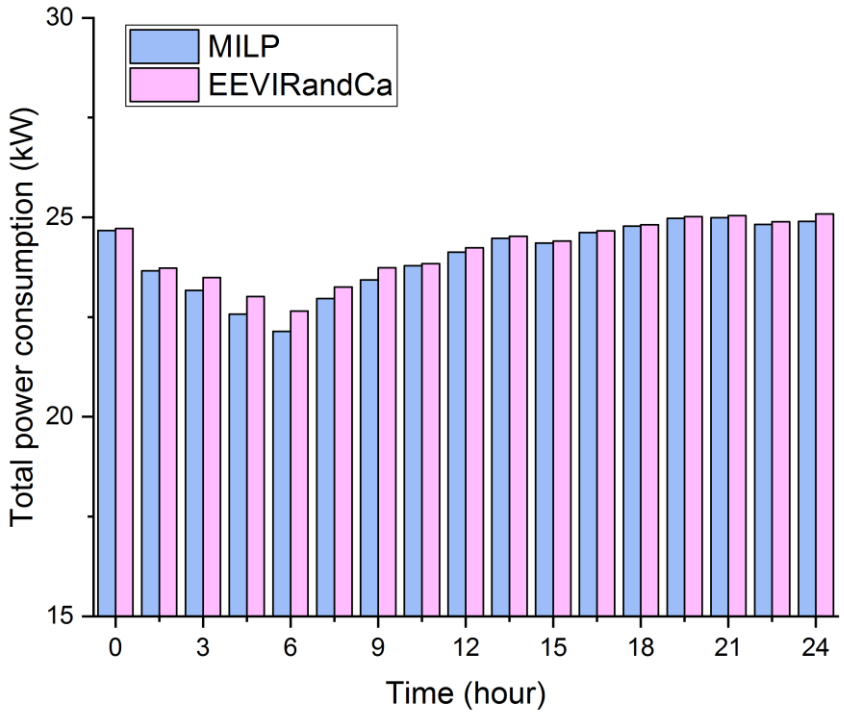


Figure 7.14 Total power consumption of EEVIRandCa compared to the integrated approach MILP model when CNVMs inter-traffic is 10% of the total backhaul traffic

7.5 Summary

NFV and content caching are two promises technologies recognised by the network operators and designers. Integrating these technologies together can leverage the merits of both technologies.

This chapter has compared the deployment of NFV and content caching in 5G networks and evaluated the associated power consumption. In addition, this chapter

has introduced an approach to combine content caching with NFV in one integrated architecture for 5G networks and has evaluated the associated power consumption. A MILP model was developed to minimise the total power consumption by jointly optimising the cache size, VM workload, and the locations of both cache nodes and VMs. The results of the developed model are investigated under the impact of CNVMs inter-traffic. The result show that the OLT nodes are the optimum location for content caching and hosting VMs during busy times of the day whilst IP over WDM nodes are the optimum locations for caching and virtualisation during off-peak time.

The results show that the virtualisation-only approach is better than caching-only approach for video streaming services where the virtualisation-only approach compared to caching-only approach, achieves a maximum power saving of 7% (average 5%) when no CNVMs inter-traffic is considered and 6% (average 4%) with CNVMs inter-traffic 10% of the total backhaul traffic. On the other hand, the integrated approach has a maximum power saving of 15% (average 9%) with and without CNVMs inter-traffic compared to the virtualisation-only approach, and it achieves a maximum power saving of 21% (average 13%) without CNVMs inter-traffic and 20% (average 12%) when CNVMs inter-traffic is considered compared with the caching-only approach.

In order to validate the MILP models and achieve real-time operation of our approaces, a EEVIRandCa heuristic was developed. The heuristic results are compared to the integrated approach MILP model with and without CNVMs inter-traffic. The heuristic and the MILP model results are in close agreement.

Chapter 8
Summary, Conclusions, and Future Work

This chapter summarises the work presented in this thesis and specifies the original contributions. In addition, this chapter suggests potential new directions for future research that could be conducted as a result of the work presented in this thesis.

8.1 Summary of contributions

This thesis investigated the high energy utilisation in 5G mobile networks and the energy efficiency challenges. It also introduced and evaluated possible solutions to cope with these challenges through the development of MILP models that were verified by real-time heuristics.

The first contribution in this thesis was to develop a new optical-based architecture for 5G networks that exploits the benefits of network function virtualisation to bring about energy efficiency. A general framework for energy-efficient NFV in 5G networks was proposed in Chapter 4. In this framework, an IP over WDM network was proposed as the backbone network whilst PON was proposed as an energy-efficient wired access network that carries the burden of backhaul/fronthaul traffic. The mobile core functions were virtualised and provided as virtualised network function (VM). In addition, the baseband function was abstracted from its underlying hardware and provided as a VM after being separated from RRH node. A MILP optimisation model was developed with the objective of

minimising the total power consumption. To capture a range of processing scenarios and account for factors such as variation in the BBU processor type or number of cycles per instruction, five scenarios were examined with different ranges of normalised workload (NWL) of BBUVMs and CNVMs. In addition, two main approaches were investigated: virtualisation in IP over WDM network only and virtualisation in both IP over WDM and PON networks. Virtualisation in the IP over WDM and PON networks approach showed an average saving in power consumption of 22% compared to the approach where virtualisation was restricted in the IP over WDM network.

To investigate the impact of the traffic between CNVMs and the traffic reduction / expansion caused by BBUVM, an MILP model was developed to extend the previous model by considering these factors with a wide range for VM workloads. The MILP model results show that scenario of virtualization in both IP over WDM and PON has less power consumption than the virtualisation in IP over WDM network only. It general, the virtualisation in both IP over WDM and PON shows a maximum power saving of 7% compared to the virtualisation in IP over WDM only recorded when high traffic (low reduction factor) flows in the network from CNVMs toward BBUVMs.

In Chapter 5, an optical-based architecture for energy-efficient network function virtualisation in 5G networks is proposed as an extension to the work done in Chapter 4 by including the wireless resource blocks and the impact of the total number of active users in the network on the total power consumption and VM distribution. A MILP optimisation model was developed with the objective of minimising the total power consumption. In the developed MILP model, the impact

of baseband processing and the ratio of backhaul to the fronthaul traffic were investigated under different amount CNVMs inter-traffic and different total number of active users during different times of the day. The MILP model results showed that the deployment of energy optimised NFV results in a total power saving of 16% (average 8%) during peak hours of the day whilst it is better to run the network without virtualisation in case of low number of active users (around 13% of the full network capacity). In addition, the results revealed that the BBUVM distribution is mainly affected by the total number of active users in the network whilst the distribution of CNVMs is mainly affected by the inter-traffic between them.

For real-time implementation and result validation, two heuristics were developed in this chapter. These are the Energy Efficient NFV without CNVMs inter-traffic (EENFVnoITr) and the Energy Efficient NFV with CNVMs inter-traffic (EENFwithITr) models. EENFVnoITr and EENFVwihtITr were developed to validate the MILP without and with the impact of CNVMs inter-traffic respectively. The results of the heuristics showed that the total power consumption of the EENFVnoITr heuristic is higher than the MILP by a maximum of 9% (average 5%) whilst it is higher than the MILP by a maximum of 9.5% (average 5%) in case of EENFVwithITr heuristic.

Energy efficient caching of contents for VoD services in 5G networks was investigated in Chapter 6 with fixed and variable cache sizes. A MILP model was developed to investigate the impact of fixed-size cache on the total power consumption. Five scenarios were applied with different fixed-size caches for each scenario. The developed MILP model results revealed that the highest fixed-size cache scenario achieved a maximum power saving of 12% compared to the case

where no cache is deployed whilst the variable cache approach achieved further power saving of 9% compared with the fixed size cache approach in a fully loaded network.

in addition, the investigation showed that OLTs are the optimum location to cache the content when the cache size is small and the total number of active users is high, whilst the IP over WDM nodes are the optimum place to cache the content when the number of users is low. As the cache size increases, more IP over WDM nodes are utilised to cache the content resulting in less participant OLTs. The MILP model results were successfully validated by two heuristics: Energy Efficient Fixed Cache Size (EEFCZ) heuristic and Energy Efficient Variable Cache Size (EEVarCZ) heuristic. The heuristics results were compared with their counterpart MILP models and they are close.

In Chapter 7, an energy efficient integrated architecture that includes both caching the content and virtualisation in 5G networks was introduced. A MILP optimisation model was developed with the objective of minimising the total power consumption. Three main approaches were investigated, virtualisation only, caching only, and integrated approach. The MILP model results showed that virtualisation-only approach is better than caching-only approach for VoD services where the virtualisation-only approach compared to caching-only approach achieves maximum power saving of 7% (average 5%) when no CNVMs inter-traffic is considered and 6% (average 4%) with CNVMs inter-traffic that represents 10% of the total backhaul traffic. On other side, the integrated approach has a maximum total power saving of 15% (average 9%) with and without CNVMs inter-traffic compared to virtualisation-only approach, and it achieves a maximum total power saving of 21%

(average 13%) without CNVMs inter-traffic and 20% (average 12%) when CNVMs inter-traffic is considered compared with caching-only approach. The MILP model results were successfully verified by an Energy Efficient Virtualisation and Caching heuristic which was developed for real-time implementation of the MILP model.

8.2 Future research directions

Many research challenges could be brought under the roof of energy efficient NFV in 5G networks including delay, latency, optimum control and management. The following are some future directions for energy efficient NFV in 5G networks.

8.2.1 Impact of latency on VM placement in energy efficient NFV in 5G

One of the key requirements in 5G networks (for certain applications) is the 1ms end to end latency. Putting this requirement in consideration adds extra constrains and challenges to the work presented in this thesis. By exploiting NFV capabilities, VMs could be placed, migrated, and replicated to meet the 5G latency requirements. This adds another degree of improvement as the latency is reduced alongside reduction in the energy consumption.

8.2.2 Backhaul traffic offloading in 5G HetNet using energy efficient NFV

In HetNet networks, Pico and femto-cells could be deployed within the coverage of macro-cells. This architecture represent a raw material for NFV to work on since functionalities of mobile core network and baseband processing could be provided as a VM placed away from the mobile office centre. In such scenario the users of pico and femto-cells communicate with each other using their cell coverage whilst they use the macro-cell as backhaul to communicate with other users in different

cells. Consequently, the traffic will be offloaded from the network main backhaul resulting in low energy consumption.

8.2.3 Renewable energy in NFV 5G networks

The renewable energy availability represented by wind farms, solar energy, and hydropower can be investigated and its impact on NFV and VMs placement can be studied. With the renewable energy deployment, another challenge is added to the current work, where the placement and distribution of VMs over the network with the availability of renewable energy can be optimised.

8.2.4 Markov models for resource provisioning in energy efficient NFV in 5G networks

Stochastic models such as Markov models are promising tools to model randomly changing queueing systems. By using Markov chain model, the trade-off between delay, VM placement and energy efficiency can be analysed and studied. This is another dimension that can extend the current work.

8.2.5 Energy efficient NFV for IoT 5G networks

5G era is the era of IoT devices where everything is connected to everything anywhere and anytime. Therefore, by connecting IoT devices to 5G networks, many challenges are added to the current work. For instance, the optimum placement of VMs to process the IoT data and reduce energy consumption. In this scenario, different levels of processing priorities would be considered such as life-dependent data which are sourced by medical IoT devices.

Appendix A
Real Time Heuristic Algorithms

INPUTS: number of RRH nodes (cells) RRH = 1 … NoRRH number of ONU nodes ONU = 1 … NoONU number of OLT nodes OLT=1 … NoOLT number of IP over WDM nodes N=1 … N number of users per cell usr(RRH) network topology G IP over WDM physical network topology Gp	
OUTPUTS: VM Server utilisation at each node Total power consumption TPC	
1:	BEGIN:
2:	get network topology (G)
3:	get the number of users(USR(r)) for each RRH node (r ∈ RRH)
4:	get the maximum workload capacity (NodeMXwl(h)) at each hosting node (h ∈ OLT)
5:	according to (USR(r)) in each cell (r ∈ RRH), calculated the download traffic (LR(r))
6:	initialise the hosting nodes (h ∈ OLT) workload NodeWL(h)=0
7:	For each RRH node (r ∈ RRH) Do
8:	calculate the BBUVM workload bbuWL(r) needed for the traffic LR(r)
9:	For each BBUVM candidate hosting node (h ∈ OLT) Do
10:	get the shortest path (dis(h,r))

11:	if dis(h,r)=2 AND bbuWL(r)+NodeWL(h)<NodeMXwl(h) Do
12:	set the traffic from node h to r; LB(h,r) = LR(r)
13:	update the hosting node h workload NodeWL(h)= NodeWL(h)+ bbuWL(r)
14:	end if
15:	End For
16:	if node (r ∈ RRH) has not been served by any BBUVM at any node (h ∈ OLT) Do
17:	For each IP over WDM node (h ∈ N) Do
18:	Get the shortest path (dis(h,r))
19:	End For
20:	sort the IP over WDM node (h ∈ N) ascendingly according to the shortest path (dis(h,r))
21:	For each IP over WDM node (h) in the sorted list (NS) Do
22:	if bbuWL(r)+ NodeWL(h)< NodeMXwl(h) Do
23:	set the traffic from node h to r; LB(h,r) = LR(r)
24:	update the hosting node h workload NodeWL(h)= NodeWL(h)+ bbuWL(r)
25:	End if
26:	End For
27:	End if
28:	End For
29:	get CNVM workload (singleCNVMwl)
30:	calculate the total traffic at each BBUVM (LB(h,r)); LBvm(h) = $\sum_{r \in RRH} LB(h,r)$
31:	For each destination BBUVM in node (d ∈ OLT ∪ N) Do
32:	if BBUVM requests traffic LBvm(d)>0 Do
33:	set an initial long distance LongDist
34:	set an initial location VMloc for CNVM

35:	set an initial total number of IPWDM candidate locations totalCNVM=0
36:	For each candidate IPWDM location (vm ∈ OLT ∪ N) Do
37:	if (singleCNVMwl +NodeWL(vm)<NodeMXwl(vm)) Do
38:	Find the shortest path dis(vm,d)
39:	if dim(vm,d)< LongDist Do
40:	LongDist = dim(vm,d)
41:	VMloc = vm
42:	totalCNVM = totalCNVM +1
43:	End if
44:	End if
45:	End For
46:	update the hosting node workload NodeWL(VMloc)= NodeWL(VMloc)+ bbuWL(VMloc)
47:	calculate the traffic from CNVM at VMloc to BBUVM at d; LC(VMloc,d)
48:	End if
49:	End For
50:	initialise the traffic (trf) in each network segment
51:	For each network node (s ∈ ONU ∪ OLT ∪ N) Do
52:	For each network node (d ∈ RRH ∪ ONU ∪ OLT ∪ N) Do
53:	if (d ∈ RRH) AND (LB(h,r)>0) Do
54:	Find the nodes (pNodes)in path between (s,d)
55:	For each two adjacent nodes (i,i+1) in the path (s,d) Do
56:	trf(i,i+1)=trf(i,i+1)+LB(s,d)
57:	else if s has CNVM and d has BBUVM AND (s ≠ d) Do
58:	Find the nodes (pNodes)in path between (s,d)

59:	For each two adjacent nodes (i,i+1) in the path (s,d) Do
60:	trf(i,i+1)=trf(i,i+1)+LC(s,d)
61:	End For
62:	End For
63:	End if
64:	End For
65:	End For
66:	get the physical topology Gph of IPWDM network
67:	Calculate number of wavelength W(i,j)in each physical link (i,j, ∈ N)in IPWDM network
68:	calculate number of aggregation ports AGR(i) for each IPWDM router (i ∈ N)
69	Calculate number of fibers f(i,j)in each physical link (i,j, ∈ N)in IPWDM network
70:	Calculate number of EDFA edfa(i,j)in each physical link (i,j, ∈ N)in IPWDM network
71:	calculate RRH nodes power consumption rrhPC
72:	calculates ONU nodes power consumption onuPC
73:	calculate OLT nodes power consumption oltPC
74:	calculate IPWDM network power consumption ipwdmPC
75:	calculate VM servers power consumption VMsrvPC
76:	calculate the total power consumption TPC
77:	END BEGIN

Algorithm A.1 Pseudocode of the Energy Efficient NFV without CNVMs inter-traffic (EENFVnoITr) heuristic model

INPUTS: number of RRH nodes (cells) RRH = 1 …NoRRH number of ONU nodes ONU = 1…NoONU number of OLT nodes OLT=1…NoOLT number of IP over WDM nodes N=1…N number of users per cell usr(RRH)	
OUTPUTS: Server utilisation at each node Total power consumption TPC	
BEGIN:	
1:	create network topology (G)
2:	get the number of users(USR(r)) for each RRH node (r ∈ RRH)
3:	get the maximum workload capacity (NodeMXwl(h)) at each hosting node (h ∈ OLT)
4:	according to (USR(r)) in each cell (r ∈ RRH), calculated the download traffic (LR(r))
5:	initialise the hosting nodes (h ∈ OLT) workload NodeWL(h)=0
6:	For each RRH node (r ∈ RRH) Do
7:	calculate the BBUVM workload bbuWL(r) needed for the traffic LR(r)
8:	For each BBUVM candidate hosting node (h ∈ OLT) Do
9:	get the shortest path (dis(h,r))
10:	if dis(h,r)=2 AND bbuWL(r)+ NodeWL(h)< NodeMXwl(h) Do
11:	set the traffic from node h to r; LB(h,r) = LR(r)
12:	update the hosting node h workload NodeWL(h)= NodeWL(h)+ bbuWL(r)
13:	end if
14:	End For
15:	if node (r ∈ RRH) has not been served by any BBUVM at any node (h ∈ OLT) Do

16:	For each IP over WDM node (h ∈ N) Do
17:	Get the shortest path (dis(h,r))
18:	End For
19:	sort the IP over WDM node (h ∈ N) ascendingly according to the shortest path (dis(h,r))
20:	For each IP over WDM node (h) in the sorted list (NS) Do
21:	if bbuWL(r)+ NodeWL(h)< NodeMXwl(h) Do
22:	set the traffic from node h to r; LB(h,r) = LR(r)
23:	update the hosting node h workload NodeWL(h)= NodeWL(h)+ bbuWL(r)
24:	End if
25:	End For
26:	End if
27:	End For
28:	set a range of maximum number of IPWDM candidate location to host CNVM (MaxCNVMno to 1)
29:	For each total number of candidate locations (CanVMLoc ∈ MaxCNVMno to 1) Do
30:	get CNVM workload (singleCNVMwl)
31:	calculate the total traffic at each BBUVM (LB(h,r)); LBvm(h) = $\sum_{r \in RRH} LB(h,r)$
32:	For each IP over WDM node (s ∈ N) Do
33:	For each candidate hosting node (d ∈ OLT ∪ N) Do
34:	find the shortest path dist(s,d)
35:	if dist(s,d)<=1 Do
36:	TRatIP(s)=TRatIP(s)+LBvm(d)
37:	End if
38:	End For

39:	End For
40:	sort the IP over WDM nodes discerningly according to the traffic (NS)
41:	For each destination BBUVM in node (d ∈ OLT ∪ N) Do
42:	if BBUVM requests traffic LBvm(d)>0 Do
43:	set an initial long distance LongDist
44:	set an initial location VMloc for CNVM
45:	set an initial total number of IPWDM candidate locations totalCNVM=0
46:	For each IPWDM candidate location (vm ∈ NS) Do
47:	if the (totalCNVM <= CanVMLoc) AND (singleCNVMwl+ NodeWL(vm)< NodeMXwl(vm)) Do
48:	Find the shortest path dis(vm,d)
49:	if dim(vm,d)< LongDist Do
50:	LongDist = dim(vm,d)
51:	VMloc = vm
52:	totalCNVM = totalCNVM +1
53:	End if
54:	End if
55:	End For
56:	update the hosting node workload NodeWL(VMloc)= NodeWL(VMloc)+ bbuWL(VMloc)
57:	calculate the traffic from CNVM at VMloc to BBUVM at d; LC(VMloc,d)
58:	End if
59:	End For
60:	initialise the inter-traffic for each two CNVM

	at nodes (s,d ∈ N); intCNVM(s,d)=0
61:	For each CNVM in IPWDM node (s ∈ N) Do
62:	if IPWDM node (s ∈ N) has a CNVM Do
63:	For each CNVM in IPWDM node (d ∈ N)Do
64:	If (s ≠ d) Do
65:	calculate the traffic intCNVM(s,d)
66:	End if
67:	End For
68:	End if
69:	End For
70:	initialise the traffic (trf) in each network segment
71:	For each network node (s ∈ ONU ∪ OLT ∪ N) Do
72:	For each network node (d ∈ RRH ∪ ONU ∪ OLT ∪ N) Do
73:	if (d ∈ RRH) AND (LB(h,r)>0) Do
74:	Find the nodes (pNodes)in path between (s,d)
75:	For each two adjacent nodes (i,i+1) in the path (s,d)Do
76:	trf(i,i+1) = trf(i,i+1)+ LB(s,d)
77:	End For
78:	else if s has CNVM and d has BBUVM AND (s ≠ d) Do
79:	Find the nodes (pNodes)in path between (s,d)
80:	For each two adjacent nodes (i,i+1) in the path (s,d) Do
81:	trf(i,i+1) = trf(i,i+1)+LC(s,d)
82:	End For

83:	else if both s and d have CNVM AND (s ≠ d) Do
84:	Find the nodes (pNodes)in path between (s,d)
85:	For each two adjacent nodes (i,i+1) in the path (s,d) Do
86:	trf(i,i+1) = trf(i,i+1)+intCNVM(s,d)
87:	End For
88:	End if
89:	End For
90:	End For
91:	get the physical topology Gph of IPWDM network
92:	Calculate number of wavelength W(i,j)in each physical link (i,j, ∈ N)in IPWDM network
93:	calculate number of aggregation ports AGR(i) for each IPWDM router (i ∈ N)
94:	Calculate number of fibers f(i,j)in each physical link (i,j, ∈ N)in IPWDM network
95:	Calculate number of EDFA edfa(i,j)in each physical link (i,j, ∈ N)in IPWDM network
96:	calculate RRH nodes power consumption rrhPC
97:	calculates ONU nodes power consumption onuPC
98:	calculate OLT nodes power consumption oltPC
99:	calculate IPWDM network power consumption ipwdmPC
100:	calculate VM servers power consumption VMsrvPC
101:	Calculate the total power consumption TPC(CanVMLoc) for this total number of CNVM candidate location (CanVMLoc)
102:	End For
103:	set initial total power consumption OpTPC
104:	set initial total CNVM location OpNoCNVM
105:	For each total number of candidate locations (CanVMLoc ∈ MaxCNVMno to 1) Do

106:	if TPC() < OpTPC Do
107:	OpTPC = TPC(CanVMLoc)
108:	OpNoCNVM = CanVMLoc
109:	End if
110:	End For
111:	get network configuration and VM server utilisation for OpNoCNVM
112:	minimum power consumption = OpTPC
END BEGIN	

Algorithm A.2 Pseudocode of the Energy Efficient NFV with CNVMs inter-traffic (EENFVnoITr) heuristic model

INPUTS: number of eNB nodes (cells) eNB = {1…NoeNB} number of ONU nodes ONU = {1…NoONU} number of OLT nodes OLT={1…NoOLT} number of IP over WDM nodes N={1…N} number of users per cell usr(eNB)	
OUTPUTS: Cache nodes distribution over the network Total power consumption TPC	
BEGIN:	
1:	get network topology (G)
2:	get the number of users(USR(r)) for each eNB node (r ∈ eNB)
3:	according to (USR(r)) in each cell (r ∈ eNB), calculated the download traffic (LR(r))
4:	calculate the regular traffic (LG(r)) and video streaming traffic (LV(r)) for each eNB node in the network (r ∈ eNB)
5:	initialise IP over WDM nodes load (nL(n)) where (n ∈ N)

6:	For each IP over WDM node (n ∈ N) Do
7:	For each eNB node (r ∈ eNB) Do
8:	get the shortest path (dis(n,r))
9:	if n is closest node to r DO
10:	get (USR(r))
11:	update the load of node n (nL(n))
12:	end if
13:	End For
14:	End For
15:	sort IP over WDM node in descend order according to their loads(nL(n)) and put them in a list (sortN)
16:	get (fIPnode) where (fIPnode = sortN(1))
17:	accommodate the mobile core node (ASR5000) at (fIPnode)
18:	initialise the mobile traffic (LAsr(fIPnode,r)), (r ∈ eNB)
19:	For each eNB node (r ∈ eNB) Do
20:	LAsr(fIPnode,r)=LGr
21:	End For
22:	get the cache node hit ratio (Ht)
23:	initialise the cache traffic (LCAur(u,r)), (u ∈ OLT ,r ∈ eNB)
24:	For each eNB node r Do
25:	if (r) has not been served Do
26:	For each OLT node (u) Do
27:	get the shortest path (dis(u,r))
28:	if u is closest node to r DO
29:	Calculate the cache traffic (LCAur(u,r)=Ht*LVr(r))
30:	End if
31:	End for
32:	End if
33:	End For

34:	initialise the Video server traffic (LSsr(s,r)), (s ∈ N ,r ∈ eNB)
35:	For each eNB node r Do
36:	calculate the traffic from the video server
37:	End For
38:	get the backhaul to fronthaul traffic ratio (alpha)
39:	initialise the traffic (trf) in each network segment
40:	For each network node (s ∈ OLT ∪ N) Do
41:	For each network node (d ∈ eNB) Do
42:	if LAsr(s,d)>0 Do
43:	Find the nodes (pNodes)in path between (s,d)
44:	For each two adjacent nodes (i,i+1) in the path (s,d) Do
45:	trf(i,i+1) = trf(i,i+1)+ alpha*LAsr(s,d)
46:	End For
47:	End if
48:	if LCAur(s,d)>0 Do
49:	Find the nodes (pNodes)in path between (s,d)
50:	For each two adjacent nodes (i,i+1) in the path (s,d) Do
51:	trf(i,i+1) = trf(i,i+1) + alpha*LCAur(s,d)
52:	End For
53:	End if
54:	if LSsr(s,d)>0 Do
55:	Find the nodes (pNodes)in path between (s,d)
56:	For each two adjacent nodes (i,i+1) in the path (s,d) Do
57:	trf(i,i+1) = trf(i,i+1) + alpha* LSsr(s,d)

58:	End For
59:	End if
60:	End For
61:	End For
62:	get the physical topology Gph of IPWDM network
63:	Calculate number of wavelength W(i,j)in each physical link(i,j, ∈ N)in IPWDM network
64:	calculate number of aggregation ports AGR(i) for each IPWDM router (i ∈ N)
65:	Calculate number of fibers f(i,j)in each physical link(i,j, ∈ N)in IPWDM network
66:	Calculate number of EDFA edfa(i,j)in each physical link(i,j, ∈ N)in IPWDM network
67:	calculate eNB nodes power consumption rrhPC
68:	calculates ONU nodes power consumption onuPC
69:	calculate OLT nodes power consumption oltPC
70:	calculate IPWDM network power consumption ipwdmPC
71:	calculates Cache nodes power consumption CachePC
72:	calculate BBU nodes power consumption bbuPC
73:	calculate mobile core node power consumption asrPC
74:	calculate video server power consumption VsrvPC
75:	minimum power consumption TPC= rrhPC+ onuPC+ oltPC+ ipwdmPC+ CachePC+ bbuPC+ asrPC+ VsrvPC
END BEGIN	

Algorithm A.3 Pseudocode of the EEFCZ heuristic model

INPUTS: number of eNB nodes (cells) eNB = 1…NoeNB number of ONU nodes ONU = 1…NoONU number of OLT nodes OLT=1…NoOLT number of IP over WDM nodes N=1…N

	number of users per cell usr(eNB)
	OUTPUTS: Cache nodes distribution over the network Total power consumption TPC
	BEGIN:
1:	create network topology (G)
2:	get the number of users(USR(r)) for each eNB node (r ∈ eNB)
3:	according to (USR(r)) in each cell (r ∈ eNB), calculated the download traffic (LR(r))
4:	calculate the regular traffic (LG(r)) and video streaming traffic (LV(r)) for each eNB node in the network (r ∈ eNB)
5:	initialise IP over WDM nodes load (nL(n)) where(n ∈N)
6:	For each IP over WDM node (n ∈ N) Do
7:	For each eNB node (r ∈ eNB) Do
8:	get the shortest path (dis(n,r))
9:	if n is closest node to r DO
10:	get (USR(r))
11:	update the load of node n (nL(n))
12:	end if
13:	End For
14:	End For
15:	sort IP over WDM node in descend order according to their loads(nL(n)) and put them in a list (sortN)
16:	get (fIPnode) where (fIPnode = sortN(1))
17:	accommodate the mobile core node (ASR5000) at (fIPnode)
18:	initialise the mobile traffic (LAsr(fIPnode,r)), (r ∈ eNB)
19:	For each eNB node (r ∈ eNB) Do
20:	LAsr(fIPnode,r)=LGr

21:	End For
22:	get the cache node hit ratio (Ht)
23:	get the maximum allowed cache nodes in the IP over WDM (MCacheNd)
24:	For each (nCacheNd) in (1 to MCacheNd) Do
25:	initialise the cache traffic(LCAur(u,r)), (u ∈OLT,r ∈eNB)
26:	get the sorted list (sortN)
27:	get the total network users (TUSR =$\sum_r$ USR(r)), (r ∈eNB)
28:	For each eNB node d Do
29:	if TUSR < 50% of the Max number of network users AND node d has not been severed Do
30:	For idx = 1 to (nCacheNd) Do
31:	get the IP over WDM node (sortN(idx))
32:	get the shortest path distance(idx) =(dis(sortN(idx),d))
33:	End For
34:	get IP over WDM node (sn ∈ sortN) that has min(distance)
35:	Calculate the cache traffic (LCAur(sn,d)=Ht*LVr(d))
36:	else if node d has not been severed Do
37:	For each OLT node s Do
38:	get the shortest path
39:	if node s is the closest node to d Do
40:	Calculate the cache traffic (LCAur(s,d)=Ht*LVr(d))
41:	End if
42:	End For

```
43:             end if
44:         End For
45:         initialise the Video server traffic (LSsr(s,r)),
            (s ∈ N ,r ∈ eNB)
46:         For each eNB node r Do
47:             calculate the traffic from the video server
48:         End For
49:         get the backhaul to fronthaul traffic ratio(alpha)
50:         initialise the traffic (trf) in each network
            segment
51:         For each network node (s ∈ OLT ∪ N) Do
52:             For each network node (d ∈ eNB) Do
53:                 if LAsr(s,d)>0 Do
54:                     Find the nodes (pNodes)in path
                        between (s,d)
55:                     For each two adjacent nodes
                        (i,i+1) in the path (s,d) Do
56:                         trf(i,i+1) = trf(i,i+1)+
                            alpha*LAsr(s,d)
57:                     End For
58:                 End if
59:                 if LCAur(s,d)>0 Do
60:                     Find the nodes (pNodes)in path
                        between (s,d)
61:                     For each two adjacent nodes
                        (i,i+1) in the path (s,d) Do
62:                         trf(i,i+1) = trf(i,i+1)
                                + alpha*LCAur(s,d)
63:                     End For
64:                 End if
65:                 if LSsr(s,d)>0 Do
66:                     Find the nodes (pNodes)in path
                        between (s,d)
```

67:	For each two adjacent nodes (i,i+1) in the path (s,d) Do
68:	trf(i,i+1) = trf(i,i+1) + alpha* LSsr(s,d)
69:	End For
70:	End if
71:	End For
72:	End For
73:	get the physical topology Gph of IPWDM network
74:	Calculate number of wavelength W(i,j)in each physical link(i,j, ∈ N)in IPWDM network
75:	calculate number of aggregation ports AGR(i) for each IPWDM router (i ∈ N)
76:	Calculate number of fibers f(i,j)in each physical link(i,j, ∈ N)in IPWDM network
77:	Calculate number of EDFA edfa(i,j)in each physical link(i,j, ∈ N)in IPWDM network
78:	calculate eNB nodes power consumption rrhPC
79:	calculates ONU nodes power consumption onuPC
80:	calculate OLT nodes power consumption oltPC
81:	calculate IPWDM network power consumption ipwdmPC
82:	calculates Cache nodes power consumption CachePC
83:	calculate BBU nodes power consumption bbuPC
84:	calculate mobile core node power consumption asrPC
85:	calculate video server power consumption VsrvPC
86:	calculate power consumption TPC(nCacheNd)
87:	End For
88:	Get min(TPC)
END BEGIN	

Algorithm A.4 The pseudocode of EEVarCZ heuristic model

INPUTS: number of RRH nodes (cells) RRH = 1…NoRRH

	number of ONU nodes ONU = 1…NoONU number of OLT nodes OLT=1…NoOLT number of IP over WDM nodes N=1…N number of users per cell usr(RRH)
OUTPUTS: Total power consumption TPC	
BEGIN:	
1:	create network topology (G)
2:	get the number of users(USR(r)) for each RRH node (r ∈ RRH)
3:	get the maximum workload capacity (NodeMXwl(h)) at each hosting node (h ∈ OLT)
4:	according to (USR(r)) in each cell (r ∈ RRH), calculated the download traffic (LR(r))
5:	calculate the regular traffic (LG(r)) and video streaming traffic (LV(r)) for each RRH node (r ∈ RRH)
6:	initialise the hosting nodes (h ∈ OLT) workload NodeWL(h)=0
7:	For each RRH node (r ∈ RRH) Do
8:	calculate the BBUVM workload bbuWL(r) needed for the traffic LG(r)
9:	For each BBUVM candidate hosting node (h ∈ OLT) Do
10:	get the shortest path (dis(h,r))
11:	if dis(h,r)=2 AND bbuWL(r)+ NodeWL(h)< NodeMXwl(h) Do
12:	set the traffic from node h to r; LB(h,r) = LG(r)
13:	update the hosting node h workload NodeWL(h)= NodeWL(h)+ bbuWL(r)
14:	end if
15:	End For
16:	if node (r ∈ RRH) has not been served by any BBUVM at any node (h ∈ OLT) Do
17:	For each IP over WDM node (h ∈ N) Do

18:	Get the shortest path (dis(h,r))
19:	End For
20:	sort the IP over WDM node (h ∈ N) ascendingly according to the shortest path (dis(h,r))
21:	For each IP over WDM node (h) in the sorted list(NS) Do
22:	if bbuWL(r)+ NodeWL(h)< NodeMXwl(h) Do
23:	set the traffic from node h to r; LB(h,r) = LG(r)
24:	update the hosting node h workload NodeWL(h)= NodeWL(h)+ bbuWL(r)
25:	End if
26:	End For
27:	End if
28:	End For
29:	set a range of maximum number of IPWDM candidate location to host CNVM (MaxCNVMno to 1)
30:	For each total number of candidate locations (CanVMLoc ∈ MaxCNVMno to 1) Do
31:	get CNVM workload (singleCNVMwl)
32:	For each BBUVM at node h (h ∈OLT ∪ N) Do
33:	get the traffic from CNVMs to BBUVMs Lph(CanVMLoc,h)
34:	End For
35:	get CNVMs inter-traffic (CNVMint)
36:	evaluate the IP over WDM network power consumption IPWDMpc(CanVMLoc)
37:	End For
38:	get min(IPWDMpc(CanVMLoc))
39:	update nodes workload NodeWL(h) node h (h ∈N)
40:	get number of CNVMs (CanVMLoc) at min(IPWDMpc(CanVMLoc))
41:	get the location of CNVMs and network routing

```
42:   accommodate the VoD server at the location of first CNVM (VoDLoc)
43:   get the total users in the network (netUsers)
44:   get the Maximum users in the network (MAXnetUsers)
45:   get the cache node hit ratio (Ht)
46:   initialise the cache traffic (LCAur(u,r)), (u ∈ OLTU N ,r ∈ RRH)
47:   set a range of maximum cache nodes at IPWDM network (MaxCacheNodes to 1)
48:   sort IPWDM nodes according to no users per node
      sIPWDM = sort(IPWDM, descend)
49:   For each cNode in the range (MaxCacheNodes to 1) Do
50:       For each RRH node r (r ∈ RRH) Do
51:           if netUsers < 0.5*MAXnetUsers AND
                  ∑u LCAur (u, r) = 0  Do
52:               For inx = 1 to MaxCacheNodes Do
53:                   get shortest path
                      dis(sIPWDM(inx), r)
54:                   if node sIPWDM(inx) is the
                          closest to r Do
55:                       Calculate the cache traffic
                      (LCAur(sIPWDM(inx),r)=Ht*LVr(r))
56:                   end if
57:               End For
58:           else if ∑u LCAur (u, r) = 0 Do
59:               For each node u (u ∈ OLT) Do
60:                   get shortest path dis(u, r)
61:                   if node u is the closest to r Do
62:                       Calculate the cache traffic
                          (LCAur(u,r)=Ht*LVr(r))
63:                   end if
64:               End For
65:           end if
66:           get traffic from VoD server LS(VoDLoc,r)
```

67:	End For
68:	evaluate the IP over WDM network power consumption IPWDMpc2(cNode)
69:	End For
70:	get min(IPWDMpc2(cNode))
71:	get number of cache nodes (cNode) at min(IPWDMpc2(cNode))
72:	get the location of cache nodes and network routing
73:	Initialise a traffic between BBUVM and RRH nodes LVur(u,r) (r ∈ RRH, h ∈ OLTU N)
74:	For each RRH node r (r ∈ RRH) Do
75:	For each node h (h ∈ OLTU N) Do
76:	if LBhr(h,r)>0 Do
78:	For each node u (h ∈ OLTU N) Do
79:	if LCAur(u,r)>0 Do
80:	route the traffic through BBUVM update LVur(u,r)
81:	if h≠u Do
82:	update backhaul traffic LVhu(h,u)
83:	end if
84:	end if
85:	if LSsr(u,r)>0 Do
86:	route the traffic through BBUVM update LVur(u,r)
87:	if h≠u Do
89:	update backhaul traffic LVhu(h,u)
90:	end if
91:	end if
92:	End For

```
93:             end if
94:         End For
95:     End For
96:     initialise the traffic (trf) in each network segment
97:     For each network node (r ∈ RRH) Do
98:         For each network node (u ∈ OLTU N) Do
99:             if LVur(u,r)>0 Do
100:                Find the nodes (pNodes)in path between
                       (s,d)
101:                For each two adjacent nodes (i,i+1) in
                           the path (s,d) Do
102:                    trf(i,i+1) = trf(i,i+1)+
                        LVur(u,r)
103:                End For
104:            End if
105:            if LBh(u,r)>0 Do
106:                Find the nodes (pNodes)in path between
                       (s,d)
107:                For each two adjacent nodes (i,i+1) in
                           the path (s,d) Do
108:                    trf(i,i+1) = trf(i,i+1)+LBh(u,r)
109:                End For
110:            End if
111:        End For
112:    End For
113:    For each network node (d ∈ OLTU N) Do
114:        For each network node (s ∈ OLTU N) Do
115:            if s≠d AND LCph(s,d)>0 Do
116:                Find the nodes (pNodes)in path between
                       (s,d)
117:                For each two adjacent nodes (i,i+1) in
                           the path (s,d) Do
118:                    trf(i,i+1) = trf(i,i+1)+
                                LCph(s,d)
```

```
119:                    End For
120:              End if
121:              if s≠d AND LVhu(s,d)>0 Do
122:                    Find the nodes (pNodes)in path between
                              (s,d)
123:                    For each two adjacent nodes (i,i+1) in
                                  the path (s,d) Do
124:                          trf(i,i+1) = trf(i,i+1)+
                                        LVhu(s,d)
125:                    End For
126:              End if
127:        End For
128:    End For
129:    For each network node (d ∈ N) Do
130:        For each network node (s ∈ N) Do
131:              if s≠d AND CNVMint (s,d)>0 Do
132:                    Find the nodes (pNodes)in path between
                              (s,d)
133:                    For each two adjacent nodes (i,i+1) in
                                  the path (s,d) Do
134:                          trf(i,i+1) = trf(i,i+1)+
                                    CNVMint(s,d)
135:                    End For
136:              End if
137:        End For
138:    End For
139:    get the physical topology Gph of IPWDM network
140:    Calculate number of wavelength W(i,j)in each
physical link(i,j, ∈ N)in IPWDM network
141:    calculate number of aggregation ports AGR(i) for
each IPWDM router (i ∈ N)
142:    Calculate number of fibers f(i,j)in each physical
link(i,j, ∈ N)in IPWDM network
143:    Calculate number of EDFA edfa(i,j)in each physical
link(i,j, ∈ N)in IPWDM network
```

144:	Calculate the total nodes workloads $\sum_{h \in OLT \cup N} NodeWL(h)$ and calculate the VMs server power consumption VMsrvPC
145:	calculate video server power consumption VsrvPC
146:	calculates Cache nodes power consumption CachePC
147:	calculate RRH nodes power consumption rrhPC
148:	calculates ONU nodes power consumption onuPC
149:	calculate OLT nodes power consumption oltPC
150:	calculate IPWDM network power consumption ipwdmPC
151:	minimum power consumption TPC= rrhPC+ onuPC+ oltPC+ ipwdmPC+ CachePC+ VsrvPC+ VMsrvPC
END BEGIN	

Algorithm A.5 EEVIRandCa heuristic model pseudocode

References

[1] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2016–2021," *White Paper*, June 2017 June 2017.

[2] I. Neokosmidis, T. Rokkas, P. Paglierani, C. Meani, K. M. Nasr, K. Moessner, *et al.*, "Techno Economic Assessment of Immersive Video Services in 5G Converged Optical/Wireless Networks," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, 2018, pp. 1-3.

[3] V. G. Nguyen, A. Brunstrom, K. J. Grinnemo, and J. Taheri, "5G Mobile Networks: Requirements, Enabling Technologies, and Research Activities," *A Comprehensive Guide to 5G Security*, pp. 31-57, 2018.

[4] Z. Zhang, Y. Gao, Y. Liu, and Z. Li, "Performance evaluation of shortened transmission time interval in LTE networks," in *Wireless Communications and Networking Conference (WCNC), 2018 IEEE*, 2018, pp. 1-5.

[5] L. Belkhir and A. Elmeligi, "Assessing ICT global emissions footprint: Trends to 2040 & recommendations," *Journal of Cleaner Production*, vol. 177, pp. 448-463, 2018.

[6] A. Z. Aktas, "Could energy hamper future developments in information and communication technologies (ICT) and knowledge engineering?," *Renewable and Sustainable Energy Reviews*, 2017.

[7] M. A. Rosen and S. Koohi-Fayegh, "The prospects for hydrogen as an energy carrier: an overview of hydrogen energy and hydrogen energy systems," *Energy, Ecology and Environment*, vol. 1, pp. 10-29, 2016.

[8] S. Wagh and P. Walke, "REVIEW ON WIND-SOLAR HYBRID POWER SYSTEM," *International Journal of Research In Science & Engineering*, vol. 3, 2017.

[9] M. Gruber, O. Blume, D. Ferling, D. Zeller, M. A. Imran, and E. C. Strinati, "EARTH — Energy Aware Radio and Network Technologies," presented at the IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, 2009.

[10] GreenTouch. [Online]. Available: https://s3-us-west-2.amazonaws.com/belllabs-microsite-greentouch/index.php@page=home.html. [Accessed: 2018]

[11] P. Lähdekorpi, M. Hronec, P. Jolma, and J. Moilanen, "Energy efficiency of 5G mobile networks with base station sleep modes," in *Standards for Communications and Networking (CSCN), 2017 IEEE Conference on*, 2017, pp. 163-168.

[12] X. Ge, J. Yang, H. Gharavi, and Y. Sun, "Energy efficiency challenges of 5G small cell networks," *IEEE Communications Magazine*, vol. 55, pp. 184-191, 2017.

[13] R. Bassoli, M. Di Renzo, and F. Granelli, "Analytical energy-efficient planning of 5G cloud radio access network," in *Communications (ICC), 2017 IEEE International Conference on*, 2017, pp. 1-4.
[14] K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li, X. Peng, *et al.*, "Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks," *IEEE access*, vol. 4, pp. 5896-5907, 2016.
[15] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, *et al.*, "What Will 5G Be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1065-1082, 2014.
[16] T. Choi, T. Kim, W. TaverNier, A. Korvala, and J. Pajunpaa, "Agile management of 5G core network based on SDN/NFV technology," in *Information and Communication Technology Convergence (ICTC), 2017 International Conference on*, 2017, pp. 840-844.
[17] H. Hawilo, L. Liao, A. Shami, and V. C. Leung, "NFV/SDN-based vEPC solution in hybrid clouds," in *Communications Conference (MENACOMM), IEEE Middle East and North Africa*, 2018, pp. 1-6.
[18] S. H. Won, M. Mueck, V. Frascolla, J. Kim, G. Destino, A. Pärssinen, *et al.*, "Development of 5G CHAMPION testbeds for 5G services at the 2018 Winter Olympic Games," in *Signal Processing Advances in Wireless Communications (SPAWC), 2017 IEEE 18th International Workshop on*, 2017, pp. 1-5.
[19] V. Q. Rodriguez and F. Guillemin, "Cloud-RAN modeling based on parallel processing," *IEEE Journal on Selected Areas in Communications*, vol. 36, pp. 457-468, 2018.
[20] G. C. Valastro, D. Panno, and S. Riolo, "A SDN/NFV based C-RAN architecture for 5G Mobile Networks," in *2018 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT)*, 2018, pp. 1-8.
[21] A. Tzanakaki, M. Anastasopoulos, I. Berberana, D. Syrivelis, P. Flegkas, T. Korakis, *et al.*, "Wireless-optical network convergence: enabling the 5G architecture to support operational and end-user services," *IEEE Communications Magazine*, vol. 55, pp. 184-192, 2017.
[22] M. Riva, H. Donâncio, F. R. Almeida, G. B. Figueiredo, R. I. Tinini, R. M. Cesar Jr, *et al.*, "An Elastic Optical Network-based Architecture for the 5G Fronthaul," in *Simpósio Brasileiro de Redes de Computadores (SBRC)*, 2018.
[23] J. Luo, Q. Chen, and L. Tang, "Reducing Power Consumption by Joint Sleeping Strategy and Power Control in Delay-Aware C-RAN," *IEEE Access*, 2018.
[24] N. Panwar, S. Sharma, and A. K. Singh, "A survey on 5G: The next generation of mobile communication," *Physical Communication*, vol. 18, pp. 64-84, 2016.
[25] A. Gupta and R. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies," *IEEE Access*, vol. 3, pp. 1-1, 2015.
[26] I. M. D. D. Neve, "An Overview on fifth generation (5G) mobile wireless Technology," *International Journal of Innovative and Emerging Research in Engineering*, vol. 3, 2016.

[27] S. Panwar, "Review Paper On: A Survey on 5g Technology," *International Journal of Electronics, Electrical and Computational System,* vol. 6, 2017.
[28] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials,* vol. 18, pp. 1617-1655, 2016.
[29] M. H. Alsharif and R. Nordin, "Evolution towards fifth generation (5G) wireless networks: Current trends and challenges in the deployment of millimetre wave, massive MIMO, and small cells," in *Telecommunication Systems:Modelling, Analysis, Design and Management*, ed: Springer US, 2016, pp. 1-21.
[30] R. K. Saha, P. Saengudomlert, and C. Aswakul, "Evolution toward 5G mobile networks-a survey on enabling technologies," *Engineering Journal,* vol. 20, pp. 87-119, 2016.
[31] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, and J. Sachs, "5G wireless access: requirements and realization," *Communications Magazine, IEEE,* vol. 52, pp. 42-47, 2014.
[32] 5GPPP, "5G Vision. The 5G Infrastructure Public Private Partnership: the next generation of communication networks and services," 2015.
[33] D. Fang, Y. Qian, and R. Q. Hu, "Security for 5G Mobile Wireless Networks," *IEEE Access*, vol. 6, pp. 4850-4874, 2018.
[34] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration," *IEEE Communications Magazine,* 2018.
[35] D. Soldani and A. Manzalini, "Horizon 2020 and Beyond: On the 5G Operating System for a True Digital Society," *IEEE Vehicular Technology Magazine,* vol. 10, pp. 32-42, 2015.
[36] S. Singh, N. Saxena, A. Roy, and H. Kim, "A survey on 5G network technologies from social perspective," *IETE Technical Review*, vol. 34, pp. 30-39, 2017.
[37] S. A. A. Shah, E. Ahmed, M. Imran, and S. Zeadally, "5G for Vehicular Communications," *IEEE Communications Magazine,* vol. 56, pp. 111-117, 2018.
[38] D. Evans, "The internet of things: How the next evolution of the internet is changing everything," *CISCO white paper,* vol. 1, pp. 1-11, 2011.
[39] I. Ullah Khan, N. Azim, S. B. Hussain Shah, Y. Fuliang, and M. Sameen, "Interconnected Computer Networks Security and Internet of Things: Wireless Sensor Networks," *IJMCA*, vol. 4, pp. 439-444, 2017.
[40] N.-D. Đào, H. Zhang, X. Li, and P. Leroux, "Radio access network coordination framework toward 5G mobile wireless networks," in *International Conference on Computing, Networking and Communications (ICNC), 2015*, 2015, pp. 1039-1043.
[41] T. Asai, "5G radio access network and its requirements on mobile optical network," in *International Conference on Optical Network Design and Modeling (ONDM), 2015* 2015, pp. 7-11.
[42] D. Aziz, K. Kusume, O. Queseth, H. Tullberg, M. Fallgren, M. Schellmann, *et al.*, "Deliverable D8. 4 METIS final project report," *Tech. Rep.*, 2015.

[43] S. Mattisson, "Overview of 5G requirements and future wireless networks," in *ESSCIRC 2017-43rd IEEE European Solid State Circuits Conference*, 2017, pp. 1-6.
[44] S. Li, L. Da Xu, and S. Zhao, "5G internet of things: A survey," *Journal of Industrial Information Integration,* 2018.
[45] M. Agiwal, N. Saxena, and A. Roy, "Towards Connected Living: 5G Enabled Internet of Things (IoT)," *IETE Technical Review*, pp. 1-13, 2018.
[46] N. Al-Falahy and O. Y. Alani, "Technologies for 5G networks: challenges and opportunities," *IT Professional,* vol. 19, pp. 12-20, 2017.
[47] P. Zhang, J. Lu, Y. Wang, and Q. Wang, "Cooperative localization in 5G networks: A survey," *ICT Express*, vol. 3, pp. 27-32, 2017.
[48] Ericsson, "More Than 50 billion connected devices," *White Paper,* 2011.
[49] Y. Zhu, "Efficient resource allocation for 5G hybrid wireless networks," UCL (University College London), 2017.
[50] E. Vlachos, A. S. Lalos, K. Berberidis, and C. Tselios, "Autonomous driving in 5G: Mitigating interference in OFDM-based vehicular communications," in *Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2017 IEEE 22nd International Workshop on*, 2017, pp. 1-6.
[51] M. Niaz, "5G Mobile Communication: Technology Enablers & Interference Mitigation Methods," 2015.
[52] M. G. Sánchez, M. P. Táboas, and E. L. Cid, "Millimeter wave radio channel characterization for 5G vehicle-to-vehicle communications," *Measurement,* vol. 95, pp. 223-229, 2017.
[53] N. H. Alkhazaali, R. A. Aljiznawi, S. Q. Jabbar, and D. J. Kadhim, "Mobile Communication through 5G Technology (Challenges and Requirements)," *International Journal of Communications, Network and System Sciences,* vol. 10, p. 202, 2017.
[54] A. Khan, Y. Javed, J. Abdullah, J. Nazim, and N. Khan, "Security issues in 5G device to device communication," *IJCSNS,* vol. 17, p. 366, 2017.
[55] X. Artiga, J. Nunez-Martinez, A. Perez-Neira, G. J. L. Vela, J. M. F. Garcia, and G. Ziaragkas, "Terrestrial-satellite integration in dynamic 5G backhaul networks," in *Advanced Satellite Multimedia Systems Conference and the 14th Signal Processing for Space Communications Workshop (ASMS/SPSC), 2016 8th*, 2016, pp. 1-6.
[56] A. Ijaz, L. Zhang, P. Xiao, and R. Tafazolli, "Analysis of Candidate Waveforms for 5G Cellular Systems," in *Towards 5G Wireless Networks-A Physical Layer Perspective*, ed: InTech, 2016.
[57] Y.-H. Lee, A.-S. Wang, Y.-D. Liao, T.-W. Lin, Y.-J. Chi, C.-C. Wong, *et al.*, "Wireless power IoT system using polarization switch antenna as polling protocol for 5G mobile network," in *Wireless Power Transfer Conference (WPTC), 2017 IEEE*, 2017, pp. 1-3.
[58] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The new 5G radio access technology," *IEEE Communications Standards Magazine,* vol. 1, pp. 24-30, 2017.
[59] W. Webb, "Opinion First Person-Comment: Communications-Outdated strategies are the wrong approach to implementing 5G," *Engineering & Technology*, vol. 12, pp. 23-23, 2017.

[60] M. Gidlund, T. Lennvall, and J. Åkerberg, "Will 5G become yet another wireless technology for industrial automation?," in *Industrial Technology (ICIT), 2017 IEEE International Conference on*, 2017, pp. 1319-1324.
[61] A. M. S. Isobe and H. Takahashi, "5G Standardization Trends at 3GPP," 2018.
[62] W. Zhang, Y. Huang, D. He, Y. Zhang, Y. Zhang, R. Liu, *et al.*, "Convergence of a Terrestrial Broadcast Network and a Mobile Broadband Network," *IEEE Communications Magazine*, vol. 56, pp. 74-81, 2018.
[63] C. Christophorou, A. Pitsillides, and I. Akyildiz, "CelEc framework for reconfigurable small cells as part of 5G ultra-dense networks," in *Communications (ICC), 2017 IEEE International Conference on*, 2017, pp. 1-7.
[64] V. G. Nguyen, A. Brunstrom, K. J. Grinnemo, and J. Taheri, "5G Mobile Networks–Requirements, Enabling Technologies, and Research Activities," 2017.
[65] Ł. Apiecionek, "Limiting Energy Consumption by Decreasing Packets Retransmissions in 5G Network," *Mobile Information Systems*, vol. 2017, 2017.
[66] J. S. Marcus and G. Molnar, "Network sharing and 5G in Europe: The potential benefits of using SDN or NFV," 2017.
[67] W. H. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5G wireless networks," *Wireless Communications, IEEE*, vol. 21, pp. 106-112, 2014.
[68] T. L. Marzetta, "Massive MIMO: an introduction," *Bell Labs Technical Journal*, vol. 20, pp. 11-22, 2015.
[69] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *Communications Magazine, IEEE*, vol. 52, pp. 186-195, 2014.
[70] L. Zhao, H. Zhao, K. Zheng, and W. Xiang, *Massive MIMO in 5G Networks: Selected Applications*: Springer, 2017.
[71] H. H. Yang and T. Q. Quek, *Massive MIMO Meets Small Cell: Backhaul and Cooperation*: Springer, 2016.
[72] L. Lu, G. Li, A. Swindlehurst, A. Ashikhmin, and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," 2013.
[73] X. Xu, M. Liu, J. Xiong, and G. Lei, "Key technology and application of millimeter wave communications for 5G: a survey," *Cluster Computing*, pp. 1-13, 2018.
[74] S. Song, K. Chang, C. Yoon, and J. M. Chung, "Special Issue on 5G Communications and Experimental Trials with Heterogeneous and Agile Mobile networks," *ETRI Journal*, vol. 40, pp. 7-9, 2018.
[75] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: opportunities and challenges," *Wireless Networks*, vol. 21, pp. 2657-2676, 2015.
[76] M. Elkashlan, T. Q. Duong, and H.-H. Chen, "Millimeter-wave communications for 5G: fundamentals: Part I [Guest Editorial]," *IEEE Communications Magazine*, vol. 52, pp. 52-54, 2014.
[77] J. Huang, C.-X. Wang, R. Feng, J. Sun, W. Zhang, and Y. Yang, "Multi-frequency mmWave massive MIMO channel measurements and

characterization for 5G wireless communication systems," *IEEE journal on selected areas in communications*, vol. 35, pp. 1591-1605, 2017.

[78] M. M. Abdin, W. Joel, D. Johnson, and T. M. Weller, "A system and technology perspective on future 5G mm-wave communication systems," in *Wireless and Microwave Technology Conference (WAMICON), 2017 IEEE 18th*, 2017, pp. 1-6.

[79] P. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5G network architecture," *Communications Magazine, IEEE*, vol. 52, pp. 65-75, 2014.

[80] S. Buzzi, I. Chih-Lin, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, "A survey of energy-efficient techniques for 5G networks and challenges ahead," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 697-709, 2016.

[81] D. Sabella, A. de Domenico, E. Katranaras, M. A. Imran, M. di Girolamo, U. Salim, *et al.*, "Energy Efficiency Benefits of RAN-as-a-Service Concept for a Cloud-Based 5G Mobile Network Infrastructure," *IEEE Access*, vol. 2, pp. 1586-1597, 2014.

[82] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Foundations and Trends® in Communications and Information Theory*, vol. 11, pp. 185-396, 2015.

[83] S. Hong, J. Brand, J. I. Choi, M. Jain, J. Mehlman, S. Katti, *et al.*, "Applications of self-interference cancellation in 5G and beyond," *IEEE Communications Magazine*, vol. 52, pp. 114-121, 2014.

[84] D. Korpi, L. Anttila, V. Syrjälä, and M. Valkama, "Widely linear digital self-interference cancellation in direct-conversion full-duplex transceiver," *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1674-1687, 2014.

[85] M. Zhou, Y. Liao, and L. Song, "Full-Duplex Wireless Communications for 5G," in *5G Mobile Communications*, ed: Springer, 2017, pp. 299-335.

[86] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, *et al.*, "A survey on 3GPP heterogeneous networks," *IEEE Wireless communications*, vol. 18, 2011.

[87] J. Hoadley and P. Maveddat, "Enabling small cell deployment with HetNet," *IEEE Wireless Communications*, vol. 19, pp. 4-5, 2012.

[88] J. Thompson, X. Ge, H.-C. Wu, R. Irmer, H. Jiang, G. Fettweis, *et al.*, "5G wireless communication systems: prospects and challenges [Guest Editorial]," *IEEE Communications Magazine*, vol. 52, pp. 62-64, 2014.

[89] S. Rajoria, A. Trivedi, and W. W. Godfrey, "A comprehensive survey: Small cell meets massive MIMO," *Physical Communication*, vol. 26, pp. 40-49, 2018.

[90] J. Zhang, "Tutorial on Small Cell/HetNet Deployment," presented at the IEEE Globecom Industry Forum, 2012.

[91] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, "Networks and devices for the 5G era," *Communications Magazine, IEEE*, vol. 52, pp. 90-96, 2014.

[92] Z. Mlika, E. Driouch, and W. Ajib, "Energy-Efficient Base Station Operation and Association in HetNets: Complexity and Algorithms," *IEEE Transactions on Wireless Communications*, 2018.

[93] P. Gandotra, R. K. Jha, and S. Jain, "Green communication in next generation cellular networks: a survey," *IEEE Access*, vol. 5, pp. 11727-11758, 2017.
[94] K. Benzekki, A. El Fergougui, and A. Elbelrhiti Elalaoui, "Software-defined networking (SDN): a survey," *Security and communication networks*, vol. 9, pp. 5803-5833, 2016.
[95] S. Sezer, S. Scott-Hayward, P. K. Chouhan, B. Fraser, D. Lake, J. Finnegan, *et al.*, "Are we ready for SDN? Implementation challenges for software-defined networks," *IEEE Communications Magazine*, vol. 51, pp. 36-43, 2013.
[96] 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, pp. 1617-1634, 2014.
[97] D. Kreutz, F. M. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proceedings of the IEEE*, vol. 103, pp. 14-76, 2015.
[98] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, *et al.*, "OpenFlow: enabling innovation in campus networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, pp. 69-74, 2008.
[99] W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie, "A survey on software-defined networking," *IEEE Communications Surveys & Tutorials*, vol. 17, pp. 27-51, 2015.
[100] R. Chávez-Santiago, M. Szydełko, A. Kliks, F. Foukalas, Y. Haddad, K. E. Nolan, *et al.*, "5G: The convergence of wireless communications," *Wireless Personal Communications*, vol. 83, pp. 1617-1642, 2015.
[101] A. Hakiri and P. Berthou, "Leveraging SDN for the 5G Networks," *Software Defined Mobile Networks (SDMN): Beyond LTE Network Architecture*, pp. 61-80, 2015.
[102] M. Chiosi, D. Clarke, P. Willis, A. Reid, J. Feger, M. Bugenhagen, *et al.*, "Network functions virtualisation: An introduction, benefits, enablers, challenges and call for action," in *SDN and OpenFlow World Congress*, 2012, pp. 22-24.
[103] W. K. Johnston III, "The Birth of Fiberoptics from" Light Guiding"," *Journal of Endourology*, vol. 18, pp. 425-426, 2004.
[104] S. P. Thompson, "The photophone," ed: Nature Publishing Group, 1880.
[105] F. M. MIMS III, "Alexander Graham Bell and the photophone: the centennial of the invention of light-wave communications, 1880–1980," *Optics News*, vol. 6, pp. 8-16, 1980.
[106] M. Bertolotti, *The history of the laser*: CRC press, 2004.
[107] J. Hecht, *The laser guidebook*: Tab Books, 1992.
[108] I. Tomkos, B. Mukherjee, S. K. Korotky, R. Tucker, and L. Lunardi, "The Evolution of Optical Networking [Scanning the Issue]," *Proceedings of the IEEE*, vol. 100, pp. 1017-1022, 2012.
[109] K. J. Lee and T. J. Aprille, "SONET evolution: the challenges ahead," in *Global Telecommunications Conference, 1991. GLOBECOM'91.'Countdown to the New Millennium. Featuring a Mini-Theme on: Personal Communications Services*, 1991, pp. 736-740.

[110] R. Horak, *Telecommunications and data communications handbook*: John Wiley & Sons, 2007.
[111] B. Mukherjee, "WDM optical communication networks: progress and challenges," *Selected Areas in Communications, IEEE Journal on*, vol. 18, pp. 1810-1824, 2000.
[112] A. Marincic and V. Acimovic-Raspopovic, "Evolution of WDM optical networks," in *International Conference onTelecommunications in Modern Satellite, Cable and Broadcasting Service (TELSIKS)*, 2001, pp. 473-480.
[113] Cisco, "Introduction to DWDM Technology," ed: Indianápolis-USA, 2000.
[114] S. V. Kartalopoulos, "Introduction to DWDM Technology: Data in a Rainbow," ed: IEEE Press, New York, 2000.
[115] ITU-T. (2012) "Spectral grid for WDM applications: DWDM frequency grid". *ITU-T Rec G.694.1*. Available: http://www.itu.int/ITU-T/recommendations/rec.aspx?rec=11482
[116] M. Maier, *Optical switching networks* vol. 324: Cambridge University Press Cambridge, 2008.
[117] A. Koçyiğit, D. GÖKIŞIK, and S. Bilgen, "All-optical networking," *Turkish journal of electrical engineering & computer sciences*, vol. 9, pp. 69-122, 2001.
[118] J. Buysse, M. De Leenheer, C. Develder, B. Dhoedt, and P. Demeester, "Cost-effective Burst-Over-Circuit-Switching in a hybrid optical network," in *Networking and Services, 2009. ICNS'09. Fifth International Conference on*, 2009, pp. 499-504.
[119] T. E. El-Gorashi and J. M. Elmirghani, "Optical Storage Area Networks," in *Towards Digital Optical Networks*, ed: Springer, 2009, pp. 285-302.
[120] P. Iovanna, F. Testa, R. Sabella, A. Bianchi, M. Puleri, M. R. Casanova, *et al.*, "Packet-optical integration nodes for next generation transport networks," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 4, pp. 821-835, 2012.
[121] R. S. Tucker, "Scalability and energy consumption of optical and electronic packet switching," *Journal of Lightwave Technology*, vol. 29, pp. 2410-2421, 2011.
[122] L. Xu, H. G. Perros, and G. N. Rouskas, "A simulation study of access protocols for optical burst-switched ring networks," in *International Conference on Research in Networking*, 2002, pp. 863-874.
[123] M. T. Anan, G. M. Chaudhry, and D. Benhaddou, "Architecture and performance of a next-generation optical burst switch (OBS)," in *Broadband Communications, Networks and Systems, 2006. BROADNETS 2006. 3rd International Conference on*, 2006, pp. 1-9.
[124] M. Reza, M. Hossain, and S. P. Majumder, "Evaluation of burst loss rate of an optical burst switching (OBS) network with Wavelength Conversion Capability," *arXiv preprint arXiv:1004.4612*, 2010.
[125] Y. Chen, C. Qiao, and X. Yu, "Optical burst switching (OBS): A new area in optical networking research," *IEEE network*, vol. 18, pp. 16-23, 2004.
[126] Y. Chen, J. S. Turner, and P.-F. Mo, "Optimal burst scheduling in optical burst switched networks," *Journal of Lightwave Technology*, vol. 25, pp. 1883-1894, 2007.

[127] X. Mountrouidou and H. Perros, "On the departure process of burst aggregation algorithms in optical burst switching," *Computer Networks*, vol. 53, pp. 247-264, 2009.
[128] C. TASKIN, "Performance Analysis in IP over WDM Networks," in *Internet Monitoring and Protection, 2007. ICIMP 2007. Second International Conference on*, 2007, pp. 15-15.
[129] G. Shen and R. Tucker, "Energy-minimized design for IP over WDM networks," *Optical Communications and Networking, IEEE/OSA Journal of*, vol. 1, pp. 176-186, 2009.
[130] F. Yuan, X. Niu, X. Li, S. Huang, and W. Gu, "Survivable virtual topology mapping for single-node failure in IP over WDM network," in *Asia Communications and Photonics Conference and Exhibition*, 2011, p. 83101S.
[131] X. Dong, T. El-Gorashi, and J. M. Elmirghani, "Energy-efficient IP over WDM networks with data centres," in *Transparent Optical Networks (ICTON), 2011 13th International Conference on*, 2011, pp. 1-8.
[132] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, "Green IP over WDM networks with data centers," *Journal of Lightwave Technology*, vol. 29, pp. 1861-1880, 2011.
[133] G. Kramer and G. Pesavento, "Ethernet passive optical network (EPON): building a next-generation optical access network," *Communications magazine, IEEE*, vol. 40, pp. 66-73, 2002.
[134] S. B. Weinstein, Y. Luo, and T. Wang, *The ComSoc guide to passive optical networks: Enhancing the last mile access* vol. 1: John Wiley & Sons, 2012.
[135] L. Tawade, S. Mhatli, and R. Attia, "Bidirectional long reach WDM-PON delivering downstream data 20 Gbps and upstream data 10 Gbps using mode locked laser and RSOA," *Optical and Quantum Electronics*, vol. 47, pp. 779-789, 2015.
[136] S. Varghese, "Fabrication and characterization of all-fiber components for optical access networks," 2008.
[137] F. Effenberger, D. Clearly, O. Haran, G. Kramer, R. D. Li, M. Oron, *et al.*, "An introduction to PON technologies [Topics in Optical Communications]," *Communications Magazine, IEEE*, vol. 45, pp. S17-S25 2007.
[138] C.-H. Lee, W. V. Sorin, and B. Y. Kim, "Fiber to the home using a PON infrastructure," *Journal of lightwave technology*, vol. 24, pp. 4568-4583, 2006.
[139] N. Ansari and J. Zhang, *Media access control and resource allocation: For next generation passive optical networks*: Springer Science & Business Media, 2013.
[140] F. Effenberger and T. S. El-Bawab, "Passive optical networks (PONs): past, present, and future," *Optical Switching and Networking*, vol. 6, pp. 143-150, 2009.
[141] I. Cale, A. Salihovic, and M. Ivekovic, "Gigabit passive optical network-GPON," in *Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on*, 2007, pp. 679-684.
[142] D. P. Shea and J. E. Mitchell, "Long-reach optical access technologies," *IEEE Network*, vol. 21, 2007.

[143] Z. Buyankhishig, "Subscriber Reconfigurable TDM PON Architectures," in *Ubi-Media Computing and Workshops (UMEDIA), 2014 7th International Conference on*, 2014, pp. 139-143.
[144] T. Orphanoudakis, H. C. Leligou, E. Kosmatos, and J. D. Angelopoulos, "Performance evaluation of GPON vs EPON for multi-service access," *International Journal of Communication Systems*, vol. 22, pp. 187-202, 2009.
[145] H. Song, B.-W. Kim, and B. Mukherjee, "Long-reach optical access networks: A survey of research challenges, demonstrations, and bandwidth assignment mechanisms," *IEEE communications surveys & tutorials*, vol. 12, 2010.
[146] C. F. Lam, *Passive optical networks: principles and practice*: Elsevier, 2011.
[147] G. Kramer, M. De Andrade, R. Roy, and P. Chowdhury, "Evolution of optical access networks: Architectures and capacity upgrades," *Proceedings of the IEEE*, vol. 100, pp. 1188-1196, 2012.
[148] B. Batagelj, V. Erzen, J. Tratnik, L. Naglic, V. Bagan, Y. Ignatov, *et al.*, "Optical access network migration from GPON to XG-PON," in *Proc. of The Third International Conference on Access Networks ACCESS*, 2012.
[149] J. i. Kani, F. Bourgart, A. Cui, A. Rafel, M. Campbell, R. Davey, *et al.*, "Next-generation PON-part I: Technology roadmap and general requirements," *IEEE Communications Magazine*, vol. 47, pp. 43-49, 2009.
[150] A. Banerjee, Y. Park, F. Clarke, H. Song, S. Yang, G. Kramer, *et al.*, "Wavelength-division-multiplexed passive optical network (WDM-PON) technologies for broadband access: a review," *Journal of optical networking*, vol. 4, pp. 737-758, 2005.
[151] Z. Al-Qazwini and H. Kim, "Symmetric 10-Gb/s WDM-PON using directly modulated lasers for downlink and RSOAs for uplink," *Journal of Lightwave Technology*, vol. 30, pp. 1891-1899, 2012.
[152] R. Q. Shaddad, A. Mohammad, S. A. Al-Gailani, A. Al-Hetar, and M. A. Elmagzoub, "A survey on access technologies for broadband optical and wireless networks," *Journal of Network and Computer Applications*, vol. 41, pp. 459-472, 2014.
[153] N. Cvijetic, D. Qian, T. Wang, and S. Weinstein, "OFDM for next-generation optical access networks," in *Wireless Communications and Networking Conference Workshops (WCNCW), 2010 IEEE*, 2010, pp. 1-5.
[154] N. M. K. Chowdhury and R. Boutaba, "Network virtualization: state of the art and research challenges," *IEEE Communications magazine*, vol. 47, 2009.
[155] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 18, pp. 236-262, 2016.
[156] IBM, "Virtualization in education," *IBM Corporation, Whitepaper*, 2007.
[157] J. Sahoo, S. Mohapatra, and R. Lath, "Virtualization: A survey on concepts, taxonomy and associated security issues," in *Computer and Network Technology (ICCNT), 2010 Second International Conference on*, 2010, pp. 222-226.

[158] M. Rosenblum and T. Garfinkel, "Virtual machine monitors: Current technology and future trends," *Computer*, vol. 38, pp. 39-47, 2005.
[159] S. N. T.-c. Chiueh and S. Brook, "A survey on virtualization technologies," *Rpe Report*, vol. 142, 2005.
[160] IBM, "VM History and Heritage". [Online]. Available: http://www.vm.ibm.com/history/. [Accessed: 2 May 2018]
[161] J. Sugerman, G. Venkitachalam, and B.-H. Lim, "Virtualizing I/O Devices on VMware Workstation's Hosted Virtual Machine Monitor," in *USENIX Annual Technical Conference, General Track*, 2001, pp. 1-14.
[162] O. AbdElRahem, A. M. Bahaa-Eldin, and A. Taha, "Virtualization security: A survey," in *Computer Engineering & Systems (ICCES), 2016 11th International Conference on*, 2016, pp. 32-40.
[163] U. Pawar and M. Bhelotkar, "Virtualization: a way towards dynamic IT," in *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, 2011, pp. 262-263.
[164] F. Bazargan, C. Y. Yeun, and M. J. Zemerly, "State-of-the-art of virtualization, its security threats and deployment models," *International Journal for Information Security Research (IJISR)*, vol. 2, pp. 335-343, 2012.
[165] D. A. Menascé, "Virtualization: Concepts, applications, and performance modeling," in *Int. CMG Conference*, 2005, pp. 407-414.
[166] A. F. a. H. d. Meer, "Virtualization and Resilience," *Network Resilience PhD Course, ETH Zürich, September 26-28, 2011*.
[167] I. B. B. Harter, D. A. Schupke, M. Hoffmann, and G. Carle, "Network virtualization for disaster resilience of cloud services," *IEEE Communications Magazine*, vol. 52, pp. 88-95, 2014.
[168] C. Weltzin and S. Delgado, "Using virtualization to reduce the cost of test," in *AUTOTESTCON, 2009 IEEE*, 2009, pp. 439-442.
[169] E. Ayanoglu, "Editorial launching IEEE transactions on green communications and networking," *IEEE Transactions on Green Communications and Networking*, vol. 1, pp. 1-2, 2017.
[170] B. K. Joshi and C. S. Thaker, "Improving Energy Efficiency Through VM Placement and Consolidation Techniques in Cloud Computing," 2018.
[171] Y. Jin, Y. Wen, and Q. Chen, "Energy efficiency and server virtualization in data centers: An empirical investigation," in *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on*, 2012, pp. 133-138.
[172] J. Elmirghani, T. Klein, K. Hinton, L. Nonde, A. Lawey, T. El-Gorashi, *et al.*, "GreenTouch GreenMeter core network energy-efficiency improvement measures and optimization," *Journal of Optical Communications and Networking*, vol. 10, pp. A250-A269, 2018.
[173] R. Jain and S. Paul, "Network virtualization and software defined networking for cloud computing: a survey," *IEEE Communications Magazine*, vol. 51, pp. 24-31, 2013.
[174] K. Sandström, A. Vulgarakis, M. Lindgren, and T. Nolte, "Virtualization technologies in embedded real-time systems," in *Emerging Technologies & Factory Automation (ETFA), 2013 IEEE 18th Conference on*, 2013, pp. 1-8.

[175] S. R. Smoot and N. K. Tan, *Private cloud computing: consolidation, virtualization, and service-oriented infrastructure*: Elsevier, 2012.
[176] P. C. Gope and S. R. Sree, "Performance Improvement of EduCloud using a refined Virtualization Technique," *International Journal of Electronics and Computer Science Engineering (IJECSE)*, vol. 1, pp. 1970-1974, 2012.
[177] M. Rosenblum and C. Waldspurger, "I/o virtualization," *ACM Queue*, vol. 9, 2011.
[178] Y. Luo, "Network I/O virtualization for cloud computing," *IT professional*, vol. 12, pp. 36-41, 2010.
[179] D. Wu, X. Liu, S. Hebert, W. Gentzsch, and J. Terpenny, "Democratizing digital design and manufacturing using high performance cloud computing: Performance evaluation and benchmarking," *Journal of Manufacturing Systems*, vol. 43, pp. 316-326, 2017.
[180] P. Kedia, R. Nagpal, and T. P. Singh, "A survey on virtualization service providers, security issues, tools and future trends," *International Journal of Computer Applications*, vol. 69, 2013.
[181] S. Soltesz, H. Pötzl, M. E. Fiuczynski, A. Bavier, and L. Peterson, "Container-based operating system virtualization: a scalable, high-performance alternative to hypervisors," in *ACM SIGOPS Operating Systems Review*, 2007, pp. 275-287.
[182] A. Scarfo, "New security perspectives around BYOD," in *Broadband, Wireless Computing, Communication and Applications (BWCCA), 2012 Seventh International Conference on*, 2012, pp. 446-451.
[183] D. L. Lunsford, "Virtualization technologies in information systems education," *Journal of Information Systems Education*, vol. 20, p. 339, 2009.
[184] D. Messinger and G. Lewis, "Application virtualization as a strategy for cyber foraging in resource-constrained environments," *Software Engineering Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, Technical Note CMU/SEI-2013-TN-007*, 2013.
[185] G. C. Obasuyi and A. Sari, "Security challenges of virtualization hypervisors in virtualized hardware environment," *International Journal of Communications, Network and System Sciences*, vol. 8, p. 260, 2015.
[186] ETSI. [Online]. Available: https://www.etsi.org/. [Accessed: 2018]
[187] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, pp. 90-97, 2015.
[188] R. Nejabati, S. Peng, M. Channegowda, B. Guo, and D. Simeonidou, "SDN and NFV convergence a technology enabler for abstracting and virtualising hardware and control of optical networks," in *Optical Fiber Communications Conference and Exhibition (OFC), 2015*, 2015, pp. 1-3.
[189] M. Xia, M. Shirazipour, Y. Zhang, H. Green, and A. Takacs, "Network Function Placement for NFV Chaining in Packet/Optical Datacenters," *Journal of Lightwave Technology*, vol. 33, pp. 1565-1570, 2015.
[190] S. Clayman, E. Maini, A. Galis, A. Manzalini, and N. Mazzocca, "The dynamic placement of virtual network functions," in *Network Operations and Management Symposium (NOMS), 2014 IEEE*, 2014, pp. 1-9.

[191] Y. Wang, Y. Jin, W. Guo, W. Sun, and W. Hu, "Virtualized optical network services across multiple domains for grid applications," *IEEE Communications Magazine,* vol. 49, 2011.
[192] J.-P. Elbers and A. Autenrieth, "Extending network virtualization into the optical domain," in *Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC), 2013*, 2013, pp. 1-3.
[193] M.-A. Kourtis, H. Koumaras, G. Xilouris, and F. Liberal, "An NFV-based video quality assessment method over 5G small cell networks," *IEEE MultiMedia,* 2017.
[194] A. Basta, A. Blenk, K. Hoffmann, H. J. Morper, M. Hoffmann, and W. Kellerer, "Towards a cost optimal design for a 5G mobile core network based on SDN and NFV," *IEEE Transactions on Network and Service Management,* vol. 14, pp. 1061-1075, 2017.
[195] R. Martínez, A. Mayoral, R. Vilalta, R. Casellas, R. Muñoz, S. Pachnicke, *et al.*, "Integrated SDN/NFV orchestration for the dynamic deployment of mobile virtual backhaul networks over a multilayer (packet/optical) aggregation infrastructure," *Journal of Optical Communications and Networking,* vol. 9, pp. A135-A142, 2017.
[196] J. Prados-Garzon, J. J. Ramos-Munoz, P. Ameigeiras, P. Andres-Maldonado, and J. M. Lopez-Soler, "Modeling and dimensioning of a virtualized mme for 5g mobile networks," *IEEE Transactions on Vehicular Technology,* vol. 66, pp. 4383-4395, 2017.
[197] D. Evans, "The Internet of Things How the Next Evolution of the Internet Is Changing Everything," *cisco White Paper,* 2011.
[198] C. Xu, M. Wang, X. Chen, L. Zhong, and A. L. Grieco, "Optimal Information Centric Caching in 5G Device-to-Device Communications," *IEEE Transactions on Mobile Computing,* 2018.
[199] T. Han and N. Ansari, "Powering mobile networks with green energy," *IEEE Wireless Communications,* vol. 21, pp. 90-96, 2014.
[200] H. A. H. Hassan, A. Ali, L. Nuaymi, and S. E. Elayoubi, "Renewable energy usage in the context of energy-efficient mobile network," in *Vehicular Technology Conference (VTC Spring), 2015 IEEE 81st*, 2015, pp. 1-7.
[201] Y. S. Soh, T. Q. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications,* vol. 31, pp. 840-850, 2013.
[202] F. Wang, C. Xu, L. Song, Q. Zhao, X. Wang, and Z. Han, "Energy-aware resource allocation for device-to-device underlay communication," in *Communications (ICC), 2013 IEEE International Conference on*, 2013, pp. 6076-6080.
[203] F. Wang, C. Xu, L. Song, and Z. Han, "Energy-efficient resource allocation for device-to-device underlay communication," *IEEE Transactions on Wireless Communications,* vol. 14, pp. 2082-2092, 2015.
[204] S. Li, Q. Ni, Y. Sun, G. Min, and S. Al-Rubaye, "Energy-Efficient Resource Allocation for Industrial Cyber-Physical IoT Systems in 5G Era," *IEEE Transactions on Industrial Informatics,* vol. 14, pp. 2618-2628, 2018.

[205] I. B. Sofi and A. Gupta, "A survey on energy efficient 5G green network with a planned multi-tier architecture," *Journal of Network and Computer Applications*, 2018.
[206] T. Sigwele, A. S. Alam, P. Pillai, and Y. F. Hu, "Energy-efficient cloud radio access networks by cloud based workload consolidation for 5G," *Journal of Network and Computer Applications*, vol. 78, pp. 1-8, 2017.
[207] E. Oki, *Linear programming and algorithms for communication networks: a practical guide to network design, control, and management*: CRC Press, 2012.
[208] J. Bisschop, *Aimms, Optimization Modeling. Paragon Decision Technology*: Paragon Decision Technology B.V., 2001.
[209] J. Lee, *A First Course in Linear Optimization — a dynamic book —*: ReEx PrEsS, 2016.
[210] M. S. Bazaraa, J. J. Jarvis, H. D. Sherali, and W. I. O. service), *Linear programming and network flows*, 4 ed.: John Wiley & Sons Inc., 2010.
[211] J. i. Matoušek and B. Gärtner, *Understanding and using linear programming*: Springer, 2007.
[212] K. Genova and V. Guliashki, "Linear Integer Programming Methods and Approaches – A Survey," *Cybernetics And Information Technologies*, vol. 11, 2011.
[213] E. L. Lawler and D. E. Wood, "Branch-and-Bound Methods: A Survey," *Operations Research*, vol. 14, pp. 699-719, 1966.
[214] M. Pióro and D. Medhi, *Routing, flow, and capacity design in communication and computer networks*: Elsevier, 2004.
[215] J. Kallrath, *Modeling languages in mathematical optimization* vol. 88: Springer Science & Business Media, 2013.
[216] R. Fourer, D. M. Gay, and B. W. Kernighan, "A modeling language for mathematical programming," *Management Science*, vol. 36, pp. 519-554, 1990.
[217] A. Makhorin, "GLPK (GNU Linear Programming Kit), 2000," *B B*, 2014.
[218] J. Forrest, "Cbc (coin-or branch and cut) open-source mixed integer programming solver," *URL: https://projects. coin-or. org/Cbc*, 2012.
[219] IBM, "12.2 User's Manual for CPLEX," 2010.
[220] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," *Network, IEEE*, vol. 28, pp. 18-26, 2014.
[221] C.-L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: a 5G perspective," *IEEE Communications Magazine*, vol. 52, pp. 66-73, 2014.
[222] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, *et al.*, "5G on the horizon: Key challenges for the radio-access network," *IEEE Vehicular Technology Magazine*, vol. 8, 2013.
[223] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5g systems," *IEEE Communications Magazine*, vol. 52, pp. 131-139, 2014.
[224] J. Costa-Requena, J. L. Santos, V. F. Guasch, K. Ahokas, G. Premsankar, S. Luukkainen, *et al.*, "SDN and NFV integration in generalized mobile

network architecture," in *2015 European Conference on Networks and Communications (EuCNC)*, ed: IEEE, 2015, pp. 154-158.
[225] I. Giannoulakis, E. Kafetzakis, G. Xylouris, G. Gardikis, and A. Kourtis, "On the Applications of Efficient NFV Management Towards 5G Networking," in *Proceedings of the 1st International Conference on 5G for Ubiquitous Connectivity*, ed: ICST, 2014, pp. 1-5.
[226] 3GPP, "Releases". [Online]. Available: http://www.3gpp.org/specifications/67-releases. [Accessed: 3 March 2016]
[227] C. Monfreid, "The LTE Network Architecture-A Comprehensive Tutorial," *Alcatel-Lucent White Paper*, 2012.
[228] Alcatel-Lucent. Interworking LTE EPC with W-CDMA Packet Switched Mobile Cores. *Alcatel-Lucent White Paper*. Available: http://www.alcatel-lucent.com
[229] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G Backhaul Challenges and Emerging Research Directions: A Survey," *IEEE Access*, vol. 4, pp. 1743-1766, 2016.
[230] P. Chanclou, A. Pizzinat, F. Le Clech, T. L. Reedeker, Y. Lagadec, F. Saliou, *et al.*, "Optical fiber solution for mobile fronthaul to achieve cloud radio access network," ed, 2013, pp. 1-11.
[231] Cisco. Cisco ME 4600 Series Optical Line Terminal Data Sheet [Online]. Available: http://www.cisco.com/c/en/us/products/collateral/switches/me-4600-series-multiservice-optical-access-platform/datasheet-c78-730445.html
[232] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, "IP Over WDM Networks Employing Renewable Energy Sources," *Journal of Lightwave Technology*, vol. 29, pp. 3-14, 2011.
[233] S. Electric, "FTE7502 EPON Optical Network Unit (10G ONU) datasheet". [Online]. Available: http://www.sumitomoelectric.com/onu-fte7502.html. [Accessed: 21 Feb 2015]
[234] Alcatel-Lucent. TRDU2x40-08 LTE 3GPP Band 20 LTE FDD Transmit Receive Duplexer Unit – 800 MHz EDD Datasheet [Online].
[235] A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Distributed Energy Efficient Clouds Over Core Networks," *Journal of Lightwave Technology*, vol. 32, pp. 1261-1281, 2014.
[236] A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Renewable energy in distributed energy efficient content delivery clouds," presented at the IEEE International Conference on Communications (ICC), 2015.
[237] GreenTouch, "GreenTouch Final Results from Green Meter Research Study", 2015. [Online]. Available: http://www.greentouch.org/index.php?page=greentouch-green-meter-research-study. [Accessed: 7 Jan 2016]
[238] L. Sumitomo Electric Industries, "FSU7100 Series OLT datasheet". [Online]. Available: http://global-sei.com/nws/01/index_01_1-gepon.html#specifications7101. [Accessed: 21 Feb 2015]
[239] Intel, "Intel® Pentium® III Processor 800 MHz, 256K Cache, 100 MHz FSB."
[240] Intel, "Intel® Pentium® M Processor 750."
[241] Intel, "Intel® Pentium® 4 Processor 661 supporting HT Technology."

[242] Z. Tayq, "Fronthaul integration and monitoring in 5G networks," Université de Limoges, 2017.
[243] S. Little, "Is microwave backhaul up to the 4G task?," *IEEE microwave magazine*, vol. 10, 2009.
[244] A. Pizzinat, P. Chanclou, F. Saliou, and T. Diallo, "Things you should know about fronthaul," *Journal of Lightwave Technology*, vol. 33, pp. 1077-1083, 2015.
[245] R. Chundury, "Mobile broadband backhaul: Addressing the challenge," *Planning Backhaul Networks, Ericsson Review*, pp. 4-9, 2008.
[246] Alcatel-Lucent. LTE Mobile Transport Evolution. *Alcatel-Lucent White Paper*. Available: http://www.alcatel-lucent.com
[247] R. Kwan and C. Leung, "A survey of scheduling and interference mitigation in LTE," *Journal of Electrical and Computer Engineering*, vol. 2010, p. 1, 2010.
[248] H. G. Myung, "Technical overview of 3GPP LTE," *Polytechnic University of New York*, 2008.
[249] C. Hoymann, W. Chen, J. Montojo, A. Golitschek, C. Koutsimanis, and X. Shen, "Relaying operation in 3GPP LTE: challenges and solutions," *IEEE Communications Magazine*, vol. 50, 2012.
[250] J. Zyren, "Overview of the 3GPP long term evolution physical layer," *White Paper*, 2007.
[251] Anritsu, "LTE Resource Guide", 2015. [Online]. Available: http://www.cs.columbia.edu/6181/hw/anritsu.pdf. [Accessed: 7 May 2018]
[252] R. F. Chisab and C. Shukla, "Performance Evaluation Of 4G-LTE-SCFDMA Scheme Under SUI And ITU Channel Models," *International Journal of Engineering & Technology IJET-IJENS*, vol. 14, 2014.
[253] M. Rinne and O. Tirkkonen, "LTE, the radio technology path towards 4G," *Computer Communications*, vol. 33, pp. 1894-1906, 2010.
[254] P. Edström, "Overhead impacts on long-term evolution radio networks," Master of Science Thesis Stockholm, Sweden, 2007.
[255] A. de la Oliva, J. A. Hernández, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Communications Magazine*, vol. 54, pp. 152-159, 2016.
[256] CPRI Specification V7. 0, Oct 2015 [Online]. Available: http://www.cpri.info/downloads/CPRI_v_7_0_2015-10-09.pdf
[257] J. P. Castro, *The UMTS network and radio access technology*: John Wiley Sons Limited, 2001.
[258] H. Kaaranen, *UMTS networks: architecture, mobility and services*: John Wiley & Sons, 2005.
[259] T. Werthmann, H. Grob-Lipski, and M. Proebster, "Multiplexing gains achieved in pools of baseband computation units in 4G cellular networks," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*, 2013, pp. 3328-3333.
[260] A. Q. Lawey, T. El-Gorashi, and J. M. Elmirghani, "Energy efficient cloud content delivery in core networks," in *IEEE Globecom Workshops (GC Wkshps)*, 2013, pp. 420-426.

[261] I. Waßmann, D. Versick, and D. Tavangarian, "Energy consumption estimation of virtual machines," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 2013, pp. 1151-1156.
[262] Cisco, "Cisco ASR 5000 Series Product Overview Release 12.0," 2013.
[263] Alcatel-Lucent, "Alcatel-Lucent 9926 Base Band Unit LR13.1.L," 2013.
[264] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, *et al.*, "How much energy is needed to run a wireless network?," *IEEE Wireless Communications*, vol. 18, pp. 40-49, 2011.
[265] Intel, "Export Compliance Metrics for Intel® Microprocessors," 2018.
[266] L. Nonde, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy Efficient Virtual Network Embedding for Cloud Networks," *Journal of Lightwave Technology*, vol. 33, pp. 1828-1849, 2015.
[267] J. Xu, K. Ota, and M. Dong, "Saving Energy on the Edge: In-Memory Caching for Multi-Tier Heterogeneous Networks," *IEEE Communications Magazine*, vol. 56, pp. 102-107, 2018.
[268] M. Li and H.-L. Chen, "Energy-Efficient Traffic Regulation and Scheduling for Video Streaming Services Over LTE-A Networks," *IEEE Transactions on Mobile Computing*, 2018.
[269] Cisco. Cisco Content Delivery Engine 250 [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/video/content-delivery-engine-series/data_sheet_c78-635849.html
[270] Fujitsu. FUJITSU PLAN EP MCX4-EN 100GbE QSFP28 Data Sheet [Online]. Available: https://sp.ts.fujitsu.com/dmsp/Publications/public/ds-py-plan-EP-MCX4-EN-100-QSFP28.pdf
[271] X. Li, X. Wang, K. Li, and V. C. Leung, "CaaS: Caching as a service for 5G networks," *IEEE Access*, vol. 5, pp. 5982-5993, 2017.