Design and optimization of next-generation carrier-grade wi-fi networks

Fatma Ben Jemaa

Doctor of Philosophy
UPMC Sorbonne Universités

Specialization

COMPUTER SCIENCE
(École Doctorale Informatique, Télécommunication et Électronique "EDITE de Paris")

Presented by

Ms Fatma Ben Jemaa

Submitted for the degree of

Doctor of Philosophy of UPMC Sorbonne Universités

Title:

Design and Optimization of Next-Generation Carrier-Grade Wi-Fi Networks

Defense: 27^{th} September 2016

Committee:

Ms Catherine Rosenberg	**Reviewer**	**Professor, University of Waterloo - Canada**
Mr Sami Tabbane	**Reviewer**	**Professor, SupCom Engineering School – Tunis - Tunisia**
Mr Maurice Gagnaire	**Examiner**	**Professor, Télécom ParisTech – Paris - France**
Mr Emmanuel Bertin	**Examiner**	**Senior Service Architect, Orange Labs – Caen - France**
Mr Rami Langar	**Examiner**	**Associate professor - HDR, UMPC Sorbonne Universités – Paris - France**
Mr Guy Pujolle	**Supervisor**	**Professor, UMPC Sorbonne Universités – Paris - France**
Mr Michel Pariente	**Supervisor**	**CEO and R&D Director, Meteor Network – Vitry sur Seine - France**

UPMC
SORBONNE UNIVERSITÉS

Thèse de Doctorat de
l'Université Pierre et Marie Curie - Sorbonne Universités

Spécialité

INFORMATIQUE
(École Doctorale Informatique, Télécommunication et Électronique "EDITE de Paris")

Présentée par

Mme Fatma Ben Jemaa

Pour obtenir le grade de

Docteur de l'Université Pierre et Marie Curie - Sorbonne Universités

Sujet de la Thèse :

Conception et Optimisation des Réseaux Wi-Fi Opérateur de Nouvelle Génération

Soutenance : le 27 Septembre 2016

Jury :

Mme Catherine Rosenberg	**Rapporteur**	**Professeur, Université de Waterloo - Canada**
M. Sami Tabbane	**Rapporteur**	**Professeur, École Supérieure des Communications (Sup'Com) – Tunis - Tunisie**
M. Maurice Gagnaire	**Examinateur**	**Professeur, Télécom ParisTech – Paris - France**
M. Emmanuel Bertin	**Examinateur**	**Architecte Service Senior, Orange Labs – Caen - France**
M. Rami Langar	**Examinateur**	**Maître de Conférences - HDR, UMPC Sorbonne Universités – Paris - France**
M. Guy Pujolle	**Directeur de Thèse**	**Professeur, UMPC Sorbonne Universités – Paris - France**
M. Michel Pariente	**Co-directeur de Thèse**	**CEO et Directeur R&D, Meteor Network – Vitry sur Seine - France**

Abstract

Over the past few years, Wi-Fi networks have been extensively deployed and have significantly evolved with the emergence of new technologies and services. Moreover, Wi-Fi is becoming an integral strategic component of wireless carriers' networks and is gaining a lot of momentum in future 5G networks. In this context, new carrier-grade requirements have to be ensured to provide a high-quality user experience and high-performance Wi-Fi networks. Face to these new challenges, we investigate, in this thesis, several issues related to the design and optimization of carrier-grade next-generation Wi-Fi networks. These issues are addressed from both the user perspective and the carrier perspective.

In the first stage, our objective is to improve the Wi-Fi user experience and offer him a personalized and seamless access to Wi-Fi networks and services. For this, we propose an extension to the IEEE 802.11 management frames to enable venue service discovery prior to Wi-Fi association while avoiding channel overhead. We also define a set of extensible service labels to uniquely and globally identify the most known venue-based services. These labels can be used in the proposed extension to advertise services in the Wi-Fi networks. Through deep analysis and comparison to existing solutions, we show that these two proposals are more efficient and help provide transparent and automated access to Wi-Fi venue-based services based on user preferences and context, thus satisfying one of the major user expectations in future carrier-managed Wi-Fi networks.

Our interests move in the second stage into dealing with network architecture and management issues in next-generation carrier Wi-Fi environment. More specifically, we first propose a novel carrier-managed Wi-Fi architecture that leverages Network Function Virtualization and Edge Cloud Computing concepts. The idea behind this architecture is to i) bring more agility and adaptability, ii) allow operators to easily implement new services while reducing CapEx and OpEx, and iii) improve user-perceived QoS by placing network functions and certain services close to end-users. This proposal is validated through a proof-of-concept implementation which provides good performances.

To address some major management issues in this architecture, we then propose placement and provisioning strategies of Virtual Network Functions (VNFs) based on QoS requirements. These strategies can also be applied to any wireless carrier's edge-central architecture, since they do not make any assumption about the underlying wireless technology. Simulation results show how a fair trade-off between two conflicting objectives, namely the optimization of resource utilization and the minimization of SLA violations is achieved. Moreover, a satisfactory level of overall QoS is ensured.

Key Words

Carrier-grade Wi-Fi, Next-Generation networks, User experience, Service discovery, WLAN Cloudlet, Edge-Central architecture, NFV, VNF placement and provisioning, QoS.

Résumé en Français

Comme le Wi-Fi est devenu de plus en plus important dans les réseaux actuels, ainsi que dans les réseaux du futur, de nouvelles exigences «opérateur» se sont apparues afin de supporter les attentes des utilisateurs et de fournir des réseaux Wi-Fi de haute performance. Dans ce contexte, nous étudions plusieurs problèmes liés à la conception et l'optimisation des réseaux Wi-Fi opérateur de nouvelle génération. Ces problèmes sont traités du côté utilisateur et du côté opérateur.

Dans la première étape, notre objectif est d'améliorer l'expérience utilisateur et de lui offrir un accès personnalisé et transparent aux réseaux et services Wi-Fi. Pour ce faire, nous proposons une extension des trames de gestion IEEE 802.11 pour activer la découverte des services locaux avant l'association Wi-Fi, tout en évitant la surcharge du canal radio. Nous définissons également un ensemble d'étiquettes de service pour identifier d'une manière standardisée les services les plus connus. A travers une analyse comparative avec les solutions existantes, nous montrons que ces deux propositions sont plus efficaces et aident à fournir un accès transparent et automatisé aux services Wi-Fi locaux en fonction des préférences et du contexte de l'utilisateur. Ceci permet de satisfaire l'une des principales attentes des utilisateurs dans les futurs réseaux Wi-Fi opérateur.

Nos intérêts s'orientent dans la deuxième étape vers le traitement des problèmes liés à l'architecture et la gestion du réseau dans un environnement Wi-Fi opérateur de nouvelle génération. Plus précisément, nous proposons, tout d'abord, une nouvelle architecture Wi-Fi qui exploite les concepts de NFV et du Edge Cloud Computing. Nous visons à travers cette architecture à $i)$ apporter plus d'agilité et d'adaptabilité, $ii)$ permettre aux opérateurs d'introduire des nouveaux services avec des coûts réduits et $iii)$ améliorer la QoS perçue par l'utilisateur en plaçant des fonctions réseau et certains services à proximité. Cette proposiion est validée par l'implémentation d'une preuve du concept et de bonnes performances sont obtenues.

Pour faire face à certains problèmes majeurs de gestion dans cette architecture, nous proposons, ensuite, des stratégies de placement et de provisionnement des fonctions de réseau virtuelles en s'appuyant sur des exigences de QoS. Ces stratégies peuvent également être appliquées à toute architecture d'un opérateur sans-fil basée sur des clouds de bords et centralisés. Les résultats de simulation montrent qu'un compromis équitable entre deux objectifs contradictoires, à savoir l'optimisation de l'utilisation des ressources et la minimisation des violations du SLA, est atteint. De plus, un niveau satisfaisant de QoS globale est assurée.

Table of Contents

Chapter 1

General Introduction

Contents

Over the last few years, wireless networks have experienced a strong growth of mobile data and video traffic. According to the Cisco Visual Networking Index (VNI) [1], the global mobile data traffic grew 74 percent in 2015 reaching 3.7 Exabytes per month at the end of 2015, up from 2.1 Exabytes per month at the end of 2014. Furthermore, it is expected that the overall mobile data traffic will grow to 30.6 Exabytes per month by 2020, an eightfold increase over 2015 (see Figure 1.1).

The increasing number of mobile devices (e.g., smartphones, tablets, netbooks, laptops and gaming consoles) is one of the major contributors to this growth. More than half a billion mobile devices were added in 2015 making a total of 7.9 billion [1]. In addition, these devices are widely used to access bandwidth-intensive services such as high and ultra-high definition video, video-on-demand, two-way video conferencing and online gaming services.

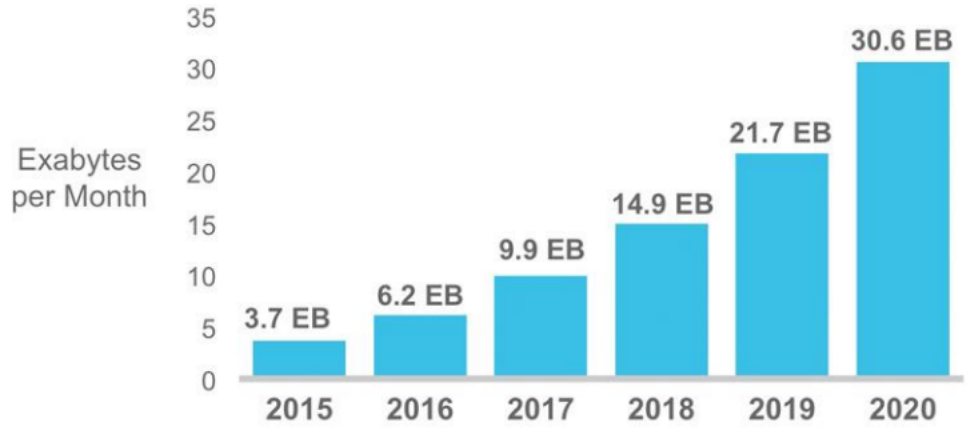


Figure 1.1: Global mobile data traffic: 2015 to 2020 [1]

To support this insatiable data demand, leveraging only cellular networks will not be sufficient. Indeed, deploying more cell sites is cost-prohibitive due to the incremental costs of backhaul and Radio Access Network (RAN). Moreover, technical enhancements (e.g., small cells, HSPA+, LTE Advanced) can bring significant benefits, but not enough to satisfy the exponential data demand growth. Consequently, Wi-Fi has been identified as a substantial complementary solution to cellular networks as it can give operators the capacity boost they need and increase the densification of their networks, either in a separate underlay network or alongside cellular small cells. Indeed, Wi-Fi represents an attractive solution to fulfill consumers' data demand due to its relatively low cost, simple architecture and its harmonized global spectrum allocation. In addition, Wi-Fi represents a widely adopted technology standard and is available on all data-centric devices. The Wi-Fi Alliance (WFA) announced in early this year that Wi-Fi device shipments have reached 12 billion units, and are expected to surpass 15 billion units by the end of 2016 [2].

From the other hand, nowadays users expect ubiquitous broadband access to their content and applications everywhere and every time, not only at home or at work but also in entertainment places and even when traveling. For that reason, Wi-Fi hotspots are increasingly being available in all public and semi-public areas such as restaurants, coffee shops, hotels, airports, stadiums and shopping malls. These hotspots are deployed by cable operators and mobile operators to offer fee-based and free Internet Wi-Fi access. In addition, Wi-Fi is being largely deployed by cities around the world and is attracting many investments from companies such as Google and Facebook to build smart cities. According to the WFA, global public hotspots count 47 million in early 2016. This includes homespots, which are residential access points. According to the Cisco VNI, commercial hotspots are a smaller subset of the overall public Wi-Fi hotspot forecast and will grow from 7.5 million in 2015 to 9.3 million by 2020.

It is worth mentioning that Wi-Fi represents a significant opportunity for mobile and fixed car-

riers as well as third-party service providers to provide a cost-effective broadband access, embrace traffic growth, enable new revenues, offer innovative services, enhance customer satisfaction, and improve brand loyalty. One of the key reasons behind the success of Wi-Fi technology is the constant evolution and innovation to keep pace with connectivity requirements of existing and emerging markets. Many advancements on core Wi-Fi technologies and standardization efforts have marked this Wi-Fi evolution. Besides, as a higher quality of user experience is becoming increasingly important in Wi-Fi networks, a new industry effort, called "carrier-grade Wi-Fi", has emerged to improve Wi-Fi network design, management, and performance to have a cellular-like quality. All these advancements and efforts represent a step towards next generation Wi-Fi networks and further enhancements are required to carry the technology well into the future.

In this context, we conducted our research work to address the design and optimization of next-generation carrier-grade Wi-Fi networks. In what follows (section 1.1), we provide an overview of the emerging technologies and trends that have revolutionized Wi-Fi landscape and can be considered as a first milestone towards 5G networks. Then, section 1.2 outlines the basic attributes of a carrier-grade Wi-Fi network. Next, we present our thesis problematics and our contributions, respectively in section 1.3 and 1.4. Finally, section 1.5 describes the organization of this manuscript.

1.1 Evolution of Wi-Fi Networks towards 5G

In this section, we show how Wi-Fi is gaining a lot of momentum in today's networks as well as in future 5G networks. We first describe the Wi-Fi offload and roaming and their potential role in current and future networks. Second, we present the emergence of Voice over Wi-Fi as a complementary solution to Voice over LTE (VoLTE). Finally, we describe the existing and on-going standards in Wi-Fi addressing 5G requirements.

1.1.1 Wi-Fi Offload and Roaming

Wi-Fi has become an integral part of the access strategy and architecture of many operators around the world alongside their cellular networks or as an alternative broadband wireless ecosystem. As a strategic technology in the current 4G landscape, Wi-Fi represents a complementary solution to mobile networks and plays a significant role in Heterogeneous Networks (HetNets), both in terms of data offloading and roaming. It is widely argued that Wi-Fi will also take an even more integrated role in the next 5G networks and the deployment of HetNets will be a key to support throughput, coverage and capacity needs.

Driven by the surging wireless data demands and the proliferation of Wi-Fi enabled devices, Wi-Fi offloading has become a major solution for mobile operators to alleviate network congestion. It permits to offload data traffic from licensed capacity-constrained networks to unlicensed Wi-Fi networks. According to the Cisco VNI [1], mobile offload to Wi-Fi as well as small-cell networks

will increase from 51 percent (3.9 exabytes/month) in 2015 to 55 percent (38.1 exabytes/month) by 2020 of the total mobile data traffic from all mobile-connected devices (see Figure 1.2).

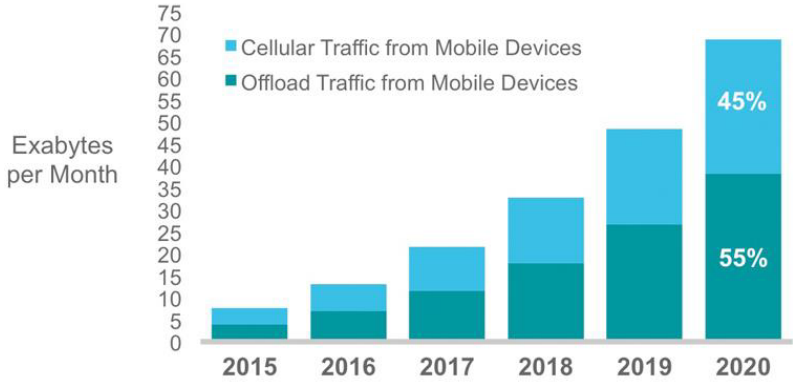


Figure 1.2: Mobile data offload over Wi-Fi and small cells: 2015 to 2010 [1]

In order to enable Wi-Fi and cellular integration and interoperability and provide a standardized architectural framework for Wi-Fi offloading, 3GPP has developed several solutions such as:

- Access Network Discovery and Selection Function (ANDSF) which enables operators to set traffic management policies (introduced in Release 8),
- Mobility with IP address preservation for selected IP flows (IFOM) (Release 10),
- Transparent IP connectivity via trusted Wi-Fi using GPRS Tunneling Protocol (SaMOG) (Release 11),
- LTE Wi-Fi Aggregation (Release 13) which is based on data aggregation at the radio access network, where an Evolved NodeB schedules packets to be served on LTE and Wi-Fi radio links.

From the other hand, seamless Wi-Fi roaming is recently launched due to Passpoint, the WFA certification program based on the Hotspot 2.0 specification [3, 4]. This enables consistent, secure and automated connectivity to Wi-Fi hotspots worldwide. In addition, the Wireless Broadband Alliance (WBA) has launched the Next Generation Hotspot (NGH) initiative to establish roaming best practices for Wi-Fi and facilitate the creation of roaming partnerships between carriers around the world. The intent of these efforts is to improve Wi-Fi user experience, fulfill service provider business objectives and thereby move towards delivering “carrier-grade” Wi-Fi solutions. From the end-user side, a seamless access to Wi-Fi hotspots as easy and secure as cellular network access is ensured. From service providers side, they will be able to extend their global network footprint through roaming agreements, increase their revenues with improved subscriber satisfaction, and better monetize their hotspots.

1.1.2 Voice over Wi-Fi

Wi-Fi is considered by mobile operators not only as an offload solution of data traffic but also as a supplement to cellular voice. Closely aligned with VoLTE, the new technology Voice over Wi-Fi (VoWiFi), also known as Wi-Fi calling, uses the same IP Multimedia Subsystem (IMS) core infrastructure, carries voice and video calls over Wi-Fi, and supports mobility between LTE and Wi-Fi accesses. Thus, a seamless user experience is guaranteed. To ensure interoperability and high-quality IMS-based telephony services over Wi-Fi access networks, a wireless device and network are required to implement a set of mandatory features defined by 3GPP and GSMA specifications [5].

The primary objective of using VoWiFi is to provide a better indoor coverage where cellular coverage is limited and access to Wi-Fi hotspots are wider and more optimum. Operators also see a value in offering Wi-Fi calling services as a differentiator from Over-The-Top (OTT) VoIP solutions. Moreover, this technology can reduce roaming costs with access to the IMS telephony services of their home network over Wi-Fi networks.

Recently, there has been a huge interest in the next generation Wi-Fi calling all around the world. Indeed, many mobile carriers (e.g., T-Mobile, AT&T, Sprint and Verizon in the USA, Vodafone and EE in the UK) have already launched VoWiFi. In France, Orange plans to launch this service during 2016 and early 2017. Wi-Fi calling is also adopted by big companies such as Google which launched last year its Project Fi where most calls are handled over Wi-Fi instead of a cellular network. According to the Cisco VNI [1], VoWiFi is going to surpass VoLTE by 2016 and VoIP by 2018 in terms of minutes of use. By 2020, VoWiFi will have 53 percent of mobile IP voice, up from 16 percent in 2015 (see Figure 1.3).

1.1.3 5G-oriented Technologies

Using advanced technologies, the fifth generation networks are expected to provide higher throughput, low latency, better Quality of Service (QoS) and security, widespread connectivity, and energy efficiency, among others. Wi-Fi has already multiple standards and on-going works which address 5G requirements.

- **802.11ac**: is the latest generation of Wi-Fi which operates in the 5 GHz band and delivers up to 1 Gbps of data rates. It offers better performances in terms of capacity, power management, and latency. It is designed to meet today's demanding applications while paving the way for new services. The major enhancements of 802.11ac are at the physical layer and include the use of wider channels (up to 160 MHz), improved modulation (up to 256 QAM), increased number of spatial streams (up to 8), Multi-User MIMO, and beamforming.
- **802.11ah** (HaLow): operates in the 900 MHz band, offering longer range and lower power connectivity necessary for applications including sensors and wearables. Wi-Fi HaLow will

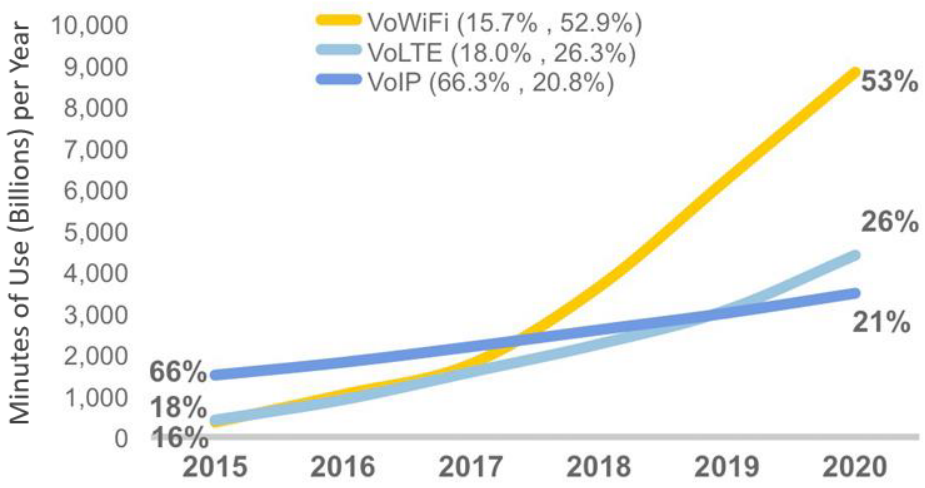


Figure 1.3: Mobile voice use: VoWiFi, VoLTE, and VoIP [1]

enable a variety of new and future power-efficient use cases in Internet of Things (IoT) environments such as smart home, smart city, connected car, and digital healthcare.

- **802.11ad** (WiGig): extends Wi-Fi to the 60 GHz frequency band to provide very high throughput (multi-gigabit) and low latency between nearby devices. It provides better performance for video and other high-throughput applications especially for line-of-site connections and high-density environments.

- **802.11ax** High Efficiency WLAN (HEW): is an on-going IEEE project and will be the successor of 802.11ac. Its objective is to improve spectrum efficiency and enhance the system throughput in high density areas such as hotspots in public venues. A key change in 802.11ax will be the jointly use of Multiple Input Multiple Output (MIMO) and Orthogonal Frequency Division Multiple Access (OFDMA).

- **802.11ay**: is the next generation wireless transmission standard in the 60 GHz band which is expected to be completed in 2017. The 801.11ay is expected to support a maximum throughput of at least 20 Gbps while maintaining or improving the power efficiency per station. It will also ensure backward compatibility and coexistence with legacy 802.11ad stations operating in the same band.

1.2 Carrier-grade Wi-Fi Networks

Wi-Fi technology is becoming a viable integral component of operators' mobile broadband strategies. Thus, they will have to deploy a high-performance carrier-grade networks that are scalable and able to support customers' demand and to provide a high-quality user experience. To achieve these objectives, carrier Wi-Fi networks should provide, as depicted in Figure 1.4, three basic high-level attributes: i) consistent user experience; ii) a fully integrated end-to-end network; and iii) network management capabilities. These attributes are defined by the WBA, one of Wi-Fi's standards-setting bodies, as a set of requirements that Wi-Fi networks need to meet in order to be branded 'carrier-grade' [6].

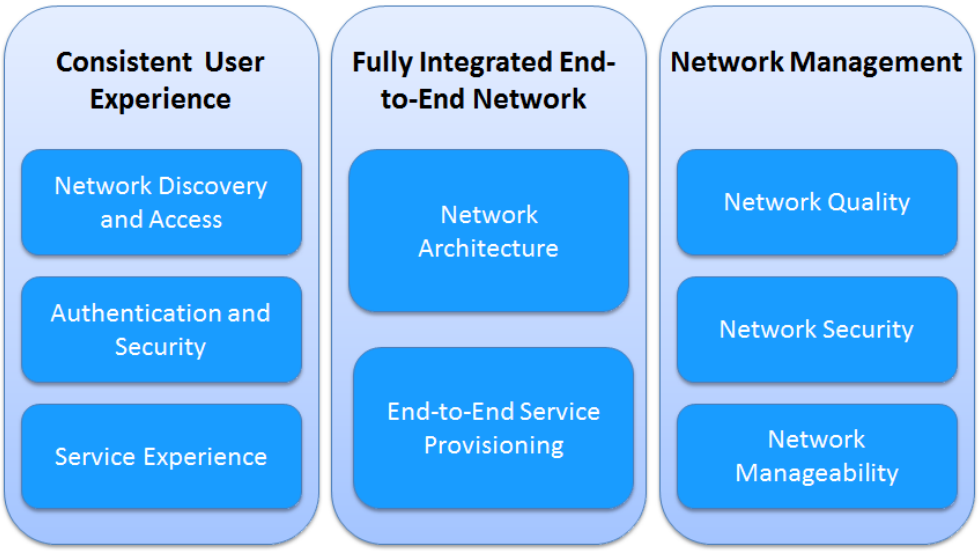


Figure 1.4: Carrier-grade Wi-Fi basic attributes [6]

From the user perspective, in carrier-managed Wi-Fi networks, a consistent user experience is expected with service and performance levels exceeding that provided in "existing legacy Wi-Fi". Indeed, devices should be able to automatically manage network discovery and access and to be securely connected to a Wi-Fi network. Furthermore, users expect a seamless and high-performance service experience with support of mobility either across Access Points (APs) within the Wi-Fi network or between Wi-Fi and cellular access. All these features are referred to as the NGH, which are based principally on the WFA Hotspot 2.0 technology (Passpoint) [3,4] whose main purpose is to enhance the Wi-Fi user experience and make Wi-Fi as easy to use and secure as cellular networks.

From the operator perspective, the Wi-Fi carrier should first provide a fully integrated end-to-end network based on an evolved architecture supporting a set of features (e.g., standard interfaces, IPv6 and scalability) and ensuring an end-to-end service provisioning (i.e., service subscription,

operator policy provisioning and enforcement, and accounting and charging). Second, an integral network management system is needed for a carrier Wi-Fi network to ensure network quality, security and manageability.

1.3 Problem Statement

Our work is carried out in collaboration with **Meteor Network** [7], one of the leading public Wi-Fi service providers in France. In this context, we investigate several problems surrounding the design and optimization of **carrier-grade Wi-Fi** networks that support future user expectations and next-generation network requirements. Both user and carrier perspectives are addressed.

- **User perspective**

We address Wi-Fi user experience to provide more personalization and transparency for access to Wi-Fi networks and services. Specifically, we are interested in Venue-Based Services that are increasingly deployed by venue owners and Wi-Fi operators in their networks to improve their users' experience and generate new revenue streams. These services are locally offered by the venue and client mobile devices should typically authenticate and associate to Wi-Fi networks to be able to discover these services. However, end-users often prefer to associate to a Wi-Fi network only if it offers the specific service they need. Moreover, they no longer want to manually search for and choose a network which offers the desired service. Therefore, a personalized rich user experience and a seamless access to Wi-Fi networks and services are becoming major user expectations in next-generation carrier-managed Wi-Fi networks.

- **Carrier perspective**

First, we address Wi-Fi network architecture that has largely been static and difficult to evolve while newer Wireless Local Area Network (WLAN) technologies and services have been emerging at a prolific rate. This architecture needs to be flexible to support a wide range of services and be able to easily adapt to future capabilities. Moreover, scalability has to be considered in the design of such architecture to support the increasing user demand and future network requirements. These characteristics have to be ensured while reducing network deployment and operation cost.

We consider Network Function Virtualization (NFV) as a promising solution in our carrier-grade architecture. Thus, we address network management and provisioning in an NFV-based carrier Wi-Fi architecture taking into account QoS requirements of network functions (e.g., response time and delay sensitivity), resource availability and network status (e.g., network delay, virtualization overhead, etc.). Optimizing these metrics represents a major concern in next-generation carrier-grade Wi-Fi networks to improve the user-perceived QoS and network performance and reduce resource utilization costs.

1.4 Contributions

In this thesis, we address three major challenges in carrier-grade next-generation Wi-Fi networks. The first challenge is related to the user experience and access to Wi-Fi services in such networks. From the wireless carrier's perspective, the two other challenges deal with architecture and network management issues. In the following, we summarize the significant contributions related to these challenges.

1.4.1 Personalized and Seamless Access to Wi-Fi Services

In this contribution, we present a solution to provide a personalized and seamless access to venue-based services offered through public Wi-Fi networks. Our objective is to improve Wi-Fi user experience and deliver to him the services he desires, where and when he desires. To do so, we propose an extension to IEEE 802.11 management frames to enable venue service discovery prior to Wi-Fi association while avoiding channel overhead. We also define a set of extensible service labels to uniquely and globally identify the most known venue-based services. These labels can be used in the proposed extension to identify advertised services in the Wi-Fi networks. Through deep analysis and comparison to existing solutions, we show how these two proposals could be efficient and useful and help to have a personalized, transparent and automated access to Wi-Fi venue-based services based on user preferences and context. This contribution is the object of a publication [8] and a proposal for an IETF draft [9].

1.4.2 NFV- and Cloudlet-based Carrier Wi-Fi Architecture

In this contribution, we propose a novel architecture for carrier-managed Wi-Fi networks that leverages Network Function Virtualization and Edge Cloud Computing concepts. We aim through this architecture to bring more flexibility and adaptability and allow operators to easily implement new services while reducing CapEx and OpEx costs. We also aim to decrease access latency by placing network functions and certain services close to end-users. To achieve these objectives, we introduce a new architecture element, called WLAN Cloudlet located in the end-user premises, that offloads MAC layer processing from APs and consolidates network functions and value-added services. All these functions and services are based on software instances. To prove the feasibility and evaluate the performance of our proposal, we develop a proof-of-concept prototype and compare it with a reference architecture. Results show that the WLAN Cloudlet solution achieves good performances, while providing at the same time many advantages in terms of cost, flexibility, and agility. This contribution is the object of publication [10] and [11].

1.4.3 VNF Placement and Provisioning in Carrier Wi-Fi Network

In this contribution, we propose placement and provisioning strategies of Virtualized Network Functions (VNFs) in the Wi-Fi carrier Cloudlet-based architecture taking into account QoS requirements. The main goals of these strategies are i) to optimize resource utilization, ii) to prevent cloudlet overload and congestion, and iii) to avoid violation of QoS and SLA requirements. For this purpose, we model performance of our system using analytical and QoS models taking into account virtualization overhead, resource workloads and availability, network delay and real-time requirements. Then, we formulate the VNF placement and provisioning problem as a Mixed-Integer Linear Program (MILP). Our problem represents a Multiple Objective Decision Making (MODM) based on two conflicting objectives, namely i) the optimization of resource utilization, especially in the capacity-constrained cloudlet system, and ii) the minimization of SLA violations. To resolve this problem, we propose three different solutions to achieve a trade-off between these conflicting objectives according to the wireless operator requirements. These solutions are evaluated through extensive simulations and encouraging results are obtained. This contribution is the object of publication [12].

1.5 Thesis Outline

This thesis is organized into three principal parts, preceded by a general introduction and followed by a conclusion.

- In the introduction chapter, **Chapter 1**, we describe the context, motivations and problems addressed in this thesis.
- **Chapter 2** represents the first part dealing with improving the user and service experience in Wi-Fi carrier networks.
- In the second part, **Chapter 3**, we present our proposed architecture for next-generation and carrier-grade Wi-Fi networks based on NFV and Edge Cloud Computing paradigms.
- In the third part, **Chapter 4**, we address management issues of virtualized network functions in this architecture. More specifically, we focus on the optimization of the placement and provisioning of these functions.
- Finally, **Chapter 5** concludes the thesis and presents perspectives for future work.

Chapter 2

Service Discovery and Access in Carrier Wi-Fi Networks

Contents

2.1 Introduction

Wi-Fi networks are increasingly deployed in public and semi-public areas such as hotels, shopping malls, and airports as well as smaller venues like restaurants and coffee shops. As Wi-Fi in public hotspots has grown in availability and popularity, adding value to wireless service offerings and enhancing user experience with context-aware services have created new opportunities for venue owners and Wi-Fi operators to boost their customers' satisfaction and generate new revenue

streams. These services, called Venue Based Service (VBS), are related to the type of the venue and are tailored to the special needs of mobile users in this venue. They aim to offer visitors with relevant and specific information (e.g., indoor way finding and discounts) and local services (e.g., printer).

Traditionally, the discovery of these services requires user authentication and association to a Wi-Fi network and manual intervention to identify services offered by the network. This may also require frequent and tedious manual user intervention to search and choose the appropriate network offering the desired service. Indeed, since the user may not know in advance which network provides the desired service and since multiple networks may be available, this will lead to long time to connect and high power consumption.

To address these issues, this chapter provides pre-association service discovery mechanisms in Wi-Fi networks with the aim of ensuring more personalization and transparency for access to Wi-Fi networks and services in a power efficient manner.

After reviewing related work from the literature in section 2.2, we describe our proposal which provides lightweight discovery mechanisms of venue services prior to Wi-Fi association and enables global and extensible identification and description scheme of these services in section 2.3. Following that, we analyze and compare our proposed solution with existing solutions in section 2.4. In section 2.5, we present use case scenarios highlighting the usability of our solution. Finally, we conclude this chapter in section 2.6.

2.2 User Experience and Service Discovery in Wi-Fi Networks: State of the Art

In this section, we first describe the most recent standards which aim to improve the Wi-Fi user experience (section 2.2.1). Second, we outline the different service discovery approaches according to which we classify the existing protocols and solutions (section 2.2.2). Finally, we summarize the reviewed solutions (section 2.2.3).

2.2.1 Existing Standards

Hereafter, we describe the most recent and relevant standards related to the user experience improvement in Wi-Fi networks, namely i) IEEE 802.11u and Hotspot 2.0 and ii) Wi-Fi Aware.

2.2.1.1 IEEE 802.11u and Hotspot 2.0 (Network Discovery)

Recent standards and technologies, such as IEEE 802.11u [13] and Hotspot 2.0 (Wi-Fi Alliance) [3, 4], have been developed to deliver a seamless Wi-Fi access and to enable automated network discovery and selection without any active intervention from the user. A key innovation of these

technologies is to provide information to a mobile device about a Wi-Fi network before deciding to join it. Particularly, this helps to assist the mobile device in selecting a network.

For this purpose, 802.11u focuses on enhancing network discovery by adding new beacon and probe response information elements (e.g., Extended capabilities, Interworking, and Roaming Consortium). In addition, it introduces a new pre-association protocol, called Access Network Query Protocol (ANQP), used by Wi-Fi client devices to query the hotspot for additional network information (e.g., Venue Name and Network Authentication type) that is not advertised in beacon and probe response frames. These queries are transported using Generic Advertisement Service (GAS), an IEEE 802.11 service that provides over-the-air transportation for frames of higher layer advertisements. The Wi-Fi Alliance has extended ANQP protocol with its own Hotspot 2.0 ANQP elements (e.g., Operator Friendly Name, WAN Metrics and Connection Capability) to provide further querying functionality.

All this information carried by beacon frames and ANQP protocol is related to the hotspot's capabilities but it does not include information about the services that are locally reachable via the hotspot. Indeed, these technologies enable only mechanisms of pre-association for network discovery and selection and do not support service discovery. Thus, the network selection could not be based on available services in the Wi-Fi network.

2.2.1.2 Wi-Fi Aware (Peer-to-Peer Service Discovery)

Wi-Fi Aware [14] is a new capability launched by the Wi-Fi Alliance in 2015 which enables proximity-based service discovery before making a connection. It is an always-on technology that helps to find nearby information and services without a connection to a wireless AP. Then, it initiates interactions between devices and people. Wi-Fi Aware is a key enabler of an interactive and personalized mobile experience, enabling users to find near video gaming players, share media content, and get contextual notifications and offers according to their preferences.

This technology is based on the Wi-Fi Alliance Neighbor Awareness Networking (NAN) Technical Specification [15]. It enables a continuous device-to-device discovery using NAN Discovery Beacon, a modified version of the IEEE 802.11 beacon management frame. When a device discovers an interesting service, the device then initiates a Wi-Fi connection.

Wi-Fi Aware is a promising "neighbor awareness" technology. However, it is based on a peer-to-peer connectivity and does not provide discovery and access to local services offered through public Wi-Fi networks.

2.2.2 Service Discovery Approaches

Service Discovery (SD) is a process enabling dynamic discovery of available services in the network and automatic configuration of devices. It provides necessary information about available services and help users and applications to access to network resources such as devices, data and

services. Many Service Discovery Protocols (SDPs) were developed in research and industrial communities. These protocols are designed to minimize human intervention and administrative overhead and improve user experience. We classify these protocols into two major categories. The first category regroups protocols used after the device is associated and connected to the network, called post-association protocols. The second category is pre-association protocols that enable devices to discover services prior to associate to the network. These protocols are specifically used in wireless networks, particularly Wi-Fi.

We review, in the following, the most relevant existing solutions corresponding to each category.

2.2.2.1 Post-association Service Discovery

With the aim to enable devices to join the network and dynamically discover and access the needed services with zero-configuration, a number of service discovery protocols and architectures was developed. Examples include Jini [16], Universal Plug and Play (UPnP) [17], Service Location Protocol (SLP) [18], Salutation [19], and Bonjour [20]. These protocols are generally used in wired networks but some of them can also be used in wireless networks (e.g., UPnP and Bonjour). Each one of these protocols is designed to address a specific set of issues.

- **Jini** [16] has been developed by Sun Microsystems and is a Java based technology. It provides a software platform enabling service discovery and invocation among java enabled devices. In a Jini environment, users are able to share resources and services in a network and to easily access to available services. To ensure the advertisement and discovery of services, Jini uses a set of Discovery protocols, namely *multicast request* protocol, *multicast announcement* protocol, and *unicast discovery* protocol [21]. Jini specification is independent of the network protocol, but the most current implementations are based on TCP and UDP.

- **UPnP** [17] is a Microsoft's peer-to-peer networking initiative. It enables service advertisement and discovery in small home and corporate environments. It provides peer-to-peer connectivity between appliances, services and wired/wireless devices. It was introduced as an extension to the plug-and-play peripheral model to support discovery and configuration of devices throughout the network. UPnP uses Simple Service Discovery Protocol (SSDP) [22] for service discovery. This protocol is used for discovering devices or services and announcing the presence of a device or availability of services in the network. To do so, it uses HTTP over unicast and multicast UDP packets. UPnP is independent of operating systems, programming languages and physical media but is designed for only TCP/IP networks.

- **Salutation** [19] is another approach for service discovery. The Salutation architecture was developed by the Salutation Consortium. The aim of this architecture is to address the problem of discovery and determining the capabilities of heterogeneous information appliances

that can be encountered in a networked environment. It is independent of operating systems, physical platforms and communication protocols.

- **SLP** [18] is an IETF standard that provides a scalable and flexible framework for service discovery on IP networks. SLP can be deployed in small networks, like home networks, without any specific configuration, as well as it can scale well in large networks with predefined policies. SLP advertises services through a service URL, which contains all information necessary to connect to a service. The protocol has the advantage of not depending on any programming language or communication protocol.

- **Bonjour** is a technology developed by Apple to provide service and device discovery among computers, electronic appliances and other networked devices (e.g., printers, faxes, etc.) over IP networks. It is based on a service discovery protocol, called DNS-based Service Discovery (DNS-SD) [20].

The aforementioned protocols do not specifically address wireless issues. Indeed, they generally rely on excessive multicast and broadcast messaging which leads to channel overhead if they are used in wireless networks. In addition, these protocols are high-layer protocols and most of them are designed solely for IP-based networks. Thus, they could not be used in Wi-Fi pre-association phase when the client device has not yet acquired an IP address. Furthermore, in this phase, it is more appropriate to use "lite" service discovery mechanisms supported by the MAC layer in order to minimize the service discovery overhead.

To address the problem of service discovery and advertisement in wireless networks, [23] proposes an architecture composed of three major components: the wireless client device, the wireless AP and a service discovery server. In this architecture, the service discovery server may use UPnP, SLP, Jini or other SDPs to deliver information about available services to the wireless AP. Then, the latter transmits the received service information to the wireless client device in one or more reserved Information Element (IE) within an IEEE 802.11 management frame (e.g., beacon frame). The extensible markup language (XML) is used to describe available services within the information field. This method enables to discover and advertise information related to services in a wireless network. Nevertheless, including many information details in broadcast frames (i.e., beacon frames) along with the use of XML language can lead to channel overhead. Moreover, it is recommended to keep the size of beacon frames as low as possible to improve the wireless channel quality and reduce the usage of transmission bandwidth.

2.2.2.2 Pre-association Service Discovery

In the following, we identify the most relevant industry initiatives as well as standardization efforts related to the pre-association service discovery in wireless networks.

a) Industry initiatives

- ***Cisco's Mobility Services Advertisement Protocol***

With the aim of facilitating service discovery and enabling the connection of users to venue-based services in a WLAN, Cisco has developed the Mobility Services Advertisement Protocol (MSAP) [24]. This protocol defines a pre-association mechanism for service advertisement.

This protocol is transported by the IEEE 802.11u GAS Public Action frames, which provide transport mechanisms for advertisement services in the pre-association state. Thereby, mobile devices can query for local services prior to authenticating to a Wi-Fi network. In addition, a lot of high-layer security methods are employed in the service advertisement validation process to prevent potential attacks.

The MSAP message exchange and the high-layer security methods may potentially increase Wi-Fi connection time and network overhead and require further treatment in the mobile device. Furthermore, MSAP is a proprietary solution and an MSAP client should be integrated with mobile devices to support this protocol. This is a major drawback that impedes scalability and interoperability.

- ***Intel solution***

In this invention [25], authors define a new methodology for network and service discovery during pre-association phase to WLAN hotspot. The objective is to enable clients to receive the advertised capabilities and services offered at any hotspot and to make use of this information to select the appropriate network.

This method consists in transmitting a probe request from a wireless client to an AP to query for available services. This probe request is based on an advertising protocol which the AP supports. This protocol is determined by the client after monitoring AP capabilities in the beacon frame. Using the same advertising protocol, the client receives a probe response, from the network access point, including the required information. Although a single client may have transmitted the probe request, the response is made available to multiple clients using broadcast and multicast messages. Thus, these clients can determine network capabilities without having to transmit a probe query.

Inventors do not specify how information about available services is described through advertisement messages. They only mention that these messages are based on IEEE 802.11u. However, this standard does not include any specification about service discovery either. Moreover, using multicast/broadcast advertisement messages can burden the transmission bandwidth.

- ***Microsoft solution***

This invention [26] enables discovery of services between devices prior to establishing a connection. This can be between wireless devices or between wireless devices and devices connected to a wireless AP via wired connection. The services provided by local devices may be discovered by sending and receiving radio messages that include an IE. The latter contains information including the type of service provided by the device, the type of service discovery protocol to be used by higher layer, and other information such as security information and user-friendly name. One or more portions of the IE may be compressed prior to the transmission.

Inventors propose to use Universally Unique Identifier (UUID) [27] to uniquely identify a particular instance of a service. It is a fixed size identifier of 128-bits defined as an unsigned integer. This identifier can neither be shortened nor extended. Moreover, it can be used only to identify the service and cannot not include further service description. To do so, additional portions must be added to the transmitted IE.

The invention also enables the use of IP-based service discovery protocols (e.g., UPnP, SSDP, SLP and Rendezvous) prior to establishing a connection with Layer 2 messages (e.g., IEEE 802.11 beacon messages) by compressing service information. Compression enables to reduce the amount of transmitted information but this generates the burden of decompressing this information from the device when receiving the IE.

This solution does not specifically target 802.11 enabled devices but also invokes other wireless technologies such as Bluetooth and UWB. Moreover, it principally focuses on local hardware services, such as printer, camera and video game device, and does not deal with software services that can be offered by the wireless network provider, such as indoor mapping and discounts.

b) Standardization Efforts: 802.11aq

Currently, there is an underway IEEE project, called 802.11aq [28], that aims to develop an amendment to enable pre-association discovery of available services in the network by IEEE 802.11 stations. This standard is scheduled to be accomplished by 2017. The objective of this project is to provide mechanisms that permit to advertise the existence of services in IEEE 802.11 wireless networks and to deliver information describing them, leveraging existing schemes such as high-layer service discovery protocols and pre-association protocols (e.g., ANQP).

2.2.3 Summary

Table 4.2 presents a comprehensive survey of the aforementioned service discovery solutions.

It is argued that a pre-association method for service discovery notably improves user experience when he desires to access a specific service in public Wi-Fi networks. Two different mechanisms are described in the literature.

- The first consists in using a service advertisement protocol which requires further message exchange between the AP and the client. This increases the connection time, requires more bandwidth and consumes more power in the client device.

- With regard to the second mechanism, it consists in including one or more IE in the IEEE management frames to provide service information. This may avoid transmitting advertisement queries. However, the existing solutions are based on heavy techniques (e.g., XML, compression) [23, 26] to enable description and transmission of service information within management frames. Moreover, these frames (e.g., beacon frames) are periodically broadcasted in the network with short intervals, and their size, therefore, has to be optimized to improve network efficiency.

Consequently, a new solution is required to offer service discovery and access to public Wi-Fi services in a seamless way and with a lightweight mechanism. Thus, better user experience and more personalized access to Wi-Fi networks and services can be enabled.

Solution	Scope	Network type	Communication Mode	Discovery Time	Discovery and Advertisement Method	Open/Proprietary
802.11u and Hotspot 2.0 [3,4,13]	Network discovery and selection	Wi-Fi network	Infrastructure mode	Pre-association	ANQP/ GAS	Open standard (IEEE and Wi-Fi Alliance)
Wi-Fi Aware [14]	Service discovery	Wi-Fi network	Peer-to-Peer mode	Pre-association	NAN	Open standard (Wi-Fi Alliance)
Jini [16]	Service discovery	Wired network	Peer-to-Peer mode	Post-association	Jini Discovery protocols	Changed from proprietary Sun product to open source Apache project in 2006
UPnP [17]	Service discovery	Wired/ wireless network	Peer-to-Peer mode	Post-association	SSDP	Open standard (UPnP Forum taken over by the Open Connectivity Foundation since early 2016)
Salutation [19]	Service discovery	Wired network	Peer-to-Peer mode	Post-association	The Salutation Manager Protocol	Open standard (The Salutation Consortium)
SLP [18]	Service discovery	Wired network	Peer-to-Peer mode	Post-association	SLP protocol	Open standard (IETF)
Bonjour [20]	Service discovery	Wireless network	Peer-to-Peer mode	Post-association	DNS-SD	Proprietary protocol (Apple)
[23]	Service discovery	Wireless network	Infrastructure mode	Post-association	High layer SDP between the AP and the SD server + one or more IE transmitted in radio messages including service information	Proprietary solution (Intel)
[24]	Service discovery	Wi-Fi network	Infrastructure mode	Pre-association	MSAP	Proprietary solution (Cisco)
[25]	Service discovery	Wireless network	Infrastructure mode	Pre-association	Advertisement protocol (not specified)	Proprietary solution (Intel)
[26]	(Hard-ware) Service discovery	Wireless network	Peer-to-Peer/ Infrastructure mode	Pre-association	An IE transmitted in radio messages including compressed service information that can be related to IP-based SDPs	Proprietary solution (Microsoft)
IEEE 802.11aq [28]	Service discovery	Wireless network	Infrastructure mode	Pre-association	Not yet specified	Open standard (IEEE)

Table 2.1: Comparison of the reviewed solutions

2.3 Personalized and Seamless Access to Wi-Fi Services

In this section, we present our proposition to enable a personalized, transparent and automatic Wi-Fi access to venue-based services. For this purpose, we handle two principal points. The first one is the discovery of local services prior to Wi-Fi association. The second point is related to the unique and global identification of the service. We describe each one in detail below.

2.3.1 Pre-association Discovery of Local Services

Allowing Wi-Fi client devices to learn more about a network and particularly its services before deciding to join it is a crucial requirement to enhance network selection processes. In several cases, the presence of a specific service in the network could be a decisive criterion for association to this network. So, with the proliferation of venue-based services, it is required to discover the available services prior to association. In order to support this function and thereby enable network selection based on available services, we propose to extend the IEEE 802.11u standard, which only enables network discovery and selection in pre-association mode. To do so, we propose lightweight extensions for service discovery to the 802.11 management frames, particularly the beacon and the probe response frames.

The beacon frame, as well as the probe response frame, carries information about the AP's capabilities in a component of the frame called an Information Element (IE). In our work, we propose to add a new IE, called Venue Based Services (VBSs) element, which contains information identifying the venue-based services accessible via the AP transmitting this element. This proposal represents a fast and battery-efficient method for venue service discovery since it may avoid performing additional service discovery queries which may take a long time and consume too much battery energy.

The proposed element provides a list of Venue-Based Services Identifiers (VBS IDs). This list is generally expected to be short since it is anticipated that many hotspots will offer only a few venue specific services. Thus, there is no risk for beacon overhead.

The element's format is shown in Figure 2.1. Note that the format of this element has been defined as a proposal for standardization, but it can also be defined as a Vendor Specific element which is used to carry information not defined in the IEEE 802.11 standard. In the latter case, an Organization Identifier field should be added to identify the entity which has defined the content of this Vendor Specific element.

The element ID should be assigned by the IEEE 802.11 standard. The length field specifies the number of octets in the following information fields in this element, i.e., its value is equal to the total length of VBS ID Unit fields. When this element is present, it contains at least one VBS ID Unit field. The VBS ID Unit field has the structure shown in Figure 2.2.

The Length of VBS ID field value is set to the length of the VBS ID field. The VBS ID field is

	Element ID	Length	VBS ID Unit #1	…	VBS ID Unit #N (optional)
Octets	1	1	variable	…	variable

Figure 2.1: Venue-Based Services element format

	Length of VBS ID	VBS ID
Octets	1	variable

Figure 2.2: VBS ID Unit field format

an UTF-8 formatted field containing information, used to identify venue-based services available in the network (e.g., VBS URN service labels defined below). UTF-8 format is defined in IETF RFC 3629 [29].

2.3.2 Unique and Global Service Identifiers

Increasingly, local value-added services are deployed in airports, shopping malls, and other public spaces via the Wi-Fi network to improve customer experience. Although these services depend on the venue type and the specific user needs in this venue, there is a set of well-known and common services often required by customers in such indoor area (e.g. indoor mapping).

In this section, we propose generic identifiers that allow us to uniquely identify such global and well-known venue-based services, while actual service deployment may depend on the venue type. We define new Uniform Resource Name (URN) service labels to identify these venue-based services using the "service" URN namespace defined in RFC 5031 [30]. This URN namespace, as defined in [30], allows well-known context-dependent services that can be resolved in a distributed manner to be identified. This approach for creating service identifiers provides globally unique, persistent and extensible identifiers. However, RFC 5031 defines only emergency and counseling services whereas it is useful to define such identifiers for other well-known services such as indoor venue-based services.

Thus, we propose to extend this RFC, which defines the first two well-known context-dependent services, 'sos' and 'counseling', by adding a third top-level service label 'vbs' (referring to venue-based services).

- **urn:service:vbs** The generic 'vbs' service type encompasses all of the services offered by the venue.

We also define additional sub-services corresponding to the most well-known indoor services that

are of general public interest.

- **urn:service:vbs:mapping** The 'mapping' service refers to indoor localization and way finding using the venue map.

- **urn:service:vbs:discount** The 'discount' service refers to discounts and special deals proposed by the venue (e.g., discount offered by a restaurant , promotions offered by a shopping mall, special price reduction offered by an hotel, etc.).

- **urn:service:vbs:printer** The 'printer' service refers to printing services that can be offered by the venue such as an hotel, a library, etc.

- **urn:service:vbs:info** The 'info' service gives information related or about the visited venue. For example, in a shopping mall, it gives the list of available shops, brands, restaurants in this mall. Or in the airport, travelers could use this service to access information about their flights.

- **urn:service:vbs:video** The 'video' service refers to video streaming or download service offered by certain venues. For example, in a stadium, it gives exclusive in-venue content such as replays and live video streaming.

These URN venue service-identifying labels have many advantages. Indeed, availability of such venue service identifiers allows network entities (e.g., AP) to convey information about the available services to user devices while ensuring consistency and compatibility between devices and service providers. Thus, it allows a user device to recognize the desired services among the received information according to the defined user preferences. In fact, URN identifiers are globally unique and well-known. This allows for more automatism and transparence relative to end-users.

In addition, these URN labels identify services independently of the particular protocol using these identifiers. In our case, we propose to use it in IEEE 802.11 management frames. Meanwhile, it may appear in protocols that allow general Uniform Resource Identifiers (URIs) such as Session Initiation Protocol (SIP), web pages and mapping protocols (e.g., Location-to-Service Translation (LoST) protocol [29] that can be used to map service identifiers to service Uniform Resource Locators (URLs)).

Finally, as URN identifiers are extensible, they may contain a hierarchy of sub-services that further describes the service (e.g., **urn:service:vbs:info:flight** to inform travelers that they can get information about their flights through the airport Wi-Fi network). Consequently, the defined scheme can be used both to identify and describe services in a standard and lightweight manner.

2.4 Analysis

In this section, we analyze and compare our proposed solution with the two existing solutions, mainly the post-association service discovery and the pre-association discovery using an advertisement protocol, from the client perspective and from the network operator perspective (see Figure 2.3).

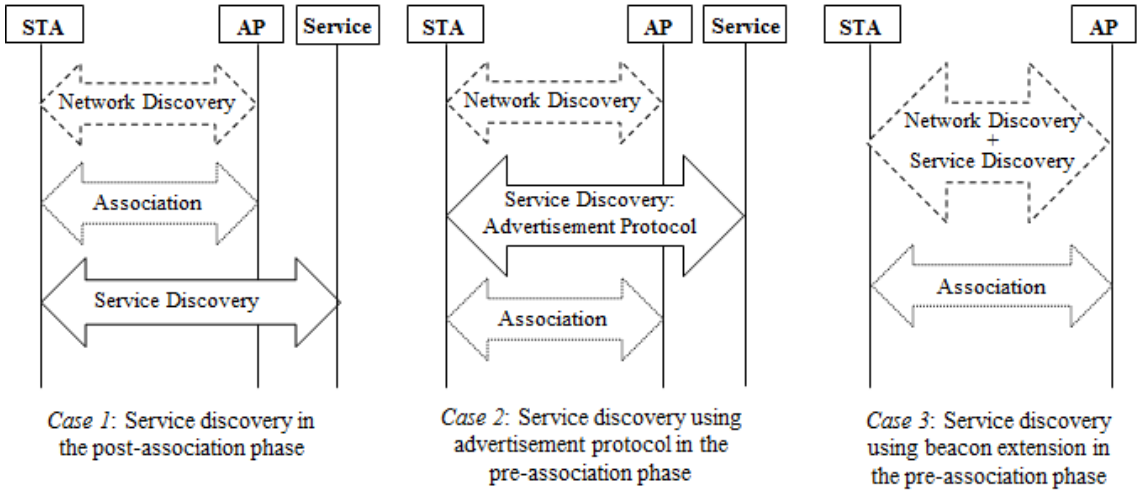


Figure 2.3: Comparison between the proposed solution (case 3) with the other existing approaches (case 1 and 2)

2.4.1 The Client Perspective

From the client's perspective, we consider three aspects: 1) link setup time, 2) power consumption, and 3) user experience and satisfaction.

2.4.1.1 Link Setup Time

The link setup time is the amount of time required to associate to a Wi-Fi network and gain the ability to send IP traffic with a valid IP address through the AP.

- In case 1, no additional time is required for service discovery in the pre-association phase because it is done after association.
- In case 2, an additional time is required for service query using an advertisement protocol.
- In case 3, no remarkable additional time is required for service discovery. This information is encapsulated in the beacon frame.

2.4.1.2 Power Consumption

As clients are usually using mobile devices operating on batteries, power consumption is a crucial factor to be considered and it is important to understand how each solution could affect and differ in power consumption on mobile devices.

- In case 1, a significant amount of power could be consumed in case of multiple Wi-Fi network association and service discovery query exchange. In fact, each time the mobile device associates to a Wi-Fi network and the client does not find the desired service, he will try to associate to another network.

- In case 2, extra power is consumed for service advertisement message exchange since energy is consumed for each transmission and reception of data.

- In case 3, no need to exchange further messages than beacon and/or probe frames but some power could be consumed for information treatment and automatic selection of the appropriate network which offers the desired service.

2.4.1.3 User Experience and Satisfaction

With the rise of using Wi-Fi enabled devices at public hotspots, delivering a better user experience is becoming a requirement for operators and in the Wi-Fi industry. There is an increased need to provide a set of features that improve and enhance user experience and deliver to users the services where and when they desire in a seamless way. Thus, it is important to analyze the implications of the three solutions on the user experience.

- In case 1, user should associate to a Wi-Fi network to be able to discover the available services. Consequently, user may associate to a network that does not offer the required service and he will be obliged to try with other networks. This can lead to user dissatisfaction.

- In case 2, if the beacon frame indicates the support for an advertisement protocol, the mobile device can use this protocol to retrieve service advertisements. When receiving the response from the server, the mobile device displays the icon(s) of service(s) provided by a Wi-Fi network on its user interface. The user then selects a displayed icon. In response, the mobile device associates to the Wi-Fi network identified in the service advertisement. The mobile device may be also configured with a policy to decide whether to retrieve service advertisements as well as to decide whether or not to associate to any given Wi-Fi network. The problem with this solution is that the mobile device should support this advertisement protocol when this technology is generally proprietary and it is supported by a limited number of mobile devices. Thus, this technology is not accessible to all users.

- In case 3, based on pre-defined user preferences and the information received in the beacon frame related to the user location (Venue Info element) and the services provided by the Wi-Fi network, the mobile device could be automatically connected to the appropriate network offering the desired service in a particular context. Thus, it allows a user to access the service without any intervention. In the case of more than one network match the user defined options, other parameters conveyed by the beacon frame (e.g., RSSI, BSS load) could be used to select the best one in terms of quality of service.

2.4.2 The Network Operator Perspective

From the network operator perspective, we consider bandwidth usage and ease of deployment aspects which are the main concerns for operators.

2.4.2.1 Bandwidth Usage

- In case 1, extra bandwidth is consumed in case of multiple Wi-Fi associations and service discovery query exchange in order to find the required service.

- In case 2, the service advertisement protocol is generally transported by IEEE Public Action frames (e.g. GAS) and this leads to extra bandwidth use.

- In case 3, an information element is added to the beacon frame but there is no risk for overhead because the list of advertised services will be short since there is a few number of venue services.

2.4.2.2 Ease of Deployment

- In case 1, deployment aspects depend on the used service discovery technology. For example, Jini technology requires a lookup server that enables service providers to advertise their services and enables clients to discover and interact with those services.

- In case 2, a server is required to provide service advertisements in the Wi-Fi network and this technology cannot work without a software client to support the advertisement protocol.

- In case 3, the solution can be deployed in the existing Wi-Fi network infrastructures without additional hardware cost. The only upgrade is to enable the support of the new beacon information element by APs. In conjunction with a software client, this could ensure automatic network selection based on a set of criteria (e.g., user preferences, user context, available services in the network, etc.).

2.5 Use Case Examples

As Wi-Fi users begin to look for venue-based services and to have a personalized and seamless user experience, it is expected that our proposal will be practical and useful in many usage scenarios. As illustrative examples, consider the two following use cases:

- *Use case 1*

A user entering into a shopping mall is interested in different promotions offered by shops in that mall. He wants to receive all the discount offers in his smartphone. His preferences are stored in his mobile device. The latter is configured with a policy to decide whether to retrieve a particular service as well as to decide whether or not to associate to any given Wi-Fi network without user intervention.

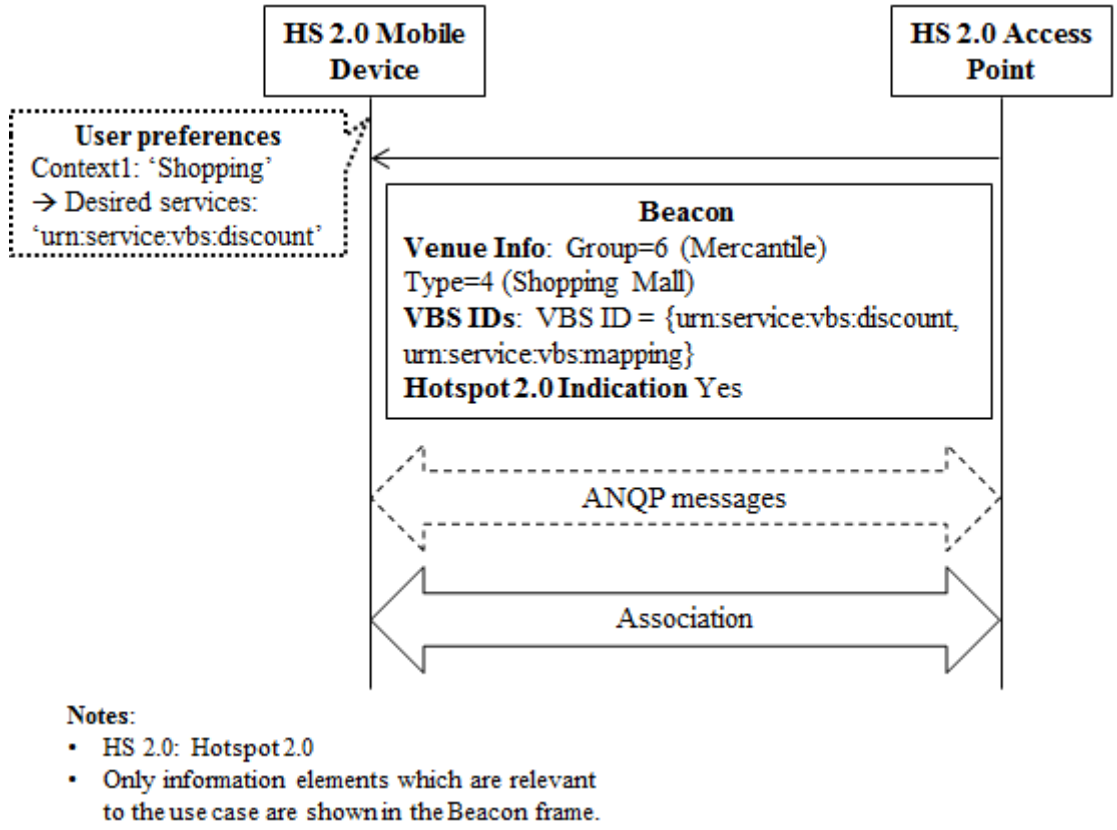


Figure 2.4: Wi-Fi pre-association message exchange (Use case 1)

When his mobile device is within the radio range of the shopping mall's hotspot, it receives the hotspot's beacon frames (Figure 2.4). In a Next Generation Hotspot environment supporting Hotspot 2.0 and 802.11u, it recognizes that the user is in a shopping mall (venue type=shopping

mall). Once it knows that the context of the user is "shopping", it will search for services that pertain to his shopping preferences. With our proposal, the mobile device will not have to issue further requests to discover the available venue services and retrieve the desired one. This information is carried by the beacon frame. Then, the mobile device detects that the "discount" service is advertised in the Venue-based Services element sent within the beacon frame.

As this Wi-Fi network matches to the user preferences, the mobile device associates to this network, and thanks to Hotspot 2.0 technologies, it can be authenticated automatically. After successful association, the mobile device launches an application (for example a browser to the URL) permitting the user to access the desired service and view the list of available discounts in his location and vicinity.

- *Use case 2*

In a retail store, a client wants to locate the item that he is looking for. He runs his VBS application and activates the "mapping" service. Then, his mobile device scans the beacon frames received from the reachable Wi-Fi networks and seeks if someone offers the "mapping" service.

The Wi-Fi network having "Store clients" as an SSID, supports this service. The mobile device retrieves this information by matching the "mapping" urn service label (i.e., urn:service:vbs:mapping) with the list of available venue services in the VBS information element carried by the beacon frame.

By using in addition Hotspot 2.0 technology, the mobile device could automatically associates and authenticates to this network. Thus, the user is connected to the Wi-Fi network offering the desired service and can access his service. He enters the name of the desired item in his user interface. Then, the application shows to him his current position in the store plan and presents a route to the item he requested.

In so doing, users are easily, immediately and autonomously connected to the desired venue-based service without any user intervention to select the network.

2.6 Conclusion

Providing more personalization and transparency for access to Wi-Fi networks and services is becoming a crucial requirement to give customers the user experience they desire and expect, especially, in parallel to the emergence of venue-based services. For this purpose, we proposed, in this chapter, a lightweight and fast mechanism to enable Wi-Fi client devices to discover local venue services in a network before deciding to join it. In addition, we defined a simple scheme to uniquely and globally identify venue services. This scheme is extensible and enable to further describe services. Thanks to these two contributions, users are able to have a personalized and seamless access to venue-based services through Wi-Fi networks.

Chapter 3

Next-Generation Carrier Wi-Fi Architecture

Contents

3.1 Introduction

Over the last decade, WLAN networks have experienced an incredible evolution with the emergence of new technologies and services which mainly enhance user experience and improve quality of service. Nevertheless, today's WLAN networks are unable to rapidly adapt to such evolution due to their rigid architectural design. In fact, this typically requires time-consuming and costly upgrades of existing infrastructures generally composed of proprietary hardware appliances. Consequently, the issue of implementing a flexible architecture while boosting innovation and reducing

the cost of network upgrades becomes a major concern to support evolving contexts and service needs.

In this chapter, we present a novel architecture for carrier-managed WLAN networks which leverages Network Function Virtualisation (NFV) and Edge Computing concepts. It is based on a WLAN Cloudlet, located in the end user premises, that offloads MAC layer processing from access points and consolidates network functions and value-added services. All these functions and services are based on software instances. This brings more flexibility and adaptability to the whole system. Thus, it allows for reducing the complexity and cost of introducing new functionalities and services in the different levels (L2-L7). Moreover, the WLAN Cloudlet introduces proximate virtualization infrastructure in the WLAN architecture which provides two major benefits. First, this plays a crucial role in reducing i) AP costs by offloading MAC layer processing to virtual machines provided by the Cloudlet and ii) other network equipment costs through consolidating multiple instances of network functions in the same hardware. Secondly, it decreases access latency by placing network functions and certain services close to end-users.

The remainder of this chapter is structured as follows. Section 3.2 discusses existing WLAN architectures and emerging concepts for next generation wireless networks. In section 3.3, we describe our proposed NFV- and Cloudlet-based WLAN architecture and we present the benefits, possible applications as well as potential challenges of this architecture. Section 3.4 describes feasibility and implementation aspects and provides performance evaluation results. Finally, we conclude this chapter in section 3.5.

3.2 Background and State of the Art

In this section, we first discuss the evolution of WLAN architectures while highlighting the major advantages and drawbacks of each one. Second, we describe the emerging concepts that will have a major impact on the future wireless networks.

3.2.1 Evolution of WLAN Architectures

Hereafter, we describe the evolution and the different types of WLAN architectures, namely the autonomous, the centralized, the distributed, and the virtualized architecture. Then, we summarize the major drawbacks of these architectures.

3.2.1.1 Autonomous WLAN Architecture

In the first generation of WLAN, APs were individually managed and independent. Moreover, they contained all the intelligence to manage Wi-Fi traffic [31, 32]. As a result, duplicating this intelligence induces high cost especially for medium and large sized networks. In addition, this solution has limited capacities and is relatively static.

3.2.1.2 Centralized WLAN Architecture

As the need for centralized monitoring and dynamic configurability grew, vendors introduced controller-based systems with "thin" low-cost APs. In this architecture, the controller is responsible for controlling, configuring, and managing the entire WLAN access network. Furthermore, all the traffic is routed from the APs to the controller [33]. According to RFC 4118 [31], centralized WLAN architectures are categorized into three main variants: i) the *Local MAC* in which the MAC functions stay intact and local to APs, ii) the *Remote MAC* in which the MAC has moved away from the AP to a remote Access Controller (AC) in the network, and iii) the *Split MAC* in which the MAC is split between the APs and the ACs. The centralized WLAN architecture is characterized by the ease of deployment especially for wide networks and it provides more security and control. However, it has two major drawbacks. Firstly, the WLAN controller represents a single point of failure. Secondly, since all transmissions require passing through the controller, the latter became a major bottleneck, thus eroding network performance.

3.2.1.3 Distributed WLAN Architecture

To resolve the issues mentioned above, the third generation of WLAN, called distributed WLAN architecture, introduced distributed data forwarding [34]. A controller still provides a central point of control for APs, however, all traffic is no longer backhauled to the controller. This solution eliminates network bottleneck but it is much more expensive than the other solutions.

3.2.1.4 Virtualized WLAN Architecture

- **Virtualization of the WLAN Controller**

Recently, a new trend has emerged in the world of WLAN by introducing virtualized WLAN architecture which has rapidly been gaining the attention of industry and academia. Some vendors offer the controller as a virtual appliance and even as a cloud-based hosted service [35–37]. The cloud-based controller centrally manages and monitors APs and user data is not going through the controller. This solution has several advantages in terms of ease of deployment and availability but it exhibits a number of shortcomings. Note that even if this solution eliminates the need of deploying on-premises hardware WLAN controllers, this does not imply cost reduction. In fact, controller functionality is distributed between the APs and the cloud. Thus, the cost of the controller is integrated into the price of the APs and the Cloud controller subscription that must be continuously renewed. Over the long haul, this solution is much more expensive than simply purchasing a WLAN controller from the beginning. Moreover, for security reasons, many organizations still require on-premises WLAN controller to easily manage network traffic and there may be regulatory needs to tunnel traffic to the controller. Finally, this solution is only suitable for low- to medium-density

locations as APs are not designed to support high-scale tunnel traffic. In contrast, centralized controllers have a dedicated hardware that makes them extremely efficient at moving traffic through the network.

- **Virtualization of the WLAN Access Point**

It is worth pointing out that virtualization did not only target the WLAN controller but also access points. CloudMAC [38] attempts to partially offload the MAC layer processing to virtual machines in the Cloud acting as virtual APs. This allows for reducing the software complexity of physical APs. Furthermore, adding support for new functionalities related to MAC processing requires updates only to the virtual APs. The deficiency of this solution is the additional latencies introduced due to the distant MAC frames processing.

Another approach of virtualization in WLAN networks is the virtualization of APs which happens inside the physical AP. It is based on the virtualization of a single radio hardware and enables the creation of multiple virtual interfaces to operate as virtual APs [39–42]. This enables Wi-Fi providers to offer multiple services and define different policies on the same physical AP. This also supports a multi-provider environment where each provider can offer separate services to his subscribers through a virtual AP. In [43, 44], authors conceive Virtual APs as logical entities which could move to another physical AP using live migration. This permits better deployment of WLAN APs and supports user mobility by keeping him connected to the same virtual AP when he moves.

These solutions lead to more complex APs and require powerful APs, thus generating high costs for WLAN deployment. Moreover, the configuration and management of such APs are more complicated.

3.2.1.5 Summary

In addition to the cost and QoS shortcomings, the aforementioned WLAN architectures does not address flexibility and evolvability issues. Indeed, any new service implementation typically requires an upgrade of each AP, which is costly and time consuming particularly in large scale networks.

To summarize, an optimal and efficient solution in terms of cost, QoS (e.g., latency), flexibility, and simple deployment of new services is required.

3.2.2 Emerging Concepts for Future Wireless Networks

As emerging concepts for future wireless networks, we describe below Network Function Virtualisation (NFV) and new Cloud Computing models.

3.2.2.1 Network Function Virtualization

NFV [45] is a new industry trend that has gained a lot of attention over the past few years. NFV leverages standard IT virtualization technology to consolidate many network functions, which tra-

ditionally reside in purpose-built equipments, onto industry standard server hardware which could be located in data centers, network nodes and in the end-user premises.

Thanks to its basic idea of decoupling software from hardware, it promises cost efficient realization of network functions in software deployed over commodity hardware, which can be instantiated in, or moved to, various locations in the network as required. Moreover, it encourages openness and innovation to quickly bring new services and new revenue streams at a much lower risk.

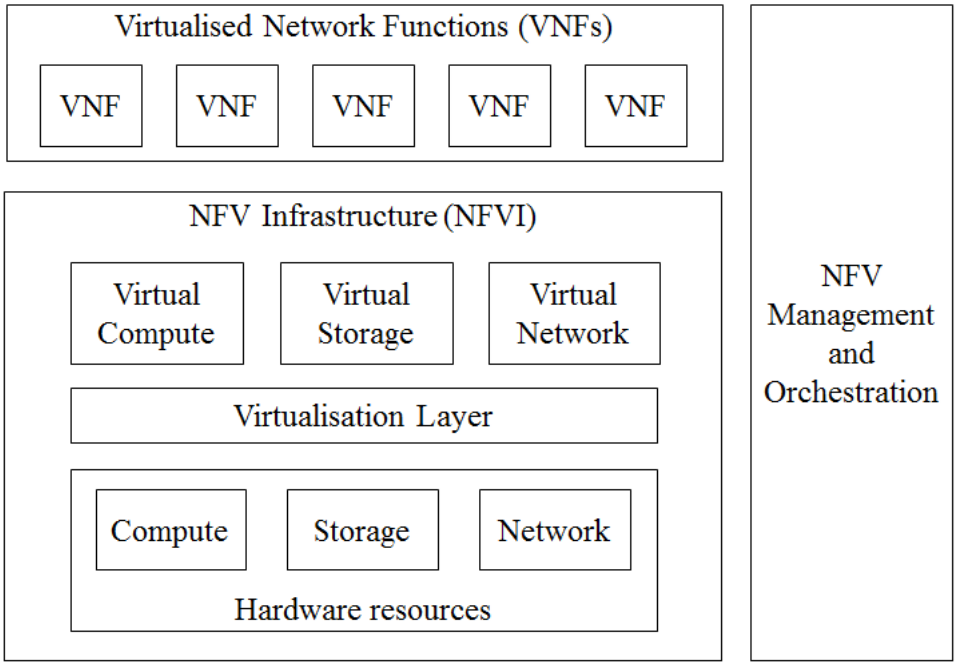


Figure 3.1: High level NFV framework

Figure 3.1 illustrates the high-level NFV framework that is composed of three main working domains:

- *NFV Infrastructure (NFVI)*, including all the physical resources (i.e., compute, storage and network resources) and how these can be virtualized to support the execution of multiple VNFs and provide a multi-tenant infrastructure.
- *Virtualized Network Function (VNF)*, the software implementation of a network function which is capable of running over the NFVI. VNFs may be dynamically deployed on the NFVI on demand within the capacity limits of the NFVI nodes.
- *NFV Management and Orchestration (MANO)*, which covers the orchestration and lifecycle management of network functions, and the management of physical resources. This includes

end-to-end network service mapping, VNF instantiation at appropriate location, hardware resource allocation and scaling, performance measurements, etc.

An ETSI Industry Specification Group (ISG) for NFV [46] was created in November 2012 by seven of the world's leading telecoms network operators. It now consists of over 270 companies which are working to develop the required standards for NFV and to set the direction for NFV implementation and deployment in an open and interoperable ecosystem.

There are many other standardization efforts related to NFV such as IETF NFV Research Group (NFVRG) [47] and ATIS NFV Forum [48]. In addition, an open source project, called Open Network Function Virtualization (OPNFV) [49], was formed to accelerate the introduction of new NFV products and services through a carrier-grade, integrated and open platform. OPNFV is considered as a complementary community to existing standards and open source bodies with a clear focus on the coordination of software development, integration and testing, documentation and API development for NFV. Particularly, OPNFV is relying on ETSI NFV ISG specifications and collaborating with open source communities (e.g., OpenStack, OpenDaylight) to achieve an industry wide NFV reference platform.

3.2.2.2 Emerging Cloud Computing Models

Recently, different terminologies have been used by the research community to describe emerging Cloud Computing models leveraged to fulfill the requirements of emerging computing and networking scenarios. This includes fog computing, edge computing and cloudlets. Although different terms are embraced, these models share similar characteristics. Indeed, they are based all on the concept of building computing and networking infrastructure that provides high-capacity and responsive virtualized resources close to end-users. In the following, we define these emerging paradigms and we highlight the major differences between them.

- ***Fog/Edge Computing***

The term "Fog computing" was first introduced by Cisco Systems [50] as a new paradigm that extends Cloud Computing to the edge of the network. In many ways, it is synonymous with "Edge computing". However, Fog computing particularly targets Internet of Things (IoT) services and applications (e.g., Connected Vehicle, Smart Cities, etc.) with location awareness, low latency and mobility support. It is based on widely distributed fog nodes that can be deployed anywhere and can include any device with computing, storage, and network connectivity (e.g., wireless access points, switches, routers, video surveillance cameras, etc.). This builds a highly virtualized platform between endpoint devices (i.e., Things) and the cloud.

Recently, a new industry and academic group, called OpenFog Consortium [51], was formed to promote the deployment of fog computing by developing an open architecture and core technolo-

gies to enable end-to-end IoT scenarios. The consortium was founded by ARM, Cisco, Dell, Intel, Microsoft and Princeton University in November 2015.

There are also some recent research works related to the edge computing concepts. For instance, in [52], authors introduce the term "edge as a service" defined as the concept of leveraging network virtualization to enable a flexible usage of access network resources (e.g., base stations, radio network controllers, wireless access points, etc.) among service providers. In [53], authors discuss dynamic allocation, migration and orchestration challenges of virtual network functions across edge networks. In [54], authors discuss possible directions for the future wireless network architecture based on edge cloud. To show the important role of edge cloud in this architecture, they present case studies dealing mainly with edge caching and local signal processing in support of limited-resource devices.

- ***Cloudlet***

Cloudlet was first introduced by Satyanarayanan et al. [55] under the context of Mobile Computing to enable mobile users to rapidly instantiate customized service software on a nearby cloudlet and then use that service over a WLAN. This allows the resource-constrained mobile device to function as a thin client and exploit virtual machine technology while avoiding long WAN latencies of Cloud Computing. It was proposed to integrate Cloudlet with access points [55], with wireless mesh networks [56], and with base stations [57]. Most of works related to the Cloudlet concept focused on Mobile Computing for resource-demanding and delay-sensitive applications.

3.3 NFV- and Cloudlet-based Carrier Wi-Fi Architecture

In this section, we first describe our proposed carrier Wi-Fi architecture. Second, we highlight the benefits and some applications of this architecture. Finally, we discuss the major challenges induced by the deployment of such architecture.

3.3.1 System Description

To address the shortcomings of existing architectures mentioned in the previous section, we propose a novel carrier-grade WLAN architecture that leverages network function virtualisation and cloud computing free of WAN delays, jitter, congestion, and failures. Hence, it is called NFV- and Cloudlet-based carrier Wi-Fi architecture. It allows greater flexibility, ease of management, and rapid creation and deployment of new services including application-level services, network functions and some specific-802.11 functions. Figure 3.2 depicts this architecture composed of a WLAN Cloudlet, Wireless Termination Points (WTPs), and a centralized Cloud-based platform.

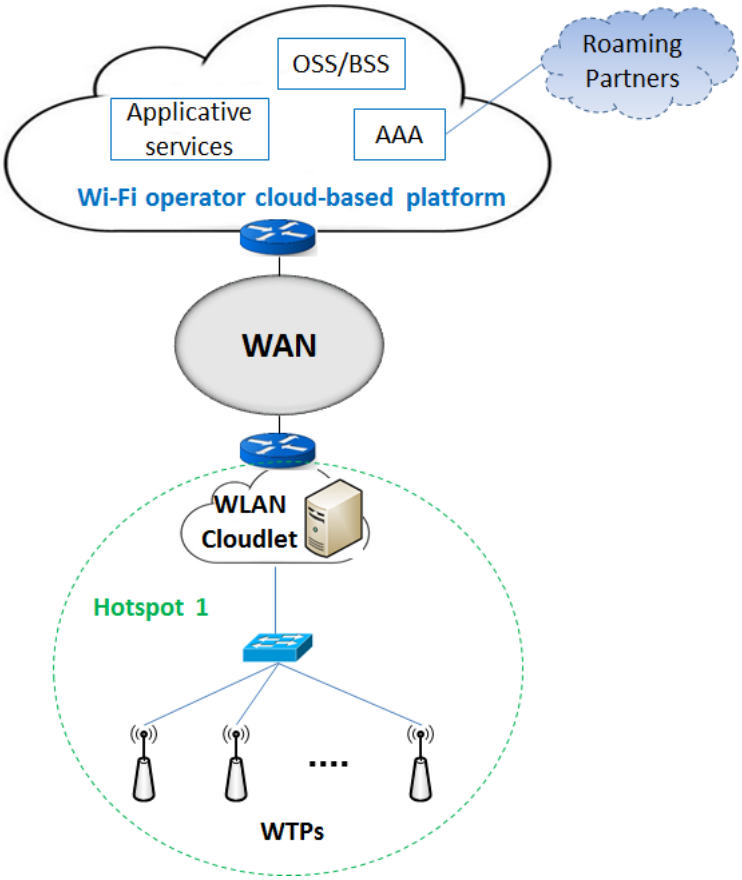


Figure 3.2: Cloudlet-based WLAN architecture

3.3.1.1 WLAN Cloudlet

WLAN Cloudlet is located in the end-user premises and consolidates some MAC-layer functions, network functions, and value-added services provisioned by software that run on an industry standard server hardware. In practice, these functions are bundled in virtual machines installed over an hypervisor. As described in Figure 3.3, WLAN Cloudlet incorporates the following types of virtual functions and services:

- ***Virtual MAC Functions***: are handled by a Software AP (SoftAP) which partially offloads WLAN MAC processing from physical access points. In fact, it runs "non-real-time" MAC functions such as *Distribution and Integration* services or responding to *Association / Authentication* MAC frames. A SoftAP can be connected to many WTPs. Consequently, a WLAN network (a hotspot) is only one SoftAP. This simplifies the management of WLAN and allows a simple update and rapid deployment of new functionalities using software modifications. Moreover, this allows the WTP to act as a thin AP by reducing software complexity.

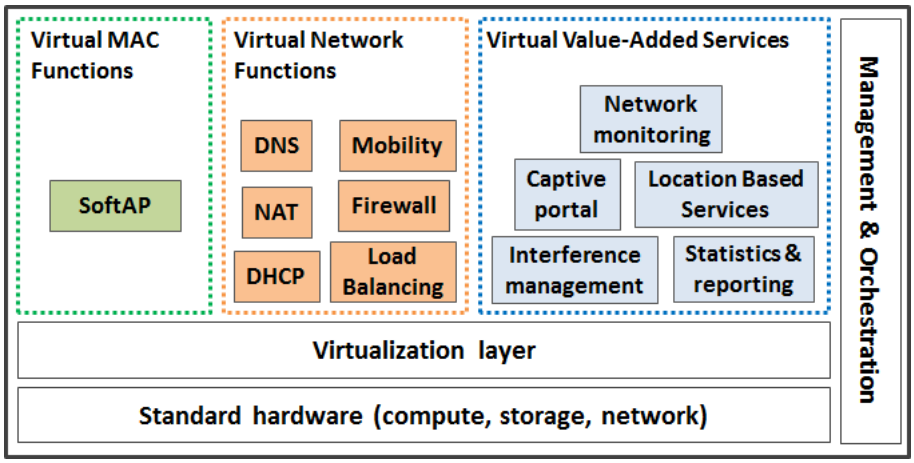


Figure 3.3: WLAN Cloudlet architecture

- ***Virtual Network Functions***: include all network functions and services that are locally deployed in a WLAN network as dedicated hardware infrastructures, such as routing, DHCP, NAT, firewall, and DNS. Moving network functions from purpose-built appliances to equivalent functionalities implemented in the Cloudlet, a Commercial Off The Shelf (COTS) hardware environment providing cloud computing capabilities, increases the flexibility to deploy new features and decreases setup and management cost.

- ***Virtual Value-Added Services***: include network and applicative services which improve QoS and user experience. Examples include Interference Management, Location-and Venue-Based Services, Statistics, etc.

3.3.1.2 Wireless Termination Points

WTPs are the physical entities that contain an RF antenna and implement 802.11 physical layer functions to transmit and receive station traffic over the air. They operate as a pass-through, forwarding MAC management frames between the WLAN clients and the SoftAP within the WLAN Cloudlet. WTPs also implement the medium access using the *Distributed Coordination Function* [58]. In addition, WTPs process MAC services with real-time constraints such as synchronization, retransmissions, generating Control frames (e.g., RTS, CTS and ACK), and beacon and probe response frames.

3.3.1.3 Cloud-Based Platform

The Cloud-based platform includes functions that need to be executed in a centralized way (e.g., OSS / BSS, AAA) or requires a lot of resources (e.g., analytics). Thus, the WLAN Cloudlet provides

interfaces to access these functions.

Furthermore, this Cloud platform can offer more scalability to our system. Indeed, processing capabilities of the WLAN Cloudlet server can be extended by cloud resources especially for non-real-time functionalities and the virtualized function can be scaled by creating additional instances of the function in the Cloud platform. This ability of WLAN Cloudlet to dynamically scale leads to support a very large number of flows.

3.3.2 Benefits and Possible Applications

3.3.2.1 Benefits

Cloudlet-based WLAN brings many benefits to the network operator and potentially to end-users. These benefits include:

- Reducing equipment costs by avoiding, especially in large scale WLANs, to have a large number of expensive fat APs to ensure network coverage since the WTPs are kept the lightest as possible.
- Reducing CapEx and OpEx by getting rid of expensive purpose-built middle-boxes (e.g., access controller used in many organizations for security and regulatory requirements, WLAN controller, firewall, etc.) and avoiding their management complexities due to NFV technology.
- Decreasing access latency by placing network functions and certain applications close to end-users.
- Bringing more flexibility and simplicity to network management. In fact, adding new MAC or network functionality requires only a software upgrade.
- Increased speed of Time to Market and faster configuration of new services.
- The possibility of introducing targeted services based on geography or venue type. Furthermore, provisioning could be made remotely in software without any site visits required to install new hardware.
- Encouraging openness and more innovation to bring new services and new revenue streams quickly at much lower risk.

3.3.2.2 Possible Application Scenarios

Cloudlet-based WLAN enables a range of applications and use cases. Hereafter, we present some examples.

- Simple upgrade to new technologies such as Hotspot 2.0 [3]. Actually, to adopt this technology, this requires a time-consuming and costly upgrade of existing infrastructure or even change of legacy equipment (i.e., APs and controllers). With our proposed solution, the deployment of this new feature will be rapid and cost-effective, since it only requires a software upgrade of MAC functions to support this standard. This upgrade can be performed centrally and there is no need to upgrade WTPs.

- Providing network gateway functions in an on-demand and customized fashion. Particularly, this enables customers to insert new network gateway functions such as security measures according to their requirements and in order to face new security threats. For example, a virtual firewall could be implemented for each group of users with specific security policies.

- Introducing new services which improve the quality of service and network throughput. As an example of these services, we mention interference management service [59, 60] which allocates the optimum frequency channel to each WTP based on information related to the radio environment gathered from WTPs.

- Simple integration of Venue-Based Services (e.g., indoor mapping, special offers) that could leverage Wi-Fi Location Based Services. This could enhance user experience and offer monetization opportunities to Wi-Fi operators.

3.3.3 Challenges

In order to realize the concept of Cloudlet-based WLAN, several challenges need to be addressed. We summarize, in the following, the main challenges.

- *Management and orchestration*. The Cloudlet-based WLAN framework should incorporate mechanisms for automated management and dynamic orchestration of virtual functions and services to enhance availability, flexibility, elasticity and to meet targeted performance constraints. As particular management mechanisms, we highlight below monitoring and optimization functionalities and VM placement.

- *Monitoring and optimization*. Adaptive monitoring and optimization approaches are needed to constantly monitor networking and computing/storage assets and optimize resources between multiple service instances relative to usage and policy constraints under dynamic conditions.

- *VM placement*. As a major optimization mechanism, a resource mapping function tailored for Cloudlet-based WLAN deployment is needed to determine where to place the VM of each service, either in the WLAN Cloudlet or in the Cloud. This function requires a prior knowledge of physical machines' capacities and VM requirements. Additionally, it may be

subject to different objectives such as optimizing system utilization and performance and reducing backbone network workload with respect to a set of criteria. This could be related to several requirements such as QoS (e.g., response time), location constraints (as some services may require to be in the proximity of end-users), and security aspects. Note that this aspect will be further studied in the next chapter.

- *Elasticity and scalability*. Automatic scaling mechanisms should be supported to satisfy service level agreements and resolve eventual bottlenecks in the WLAN Cloudlet. For example, virtual functions could be scaled by creating additional instances of the function in the Cloud platform or by migrating the VM to the Cloud.

3.4 Feasibility and Implementation

This section explores the feasibility of the Cloudlet-based WLAN architecture through a practical use case. Based on this reference case, we have built a proof-of-concept prototype and evaluated its performance.

3.4.1 Case Study

We assume the following basic scenario: a user desires to associate to a Wi-Fi network through his device to watch a video available on a video server. To describe the functional aspect of this scenario, we define a service chaining graph. To do this, we use the Virtual Network Function Forwarding Graph (VNF FG) [61], which is defined by the ETSI as one of the NFV components. This graph defines the sequence of network functions that packets go through and provides the logical connectivity between virtual appliances. A VNF FG can also include physical network functions to provide an end-to-end network service. As illustrated in Figure 3.4, in our scenario, the VNF FG is composed of the WTP, a set of virtual functions (SoftAP, DHCP, DNS, firewall and NAT), a gateway, and a video server.

3.4.2 Implementation Aspects

In our testbed for the proof-of-concept, the WTP is implemented using a machine equipped with Intel Pentium dual CPU 2.00 GHz processor with 2GB of RAM, Linux Ubuntu OS (release 14.04.2 LTS) and a TL-WN722N wireless card using Atheros 9271 chipset which uses ATH9K_HTC as a driver. This card supports the monitor mode, which allows for receiving and transmitting raw MAC frames.

The WLAN Cloudlet is a machine equipped with an Intel Core i5-3337U 1.8 GHz processor with 4GB of RAM. It is connected to the WTP through a Gigabit Ethernet switch and it uses VMware Player that runs two VMs.

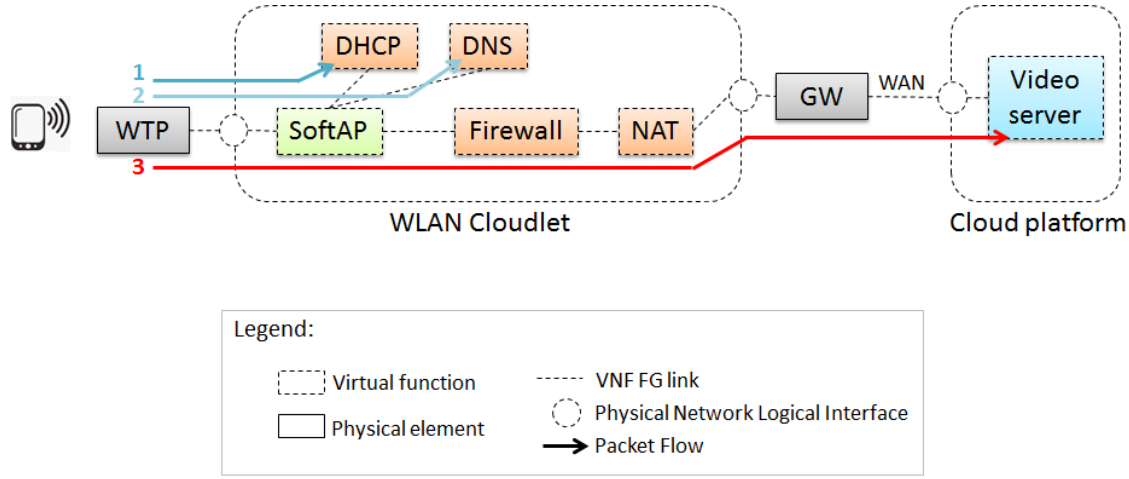


Figure 3.4: Use Case: VNF forwarding graph

The first VM is the Software AP which uses Xubuntu OS and communicates with the WTP using OpenCAPWAP protocol [62, 63], an open source version of CAPWAP (RFC 5416) [64]. This protocol supports the Split MAC approach in which only the real-time IEEE 802.11 services, such as the beacon and probe response frames, are handled on the WTP and all remaining MAC management frames and distribution and integration services reside on the SoftAP. We have chosen this approach to make the WTP as light as possible while satisfying 802.11 timing constraints. Moreover, most of the MAC services can be updated centrally.

The second VM is a FreeBSD guest machine running pfSense [65], an open source firewall which is able to run a set of services such as DHCP, DNS, routing and NAT.

3.4.3 Performance Evaluation

We report, in the following, experimental results concerning the performance of our Cloudlet-based WLAN prototype implementation. Using the testbed described above, we measured the delay and throughput metrics between different elements in our architecture. We describe below the different sets of measurement as well as the obtained results for each metric.

3.4.3.1 Delay

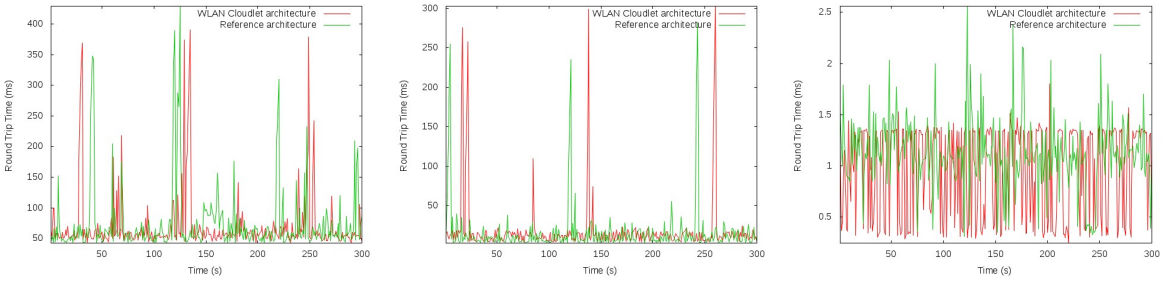
To evaluate delay performance, we measure the Round Trip Time (RTT) metric given by the ping tests. In these tests, we compare our proposal to a reference architecture which includes a standalone AP based on Hostapd [66] (using the same hardware as the WTP) and is connected to the virtual firewall. The delay measurements are performed in three different segments of the network referred to, in the following, as case ‘a’, case ‘b’, and case ‘c’.

- Case ‘a’: We consider the delay between the user device and the video server.

- Case ‘b’: We measure the RTT for both architectures between the user device and the gateway in order to have more accurate values without the impact of dynamic nature of Internet network.

- Case ‘c’: We focus on the delay between the WTP (Hostapd AP in the reference architecture) and the gateway, thus excluding the radio link delay.

Figure 3.5 shows the RTT values measured each second during two minutes for the three cases. The minimum, average, maximum and standard deviation values of RTT are presented in Table 3.1.



(a) RTT between the user device and the video server (case ‘a’)

(b) RTT between the user device and the gateway (case ‘b’)

(c) RTT between the WTP/ Hostapd AP and the gateway (case ‘c’)

Figure 3.5: RTT measurements

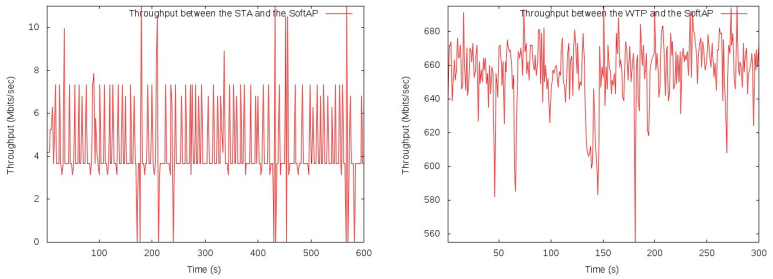
In case ‘a’ and ‘c’, the curves for both architectures are almost indistinguishable and RTT values are very close. However, in case ‘b’, the results reveal more visibly that the delay in our architecture is slightly higher than the reference architecture. This is mainly due to the presence of separate physical and software access points in our architecture, while, in the reference architecture, we have only a standalone access point. Thus, user traffic is passing through one more hop in the proposed architecture than the reference one. Moreover, this delay is also due to the virtualization layer. The results, in general, show that the WLAN Cloudlet solution achieves delay performance metrics comparable to the reference architecture.

		Min	Avg	Max	Std Dev
Case 'a'	**WLAN Cloudlet architecture**	43.310	71.069	390.764	52.655
	Reference architecture	42.672	75.609	429.529	57.865
Case 'b'	**WLAN Cloudlet architecture**	4.180	18.632	303.343	37.777
	Reference architecture	2.897	16.334	282.601	32.785
Case 'c'	**WLAN Cloudlet architecture**	0.247	0.964	1.808	0.458
	Reference architecture	0.309	1.149	2.560	0.484

Table 3.1: RTT values (ms)

3.4.3.2 Throughput

In the second set of experiments, we evaluate the throughput performance in the proposed WLAN Cloudlet architecture between the user device and the SoftAP (Figure 3.6-(a)), on the one hand, and between the WTP and the SoftAP (Figure 3.6-(b)), on the other hand. To do this, we use Iperf tool which enables to measure the maximum TCP bandwidth. We perform tests during 10 minutes when results are reported every two seconds in the first experiment and during 5 minutes when results are reported each second in the second experiment. So, we have in total 300 throughput measurement values for each experiment. The minimum, average, maximum and standard deviation of these values are depicted in Table 3.2 for both segments of network mentioned above.



(a) Throughput between user device and the SoftAP

(b) (b) Throughput between the WTP and the SoftAP

Figure 3.6: TCP Throughput measurements

	Min	Avg	Max	Std Dev
Between the user device and the SoftAP	3.15	4.45	11	1.92
Between the WTP and the SoftAP	555	656.64	695	19.72

Table 3.2: Throughput values (Mbps)

We notice that the maximum of TCP throughput value between the client and the SoftAP is 11 Mbps (this value is barely reached three times) and the average value is 4.45 Mbps. These values are expected because the used WLAN standard in the tests is the IEEE 802.11b which has as a maximum theoretical throughput 11 Mbps. We have to mention that, in our experiments, the wireless client is placed close to the WTP and these experiments are performed in an operational WLAN network where interferences with other networks cannot be avoided.

In order to eliminate the radio link effect and evaluate the maximum bandwidth of our internal system, we measure the throughput between the WTP and the SoftAP. Figure 3.6-(b) and Table 3.2 show that the maximum throughput is 695 Mbps and the average value is 656.64 Mbps. Consequently, in this current testbed, the SoftAP can handle about 130 users with a maximum of 5 Mbps for each one.

According to the aforementioned results, the WLAN Cloudlet solution does not affect network performance metrics and, at the same time, provides many advantages in terms of cost, flexibility and agility.

3.5 Conclusion

Operating large-scale WLAN deployments with heterogeneous hardware and software components, especially when technology and service innovation accelerates, is becoming very challenging for network operators. In this chapter, we presented an NFV- and Cloudlet-based WLAN architecture in which functions on the MAC, network and service layers are virtualized onto an industry standard server located in the end-user premises. Such architecture provides many benefits to network operators by reducing time and cost to integrate new services and afford flexibility and adaptability that would enhance user experience in next generation networks.

As a major challenge induced by this proposal, we will tackle, in the next chapter, the problem of optimizing the placement of virtual functions in the proposed architecture.

Chapter 4

Service Management in NFV-oriented Carrier Wi-Fi architecture

Contents

4.1 Introduction

Over the last decade, Cloud Computing has gained significant popularity as a cost-effective, scalable and flexible model for hosting services in powerful data centers which provide on-demand

computing, storage and networking resources. In particular, wireless network operators are considering the usage of Cloud Computing as a potential solution to enable more agility for the creation of new services and cope with the surging mobile data traffic [1]. Furthermore, NFV represents one of the key technology drivers of the next generation 5G networks. Indeed, by virtualizing core network functions as well as radio-access network functions, NFV can effectively reduce the cost to deploy and operate large wireless networks.

Face to this technology revolution, a new underlying cloud architecture that will be leveraged to deploy VNFs is emerging. This architecture, referred to as the edge-central cloud architecture or the two-tier architecture, is based on the combination of edge clouds (i.e., a cloudlet), close to end-users, and a core cloud, located in a centralized data center. This provides low latency and high-capacity virtualized services and resources as close as possible to end-users and ensures improved scaling and load balancing with centralized cloud resources. Moreover, it supports location- and context-aware services to improve the quality of users' experience. In such carrier-grade cloud landscape, specific requirements need to be considered for VNF management and deployment. Thereby, efficient VNF placement and provisioning strategies are needed to ensure a superior customer experience. However, deciding where to place VNFs either on the cloudlet or on the cloud is not trivial. Indeed, there are many functions (e.g., RAN functions) with strict real-time requirements and latency constraints that need to be placed on the edge cloudlet. But, the latter is constrained by capacity limits. Moreover, in order to satisfy VNF response time requirements and to improve the user-perceived QoS, several factors should be considered like the virtualization overhead, utilization level of physical hosts, and network delays.

To address these issues, we propose VNF provisioning and placement strategies to determine the required resources and placement of VNFs in the edge-central carrier cloud infrastructure taking into account QoS requirements (i.e., response time, latency constraints and real-time requirements). The main goals of our proposal are i) to optimize resource utilization, ii) to prevent cloudlet overload and congestion, and iii) to avoid violation of QoS and Service Level Agreement (SLA) requirements. For this purpose, we use queuing and QoS models along with optimization techniques to efficiently allocate resources and place VNFs.

In our proposal, we specially address carrier Wi-Fi networks based on edge-central cloud architecture as described in the previous chapter. However, our solution can be as well applied to any kind of carrier wireless network including mobile networks, since we do not make any assumption about the particular type of the wireless technology.

The remainder of this chapter is structured as follows. In section 4.2, we study the problem of Virtual Machine (VM) placement in virtualized environments and related works relevant to ours. Next, we present, in section 4.3, our proposal dealing with the placement and provisioning of VNFs in edge-central wireless network architecture based on QoS requirements. Section 4.4 presents the performance evaluation and discusses the simulation results of our proposed solutions. We finally

provide concluding remarks in section 4.5.

4.2 Virtual Machine Placement in Virtualized Environments: State of the Art

Virtualization [67] is a key technology for Cloud Computing [68, 69] and for future networks. It allows abstraction of physical resources and provides virtualized resources for high-level applications in order to improve agility and flexibility and reduce costs. Basically, virtualization has different trends such as server virtualization, storage virtualization, and network virtualization. We are particularly interested to the server virtualization that represents the mapping of single physical resources to multiple logical partitions commonly called Virtual Machines (VMs).

VMs are the key component of Cloud Computing and they provide virtual resources such as CPU, memory, storage, and network interfaces in the same way as physical resources do. In a virtualized environment, a VM instance can be dynamically created, scaled up, and migrated to another location as demand and conditions vary. Thus, virtualization improves availability and scalability and is a well suited technology especially for dynamic cloud infrastructures. However, it exposes major resource management issues.

Indeed, efficient placement of VMs on physical hosts is one of the most important resource allocation and management issues in cloud infrastructures. VM placement may have different objectives ranging from QoS, resource utilization improvement, power efficiency, to economical profit. Optimal dynamic VM placement is also required for load balancing, server consolidation and hot spot mitigation.

Furthermore, as cloud infrastructures evolve, Virtual Machine Placement (VMP) problem is becoming more and more crucial. In fact, VM placement issue is not restricted to a single cloud deployment. It also addressed in different multi-cloud scenarios that have recently emerged (see Figure 4.1). Indeed, in addition to the two basic models of cloud infrastructures, private and public clouds that can be used in isolation, there is a variety of emerging combinations where resources from different Infrastructure Providers (IPs) can be combined in novel ways. Some examples of these combinations are hybrid cloud architecture (i.e., an organization operating a private cloud is able to externalize workloads to public IPs), federated architecture (i.e., an IP can lease capacity from other providers as well as offer spare capacity to a federation of IPs), and multi-cloud architecture (i.e., service providers working directly with multiple external IPs) [70]. Besides, large companies like Amazon, Google, Microsoft, IBM, and Oracle have begun to establish new data centers for hosting cloud computing applications in various locations around the world. On the one hand, this provides redundancy and achieves reliability in case of site failures. It permits, on the other hand, to satisfy customer requests from locations close to these users and thereby helps in reducing network capacity needs, particularly for high-bandwidth applications (e.g., Cloud gaming),

and permits to meet SLA requirements (e.g., response time).

The appetite among enterprises for a range of execution environments to satisfy the requirements of different workloads emphasizes the need to a good VMP strategy with policy automation. This means making accurate choice between different locations to run VMs, depending on costs, latency, security, locality, power efficiency, etc.

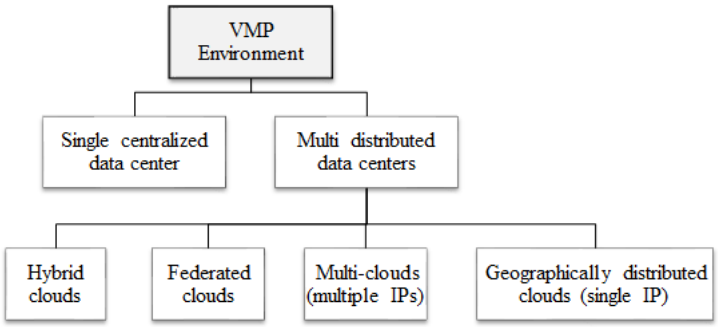


Figure 4.1: Classification of VMP Environments

The VNF placement problem, addressed in this chapter, is a kind of VMP problem since VNFs are virtual machines that run network functions. The VMP problem represents a well studied topic in the literature. Thus, we present, in the first place (section 4.2.1), a deep analysis of existing works related to this problem. To do so, we identify the different elements of problem formulation, optimization and evaluation methods. By analyzing these different methods to formulate and resolve the VMP problem in the literature, this helps us to have a clear picture of the VMP research landscape; to understand the methodology adopted to deal with the VMP problem; and thereby to efficiently address our context requirements.
In the second place (section 4.2.2), we present the most relevant works related to our problematic, namely the VNF placement in an edge-central cloud architecture.

4.2.1 Analysis of Existing VMP Approaches

We adopt a logical approach to analyze the different reviewed works based on two major steps (see Figure 4.2). The first step consists of studying methods and elements for the VMP problem formulation. Thus, we identify the different modeling approaches and mathematical tools adopted in the literature as well as the VMP objectives and constraints according to the user or service provider requirements. The second step of analysis is related to the optimized VMP computation dealing with the different used optimization strategies and methods as well as the considered VMP metrics to evaluate the obtained solutions.

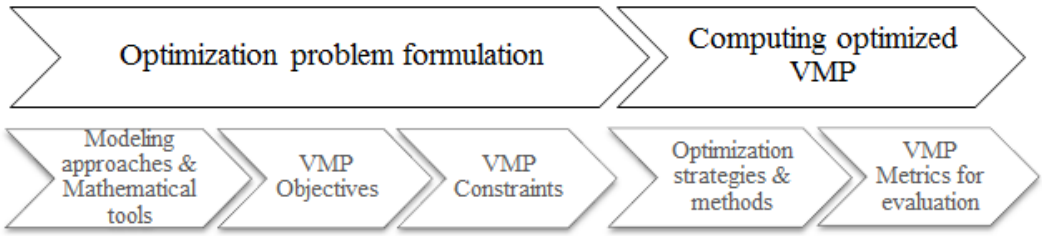


Figure 4.2: Analysis Methodology

4.2.1.1 Optimization Problem Formulation

We start here by the first step of analysis related to the formulation aspects of the VMP optimization problem. For that, we describe the modeling approaches and mathematical tools, the VMP objectives, and the VMP constraints.

4.2.1.1.1 Modeling Approaches and Mathematical Tools

1) Modeling Approaches

The VMP problem has largely been defined as a Combinatorial Optimization Problem (COP). The most commonly used COPs in the literature to model this problem include, but not limited to, the Bin-Packing Problem (BPP), the Generalized Assignment Problem (GAP), the Quadratic Assignment Problem (QAP), the Multi-Commodity Flow Problem (MCFP), and the bipartite graph matching.

The *BPP* [71] was used by many previous works [72–80] in the context of VMP, considering that the set of VMs as items of different sizes to be placed into a minimum number of Physical Machines (PMs) analogous to bins with fixed capacity. *Multi-Dimension Bin Packing Problem (MD-BPP)* [81], also known as the *Vector Bin Packing Problem (VBPP)*, is a generalization of the classical BPP in which the PM (bin) capacity and VM (item) sizes are given by multiple resource dimensions (e.g., CPU, memory). This approach was adopted by different works [82–90].

The bin-packing approach is especially useful to minimize the cost of running the data center and to reduce energy consumption as it aims to tightly pack the required VMs into the least possible number of PMs. However, packing as many VMs as possible onto a single PM may not be advisable in terms of QoS since this can lead to SLA violations due to overloading PMs.

The *GAP* [71] is another classical combinatorial optimization problem that was used or extended by a lot of works in the literature to model the VMP problem [91–96]. The objective

of this problem is to find the minimum cost assignment of n jobs (i.e., VMs) to m agents (i.e., PMs) such that each job is assigned to exactly one agent and a resource constraint is satisfied for each agent. The version of the problem in which more than one resource constraint is considered for each agent is called *Multi-Resource Generalized Assignment Problem (MR-GAP)*.

When considering network topology and communication traffic among VMs, the VMP problem is generally abstracted as *QAP* [97, 98] such as the case in [74, 99–103]. This problem consists of allocating a set of facilities (i.e., VMs) to a set of locations (i.e., PMs), with the cost being a function of the distance and flow between the facilities, plus costs associated with a facility being placed at a certain location. The objective is to assign each facility to exactly one location and vice-versa such that the total cost is minimized. The *Generalized Quadratic Assignment Problem (GQAP)* is a generalization of the QAP, with the difference that multiple facilities can be assigned to a single location subject to resource capacity at locations.

The VMP is also modeled in some works [102, 104] as a variant of the *MCFP* as they aim to minimize the total network power consumption.

Finally, the allocation of a workload to a substrate infrastructure can be viewed as a *graph matching problem* where both physical infrastructure and workloads are modeled as two graphs [104–108]. In the particular case of geographically distributed clouds, [107] and [108] formulate the VMP problem as a bipartite graph $G = (V1, V2, E)$, where $V1$ denotes the set of data centers, $V2$ denotes the location of customers, and E denotes the communication paths between customers and data centers.

It is to be noted that, in some works, VMP is considered as a combination of two optimization problems. For example, in [74], the optimization of physical servers by VM placement is abstracted as a BPP, while the optimization of network resources by using network topology and communication traffic is abstracted as a QAP. In [102], the distance-aware VM-group to server-rack mapping is modeled as a QAP, while the power-aware inter-VM traffic flow routing is modeled as a MCFP.

2) Mathematical Tools

We present, in the following, some of the most important formulations of the VMP problem and classify them according to their mathematical sources.

- ***Graph representation***

The VMP problem is generally formulated using graph theory when network topology and traffic demands are considered and the objective is to minimize communication costs and overheads or to satisfy SLA requirements in terms of response time. In the literature, graphs

are used to model different parts of the problem. In some works [109–111], only physical infrastructure is represented as a graph with vertices denoting physical servers and network switches (within a data center) and edges corresponding to communication links. In other works [99, 112], only workloads are modeled as a graph with a set of VMs with communication requirements between them.

- ***Linear Programming***

A Linear Programming (LP) problem is an optimization problem wherein the objective and all of the constraints are linear functions of the decision variables. It is widely adopted due to its simplicity and different methods of resolution. For example, in [113–117], authors provide LP model to formulate the VMP problem as the objective function and constraints are considered as simple linear functions.

- ***Quadratic Programming***

The next level of complexity beyond linear programming is quadratic programming. This model includes nonlinearities of a quadratic nature into the objective function. It is used by few works [101, 118] when the quadratic nature is due to considering network topology between hosts.

- ***Game theory***

Recently, game theory has been applied to solve resource competition problems in cloud computing. Ye and Chen [119] studied non-cooperative games amongst VMs on multidimensional VMP problem within a cloud data center. Zhang et al. [118] modeled the VMP problem as a multi-person non-cooperative game in the context of geographically distributed cloud infrastructures. Game theory has also been heavily applied in the particular case of cloud federations where resource allocation is mainly modeled as a cooperative game between federated cloud providers [120–124].

4.2.1.1.2 VMP Objectives

The VMP problem aims to find the optimal mapping of VMs to PMs driven by a particular objective. We describe below the most relevant objectives that have been adopted by existing approaches to solve the VMP problem.

- ***Performance objectives***

The VM placement on physical servers has been traditionally driven by mainly performance objectives which consist of optimizing SLA requirements based on certain parameters such as response time and throughput. SLA requirements are generally mapped to resource requirements which are often outlined as CPU usage, memory, and bandwidth.

pMapper [125] considers performance benefit maximization as one of the objectives of the application placement problem while performance is measured in terms of response time. Iqbal et al. [126] propose algorithms to detect violation of response time requirements and resolve bottlenecks that are caused by over utilization of CPU, memory, and I/O resources. The objective function in [127] is to maximize the total infrastructure performance of the deployed VMs measured in terms of computing capacity and throughput (i.e., the number of completed jobs per minute). In [99], Meng et al. aim to minimize average traffic latency caused by network infrastructure and, consequently, save bandwidth usage between VMs within a single data center. Alicherry and Lakshman [110] also aim to minimize the maximum latency in communication between VMs but in the context of distributed clouds.

- ***Minimize energy consumption***

Reducing energy consumption has becoming a major issue in cloud data centers as it represents an important part of the total operating cost and it has a significant adverse impact on the environment. Thus, cloud providers are increasingly aware by the importance of placing VMs onto physical servers efficiently so as to minimize the number of active physical resources and thereby energy consumption. There is a large initiative in the industry as well as academia to develop solutions that will help to create "green" data centers and optimize energy consumption when consolidating hardware resources under performance constraints.
Most of works on energy-efficient VMP have mainly focused on minimizing the number of hosts (i.e., computer servers) either in a single data center [125, 128–136] or in distributed data centers [80, 118]. However, some recent studies considered also network resource optimization within a single data center [103, 109, 111, 137, 138] taking into account the network topology as well as network traffic demands to meet the objective of network power reduction.
The survey [139] provides an overview of power-aware VMP strategies in a cloud data center. The survey [140] focus on energy-efficient server consolidation via live migration of VMs.

- ***Minimize migration cost***

The ability of VM migration between physical machines in a data center or even in different data centers is leveraged to dynamically optimize the VM placement. VM migration is essential to increase the flexibility in VM provisioning, avoid bottlenecks, and guarantee the service availability. However, VM migration has a direct impact on the virtual machine's performance [141].
Some recent research works take into account the migration overhead in the VMP optimization problem and adopt different migration cost models. In [113], the cost for VM migration is approximated by looking at the time required to shut down a VM in one cloud provider and start a new VM with the same configurations in another provider. According to [72], each migration is characterized by a migration duration and a migration cost. The latter is estimated by quantifying the decrease

in throughput because of live migration and estimating the revenue loss because of the decreased performance. In [117], T.C. Ferreto et al. consider that performance penalty of VM migration can be admissible when the VM capacity is being increased, since it will eventually result in a better performance after the migration, or when the capacity is being decreased as the performance may not affect the current workload demand. Nevertheless, they consider that only workloads with steady capacity harm the performance on the workload execution due to the migration cost. So, they adapted their problem to keep them in the same physical servers without migrating. [142] addresses dynamic re-allocation of VMs while minimizing the migration cost defined in terms of metrics, such as CPU and memory usage. M. Sharifi et al. [136] take into account the power cost of VM migration in their VMP strategy as they aim to minimize the total power consumption.

4.2.1.1.3 VMP Constraints

When formulating the optimization problem of VMP, several constraints can be taken into account. We cite below the most relevant ones.

- ***Resource demand and capacity constraints***

In order to achieve a valid VMP and an efficient resource allocation, two principal constraints should be satisfied. First, each VM requires a certain amount of resources such as CPU, memory, storage, and bandwidth to meet its performance goals. Second, physical hosts are limited by a certain capacity that should not be exceeded such as computation capacity, network I/O, memory, and disk capacity.

- ***SLA performance requirements***

In an SLA-based environment with fixed performance guarantees, the latter are considered as constraints in the optimization problem instead of being as metrics to be maximized. Performance constraints could be expressed as various metrics. Besides performance metrics which can be translated into resource requirements such as response time and throughput, there are other particular performance constraints. For example, in [118], a maximum delay between a client location and a data center is specified as an SLA performance constraint. In [143], A. Amokrane et al. define a location constraint as some services may require to be placed in the proximity of end-users to not degrade its performance or can only be placed in a particular set of data centers whereas others may not have such location constraints and can be placed in any data center. In [144] and [100], authors take into account a communication constraint that assigns VMs with large mutual traffic exchange to PMs in close proximity to each other. This constraint permits to guarantee network performance in terms of traffic latency and bandwidth which affects the overall performance.

- ***Power budget constraint***

Instead of minimizing the power consumption, a power budget can be fixed as a constraint in the VMP optimization problem [125, 145]. In [145], power budget is treated as a special kind of system resource whose consumption is calculated based on the load level of the VM, i.e., the utilization of CPU and the memory. So, the constraint on the power budget implicitly constrains the consumption of the system resources.

4.2.1.2 Computing optimized VM placement

We move now to the second step of analysis dealing with the computation of the optimized VMP. To do so, we present the different used optimization strategies and methods and the relevant metrics to evaluate the obtained solutions.

4.2.1.2.1 Optimization Strategies and Methods

Solving the VMP problem is NP-hard, as it falls mainly either in the category of BPP, GAP, QAP, or MCFP optimization problem (see section 4.2.1.1.1) and all these problems are known to be NP-hard [146, 147]. Therefore, large problem sizes cannot be solved optimally in reasonable time. As methods for solving the VMP problem, three major types of approaches have been used: exact methods, heuristics, and metaheuristics.

- ***Exact methods***

Exact methods are guaranteed to compute an optimal solution. However, the running time often grows exponentially with the problem size (i.e., number of VMs and hosts), and often only small sized problem can be practically solved optimally [148]. Among the exact methods, we mention branch-and-bound and dynamic programming, and (integer) linear programming based methods. The latter have been used by some works [83, 113, 114, 127] to solve optimally the VMP problem especially for small scale problems. To implement and solve these methods, there are multiple solvers that can be used such as GNU Linear Programming Kit (GLPK) [149], IBM ILOG CPLEX [150], and Gurobi [151].

- ***Heuristics***

Computation time for solving the VMP is crucial especially in dynamic environments. Moreover, the used algorithm to solve the optimization problem should be scalable in large sized problems. Consequently, heuristics or approximation algorithms are proposed to solve the VMP problem to hopefully find a 'good' feasible solution in short computation time (bounded by a polynomial in the input size). For the bin packing problem, there are a set of commonly used algorithms, namely, First-Fit (FF), Best-Fit (BF), First-Fit Decreasing (FFD), and Best-Fit Decreasing (BFD) [71, 152].

For example, in [145], a FF heuristic is used to solve the VMP problem in the data center formulated as a multi-level generalized assignment problem. In [80], authors propose a derivation of the BF heuristic to place VMs in the data center. Several other greedy heuristics have also been proposed in the literature to solve the VMP such as [127], [109], [74], [84] and [102]. Finally, we mention the linear programming relaxation based heuristic used in [114], [89] and [96].

- ***Metaheuristics***

As formally defined in [153], "*A metaheuristic is an iterative generation process which guides a subordinate heuristic by combining intelligently different concepts for exploring and exploiting the search space, learning strategies are used to structure information in order to find efficiently near-optimal solutions.*" This class of algorithms includes, but is not restricted to, Ant Colony Optimization (ACO), Evolutionary Computation including Genetic Algorithm (GA), Iterated Local Search, Simulated Annealing, and Tabu Search. The most used algorithms to solve the VMP problem in the literature are the ACO and GA which are both nature-inspired algorithms. They incorporate a learning component in the sense that they implicitly or explicitly try to learn correlations between decision variables to identify high quality areas in the search space. In [129], a modified version of the ACO is proposed to solve a multi-objective VMP problem in large-scale data centers where both power consumption and resource wastage should be minimized. [128] used another version of ACO to address the workload consolidation modeled as a MD-BPP. In [103], ACO and 2-opt local search are combined to solve the VMP problem.
A GA is proposed in [138] for VMP problem that considers the energy consumption in both the servers and the communication network in the data center. In [154], a modified GA with fuzzy multi-objective evaluation are proposed to deal with the VMP problem formulated as a multi-objective optimization problem aiming to simultaneously minimize total resource wastage, power consumption, and thermal dissipation costs.

4.2.1.2.2 VMP Metrics for Evaluation

Different metrics are used to evaluate the performance and quality of the proposed algorithms and to compare different VMP approaches. We describe below the major metrics used in the research works.

- ***Computation time***

Execution (or running) time is the time taken by an algorithm to compute the accurate placement plan. It is a crucial factor particularly for dynamic on-line VMP, on the one hand, and to evaluate the scalability of the algorithm to large environments and large number of VMs [125], on the other hand.

- ***QoS-related metrics***

QoS-related metrics measure the impact of a placement plan on the service quality after the placement of VMs has been done. In [91] and [129], throughput (in terms of transaction or job per unit of time) is measured to evaluate the VM performance. Another QoS metric that has been used by [99] and [155] is latency caused by network infrastructure and that should be minimal. In [125], N. Bobroff et al. used the rate of SLA violations as a metric to compare their proposed dynamic consolidation algorithm to a static consolidation approach. Demand satisfaction is another metric that determines whether sufficient resources are provided to the service to meet its SLA. In [105], this metric is measured relative to the demand of server resources (e.g., CPU, memory) by calculating the ratio of requested resource demand to allocated resources. However, in [109], this metric is measured relative to the demand of network resources by calculating the proportion of unsatisfied network demands.

- ***Cost-related metrics***

The cost of the VMP is related to the amount of substrate resources used by the VMs. It can be calculated explicitly using, for example, the hourly price for running a VM and the communication cost used respectively by [91] and [105]. It can also be calculated implicitly by determining the number of used hosts [73, 83] and the number of active network elements [102].

- ***Power consumption***

Power consumption can be considered as an implicit cost-related metric. Nevertheless, it is classified separately given its increasing importance nowadays in the research works. There are two types of this metric raised in the literature: power consumption of servers [83, 86, 125] and power consumption of network elements [102, 109].

4.2.2 Related Literature Review

A plethora of research works has addressed the VMP problem within a data center, in distributed clouds, or in hybrid clouds. The latter two architectures are the closest to our edge-central cloud architecture. Indeed, our architecture can be considered as a kind of distributed clouds. Also, it has the same aspect of a hybrid architecture where an organization operates a private local cloud and is able to externalize workloads to public distant IPs. However, in our case, both the local and distant clouds are managed by the same organization. We explore, in the following, relevant research works related to these two topics, namely VMP in distributed clouds and hybrid clouds, as well as initial studies on placement of VNFs.

4.2.2.1 VM Placement across Geographically Distributed Clouds

Several research works have been conducted for decision on the placement of VMs in geographically distributed data centers. For instance, Son et al. [156] propose an SLA-based cloud computing framework to facilitate location- and load-aware resource allocation. According to a utility function that involves machine workload and the expected response time, the user's VM is allocated in the physical machine that is closest to the user and has a light workload to guarantee a reasonable response time. However, this work places VMs one by one and does not address the whole system performance optimization.

In [157], Tordsson et al. propose a cloud brokering approach that optimizes placement of VMs across multiple clouds while maximizing an abstract VM performance metric with respect to a maximum total cost and a set of user defined constraints (e.g., load balancing, hardware configuration, etc.). VM hardware configurations are limited to a fixed number of instance types as they are offered by cloud providers. Moreover, geographical location regarding end-users is not considered in this work.

Unlike [157] which only addresses static scenario, Zhang et al. [158] present a framework for dynamic service placement problems in multiple data centers to minimize the total resource cost while satisfying SLA requirements, taking into consideration the fluctuation of both demand and resource price. As SLA constraint, they specify a maximum delay to achieve between a data center and a user location. However, they do not deal with resource provisioning to satisfy response time requirements for each VM.

In [159], authors propose a network-aware algorithm for allocation of virtual machines in distributed cloud systems. Their objective is to minimize the latency in communication between the VMs allocated for a user request when they are split over multiple data centers. Authors consider that the use of distributed cloud architectures enables to serve customer requests from locations close to them and thereby reduce network capacity needs and access latency. Consequently, they only consider latency between VMs and do not take into account latency between user location and the different selected data centers for the placement of user's VMs.

The aforementioned works are mainly conducted under the context of a cloud computing environment where IPs tend to build large data centers in geographically distributed locations to achieve reliability while minimizing operational cost, on the one hand, and where Service Providers (SPs) leverage geo-diversity of data centers to serve customers from multiple geographical regions, on the other hand. This context differs from our study scope which focuses on carrier network architecture leveraging NFV concepts. Indeed, the first context generally involves the placement of many and small VMs in large-scale data center infrastructure while the latter involves the placement of few and resource-intensive VNFs in a two-tier cloud infrastructure with small edge cloudlets.

4.2.2.2 VM Placement in Hybrid Clouds

Several works have been conducted to address the VMP in hybrid cloud scenario [160–165]. The main objective of these works is to ensure efficient utilization of the on-premise resources and to minimize the cost of running the outsourced tasks in the cloud, while fulfilling the applications' quality of service constraints. In our context of study, there are no costs relative to the price of VM deployment in external clouds since the whole two-tier cloud infrastructure belongs to the same organization (i.e., the wireless network carrier). However, network delay between the on-premise and the distant cloud can generate costs related to SLA violation penalties. Nonetheless, the aforementioned works can not be applied to our problematic since our objective is to minimize the maximum utilization of the edge cloud. In contrast, the cited works aim to maximize utilization of the internal data center. This approach used in hybrid clouds will be compared to our solution in the evaluation section.

4.2.2.3 VNF Placement

As NFV is becoming a hot topic across industry and academia, the problem of VNF placement has recently gained the attention of some research works. For instance, Bari et al. [166] propose a model to optimize the VNF placement problem in Internet Service Provider (ISP) and enterprise networks while minimizing operational costs, mainly node and link resource utilization. As SLA constraint, they only consider propagation delay and do not include processing delay at each node.

In the same context, Bernadetta et al. [167] propose a VNF chaining and placement model that optimize network level (i.e. link utilization) and NFVI-level (i.e. allocated computing resources) performance metrics. For this, they consider the latency bounds at both the VNF node and the end-to-end levels. To determine the VNF forwarding latency metric, two different regimes, called 'standard' and 'fastpath', are used. In the first one, forwarding latency is considered as a linear function of the aggregate bit-rate at the VNF. In the second one, the forwarding latency is constant up to a maximum aggregate bit-rate after which packets are dropped. Both of these regimes are based only on the traffic load and do not take into account utilization level of the physical host (i.e. the number of VMs placed on the NFVI node) which can dramatically increase the VNF latency due to the virtualization overhead and resource sharing of the same physical host. Moreover, it is worth mentioning that both of [166] and [167] do not consider the latency between VNFs and end-users and the fact that some VNFs may require to be placed in the proximity of end-users.

In [168], authors address the placement of virtual mobile core network functions (i.e, S-GW, PDN-GW, MME and HSS) excluding VNFs on the radio access network. Their optimization target is to minimize the cost of occupied link and node resources while taking as constraints VNF requirements in terms of bandwidth, processing and storage resources. However, they do not consider latency constraint on the VNF nodes and the end-to-end network.

In the same context of mobile networks, Taleb et al. [169] propose algorithms to place VNFs of both PDN-GWs and S-GWs on a given topology of distributed datacenters. They deal with two conflicting objectives, namely the insurance of QoE via the placement of VNFs of PDN-GWs closer to User Equipments (UEs) and the avoidance of the relocation of S-GWs via the placement of their VNFs far enough from UEs. While their results are promising, their scope is very limited to two particular mobile core network functions (i.e., S-GW and PDN-GW), on the one hand, and VNF resource requirements are not addressed, on the other hand.

In [170], authors investigate the VNF placement problem in the RAN domain which can include functions such as load-balancing, firewall, and virtual radio nodes. Their objective is to minimize the cost of mapping virtual functions to substrate network (nodes and links) while satisfying VNF requirements in terms of CPU, memory, storage, radio, and bandwidth resources. However, neither latency of VNF nodes nor end-to-end latency perceived by the user are considered.

In addition to the mentioned shortcomings, none of the aforementioned works deals with the resource provisioning problem. Indeed, specific resource requirements for each VNF are determined in advance or a set of predefined VM templates are used to embed VNFs. However, these values can change as a function of workload variation and physical host performance and utilization level. So, a joint placement and provisioning problem should be considered to satisfy VNF SLA requirements with respect to the underlying infrastructure performance and status.
To the best of our knowledge, we are the first to address VNF placement and provisioning problem over an edge-central cloud architecture while considering SLA requirements (VNF processing delay and real-time constraints), virtualization overhead and utilization level of the physical hosts.

4.3 QoS-driven VNF Placement and Provisioning in Edge-Central Carrier Cloud Architecture

In this section, we describe our performance and QoS models used in our optimization framework (section 4.3.1) and we provide mathematical formulation (section 4.3.2) and solutions (section 4.3.3) of our problem. Table 4.1 presents key symbols used in this section along with their definitions.

4.3.1 System Modeling

We present hereafter the performance and QoS models on which is based our QoS-driven VNF placement and provisioning methods for wireless carrier edge-central cloud system.

Symbol	Definition
N	Total number of VNFs
j	is equal to *1* if it refers to the cloudlet server and is equal to *2* if it refers to the cloud server
n_1, n_2	Number of VNFs that will be placed in the cloudlet and the cloud respectively
λ	Arrival rate of user requests
μ_j	Average service rate of requests by a unit of processing capacity (i.e., one CPU core) in server j
μ_{VMM_j}	Average service rate of requests by the VMM in server j
μ	Average service rate of requests by the network link between the cloudlet and the cloud servers
p_0, q_0	Probability that a request completing its service in the cloudlet (resp. in the cloud) moves to the cloud (resp. to the cloudlet)
p_s	Probability that a user request is accomplished
p_i, q_i	Probability that a request passing through the VMM in the cloudlet (resp. in the cloud) goes to the Virtual Machine i
C_j	Processing capacity of server j (i.e., number of CPU cores)
x_{ij}	Binary variable to determine if the i^{th} VNF is placed on server j (1) or not (0)
ϕ_{ij}	Number of CPU cores of server j allocated to the i^{th} VNF
y_i	Pseudo binary variable to determine if Max_Delay_i is respected (0) or violated (1)
U	Utilization rate of the cloudlet server capacity

Table 4.1: Notations and Definitions

4.3.1.1 Performance Model

Performance requirements of VNFs (instantiated within VMs) are generally specified in terms of average response time. This metric depends on various factors. First, by allocating more computing resources (i.e., CPUs), virtual machines usually improve their processing power and response time. Thus, sufficient resources have to be allocated to each VNF to guarantee the requested average response time.

Second, since a VNF potentially shares resources with other VMs on the same physical host, the response time also depends on the utilization level of the physical host (e.g., number of instantiated

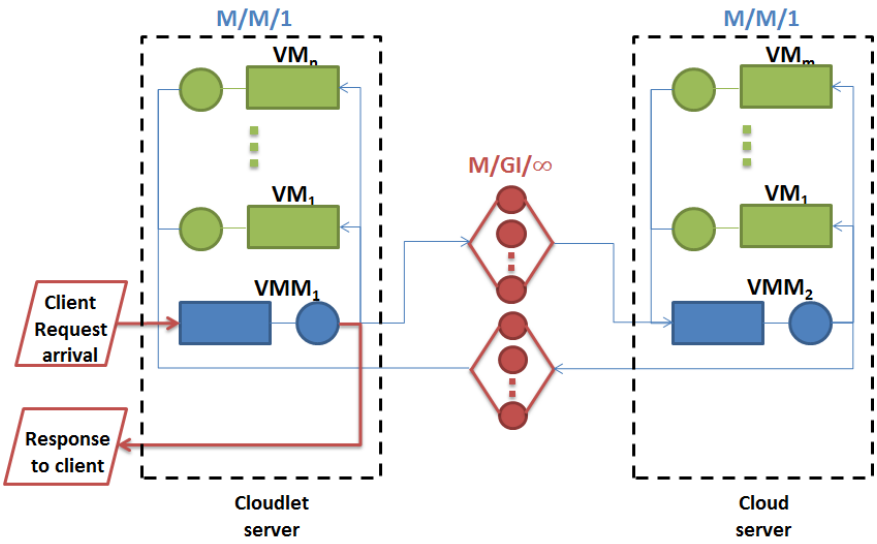


Figure 4.3: Two-tier cloud system model

VMs, number of user requests,etc.). If the latter is under a heavy workload, the VNF response time will be degraded.
Moreover, all incoming and outgoing user traffic passes through a virtualization layer (i.e., Virtual Machine Monitor (VMM) or an hypervisor), and incurs additional latency. Despite the several technologies used to reduce virtualization overhead (e.g., [171]), this still has an impact on the response time. Unless this is not explicitly modeled, response time will be inaccurate.
Finally, placing VNFs in the centralized cloud is likely to result in better response times due to the high performance capacity of the cloud. However, the network delay plays very important role in the client-perceived QoS.

Taking into account these different factors, we model performances of our system using analytical queuing models. Figure 4.3 represents this model.

We define N as the number of VNFs to be initially placed in the carrier two-tier cloud network. Let n_1 and n_2 be the number of VNFs that will be placed on the edge cloudlet and the central cloud respectively. We model both the cloudlet and the cloud servers as two open Jackson networks [172]. We assume that the distribution of inter-arrival times of user requests coming from outside have a Poisson distribution. We denote the arrival rate of requests by λ. In each network, the VMM as well as VNF machines are modeled as $M/M/1$ queues.

Let p_0 (resp. q_0) be the probability that a user request, having finished being served by the cloudlet server (resp. the cloud server), goes (resp. goes back) to the cloud server (resp. cloudlet server) and p_s be the probability that a user request is accomplished. User requests arriving from

the exterior first move towards VMM_1, the VMM of the cloudlet server.

We denote by $\{p_i; 1 \leq i \leq n_1\}$ and $\{q_i; 1 \leq i \leq n_2\}$ the probability that a request having passed through VMM_1 (resp. VMM_2, the virtual machine manager of the cloud server) goes to a VNF_i machine within the cloudlet server (resp. the cloud server). Applying the *Little's Law* [172], the mean response time of VMM_1 is:

$$\mathcal{R}_{VMM_1} = \frac{1}{\mu_{VMM_1} - \frac{\lambda}{p_s}} \tag{4.3.1}$$

and the mean response time of VMM_2 is:

$$\mathcal{R}_{VMM_2} = \frac{1}{\mu_{VMM_2} - \frac{\lambda p_0}{p_s q_0}} \tag{4.3.2}$$

where μ_{VMM_j}, $j \in \{1, 2\}$, denotes the mean service rate of requests at VMM_i. Equation 4.3.1 and 4.3.2 actually represent the average delay induced by the virtualization layer (i.e., VMM). Besides, the mean response time of a VM_i instantiated on the cloudlet server is:

$$\mathcal{R}_{VM_{i1}} = \frac{1}{\phi_{i1}\mu_1 - \frac{\lambda p_i}{p_s}}; 1 \leq i \leq n_1 \tag{4.3.3}$$

and the mean response time of a VM_i instantiated on the cloud server is:

$$\mathcal{R}_{VM_{i2}} = \frac{1}{\phi_{i2}\mu_2 - \frac{\lambda p_0 q_i}{p_s q_0}}; 1 \leq i \leq n_2 \tag{4.3.4}$$

where $\mu_j, j \in \{1, 2\}$, denotes the mean service rate of requests per unit of processing capacity and ϕ_{ij}, $j \in \{1, 2\}$, denotes the number of processing units allocated to VM_i on server j. Note that $\{p_i; 1 \leq i \leq n_1\}$ and $\{q_i; 1 \leq i \leq n_2\}$ respectively depend on the number of VNFs placed on the cloudlet and the cloud. For the sake of simplicity, and while taking into account the total number of VNFs placed in each server, we assume that:

$$p_i = \frac{1 - p_0 - p_s}{n_1} \text{ and } q_i = \frac{1 - q_0}{n_2}, \forall i \in [1, N] \tag{4.3.5}$$

By defining the constants:

$$b_1 = \frac{\lambda(1 - p_0 - p_s)}{p_s} \text{ and } b_2 = \frac{\lambda p_0(1 - q_0)}{p_s q_0} \tag{4.3.6}$$

we can write equation 4.3.3 and 4.3.4 as:

$$\mathcal{R}_{VM_{ij}} = \frac{1}{\phi_{ij}\mu_j - \frac{b_j}{n_j}}, \forall i \in [1, N] \text{ and } \forall j \in \{1, 2\} \tag{4.3.7}$$

Finally, based on equation 4.3.1, 4.3.2 and 4.3.7, the mean response time of VNF_i, $i \in [1, N]$, on server j is:

$$\mathcal{R}_{VNF_{ij}} = \mathcal{R}_{VMM_j} + \mathcal{R}_{VM_{ij}} \tag{4.3.8}$$

where VM_{ij} is the virtual machine instantiated for VNF_i on server j.

As a result, this model allows us to determine the number of processing units (i.e., CPU cores) required to achieve the requested average response time for each VNF taking into account utilization level of the physical host (i.e., arrival rate of user requests and the number of hosted VMs) and virtualization overhead.

To model the network link between the cloudlet and the cloud, we use $M/GI/\infty$ queue in which we assume that requests arrive following a Poisson process [173] with parameter $\lambda p_0/p_s$ and the service times are General Independent (GI) with rate μ. In such queue system, when a request arrives, it is immediately served and does not wait and the average response time of this queue depends only on the mean of the service time distribution. Hence, the average network delay can be calculated as follows:

$$\mathcal{D}_{Network} = \frac{1}{\mu} \tag{4.3.9}$$

4.3.1.2 QoS Model

Real-time behavior is an important aspect of network functions. Indeed, these functions have different timing requirements to maintain end-to-end QoS. For instance, management functions (e.g., OSS and off-line charging systems) have high tolerance to delays. Networks functions in the control plane (e.g., policy management, firewalling and AAA) as well as in the application plane (e.g., analytics solutions, location-based services) can tolerate small timing delays. Networking infrastructure functions (e.g., access points, routers and switches) and packet processing functions (e.g., CDN and DPI) have a very low tolerance for timing delays with a major impact on the QoS and the user experience. Inspired by the classification of these network functions presented in [174], we define a QoS model (see Table 4.2) which includes three types of network functions according to their delay sensitivity: *real-time*, *near real-time*, and *non real-time* functions. Then, for each type we define i) a priority level, according to which network functions have to be placed in proximity to end-users, ii) a maximum tolerated delay, and iii) a penalty metric representing severity of violating this delay.

4.3.2 Problem Description and Formulation

In this section, we formally define the VNF Placement problem in edge-central carrier cloud architectures as a Mixed Integer Linear Program (MILP).

Type of network function	Priority level	Maximum delay	Penalty metric
Real-time	1	10 ms	5
Near real-time	2	30 ms	3
Non real-time	3	100 ms	1

Table 4.2: QoS model for network functions

The problem of placing a given VNF_i across the infrastructure involves two steps:

- First, assign VNF_i to a server j (i.e., the cloudlet or the cloud). Hence, we define the decision variable x_{ij} as:
$$x_{ij} = \begin{cases} 1 & \text{if } VNF_i \text{ is placed on server } j \\ 0 & \text{else} \end{cases} \tag{4.3.10}$$
- Second, allocate the required processing resources to each VNF_i placed on server j. To do so, we define the resource allocation variable ϕ_{ij} as the number of CPU cores allocated to the VM running VNF_i.

In addition, we define a pseudo binary variable y_i which determines if the maximum delay Max_Delay_i defined by the QoS model of VNF_i is respected ($y_i = 0$) or violated ($y_i = 1$). We assume that the delay between end-users and the cloudlet is negligible as the latter is placed in proximity to them. Thus, we only consider the network delay between the cloudlet and the cloud when VNF_i is placed on the latter. Then, y_i is defined as follows:

$$y_i = \begin{cases} 1 & \text{if } x_{i2}\mathcal{R}_{VNF_{i2}} + 2x_{i2}\mathcal{D}_{Network} \geq Max\_Delay_i \\ 0 & \text{else} \end{cases} \tag{4.3.11}$$

In our problem, a set of **constraints** should be respected:

- A VNF_i should be placed either in the cloudlet or the cloud:
$$\sum_{j=1}^{2} x_{ij} = 1, \forall i \in [1, N] \tag{4.3.12}$$
- If VNF_i is placed on server j, a certain amount of resources $\phi_{ij} \neq 0$ should be allocated to this VNF, else ϕ_{ij} will be equal to 0. We model this constraint as follows:
$$\begin{gathered} x_{ij} \leq \phi_{ij} \leq C_j x_{ij} \\ \forall i \in [1, N] \text{ and } \forall j \in [1, 2] \end{gathered} \tag{4.3.13}$$
where Cj is the capacity of server j.

- Each server has a capacity limitation C_j. Moreover, a utilization rate U of the cloudlet server capacity C_1 should not be exceeded to prevent over provisioning and bottleneck creation at the cloudlet node. These constraints are presented as follows:

$$\sum_{i=1}^{N} \phi_{i2} \leq C_2, \text{ and } \sum_{i=1}^{N} \phi_{i1} \leq C_1 U, \forall i \in [1, N] \tag{4.3.14}$$

- The response time of each VNF_i placed on server j should not exceed a threshold value T_i defined in the SLA of this function. Based on equation 4.3.7, this constraint is presented as follows:

$$\begin{gathered} x_{ij}\mathcal{R}_{VMM_j} + \frac{x_{ij}}{\phi_{ij}\mu_j - \frac{b_j x_{ij}}{\sum_{i=1}^{N} x_{ij}}} \leq T_i, \\ \forall i \in [1, N] \text{ and } \forall j \in [1, 2] \end{gathered} \tag{4.3.15}$$

and can also be written as:

$$\begin{gathered} x_{ij} \sum_{i=1}^{N} x_{ij} \leq (T_i - \mathcal{R}_{VMM_j} x_{ij})(\phi_{ij}\mu_j \sum_{i=1}^{N} x_{ij} - x_{ij} b_j) \\ \forall i \in [1, N] \text{ and } \forall j \in [1, 2] \end{gathered} \tag{4.3.16}$$

This represents a non-linear quadratic constraint which can be linearized using the Big-M reformulation [175].

- Finally, we model equation (4.3.11) using the following constraints:

$$My_i \geq x_{i2}\mathcal{R}_{VNF_{ij}} + 2x_{i2}\mathcal{D}_{Network} - Max\_Delay_i \tag{4.3.17}$$

$$\begin{gathered} M(1 - y_i) \geq -(x_{i2}\mathcal{R}_{VNF_{ij}} + 2x_{i2}\mathcal{D}_{Network} \\ - Max\_Delay_i), \forall i \in [1, N] \end{gathered} \tag{4.3.18}$$

where $M = \max(Max\_Delay_i, i \in [1, N])$. Based on equations 4.3.2, 4.3.7 and 4.3.8, constraints 4.3.17 and 4.3.18 can be transformed to quadratic constraints which can also be linearized using the Big-M reformulation.

Our optimization problem is based on three main **objectives**:

i) Minimizing the maximum utilization of the edge cloudlet U to prevent over provisioning and to have spare resources for eventual new function placement and thus minimize VM migration:

$$Min\ U \tag{4.3.19}$$

In our work, we assume that the centralized distant cloud has extensive resources. Thus, we only consider utilization level of the cloudlet as it has very limited resource capacity.

ii) Minimizing allocated computing resources in the centralized cloud to only allocate sufficient resources:

$$Min \sum_{i=1}^{N} \phi_{i2} \tag{4.3.20}$$

iii) Minimizing QoS violation in terms of real time requirements of each VNF:

$$Min \sum_{i=1}^{N} y_i P_i \tag{4.3.21}$$

where P_i is the penalty metric defined by the QoS model of VNF_i.

The third objective can be achieved by placing all the real time constrained VNFs in the cloudlet. However, in this case, it will be over-provisioned and the first objective will be violated. This results in a conflict between the two objectives and a trade-off has to be found.

4.3.3 Solutions Description

Our problem represents a Multiple Objective Decision Making (MODM) [176] (also known as Multiple Criteria Decision Making). We present, in the following, three solutions to resolve this MODM problem.

4.3.3.1 Trade-off between Cloudlet Utilization and QoS Violation (`To-CUQV`)

As the three objective functions are linear, the wireless network carrier can specify his preferences in terms of the relative importance for each objective, especially the two conflicting ones. Thus, the problem can be solved using the weighting sum method [176, 177]. To more accurately reflect the relative importance of each objective by using weights, the different objective functions should have similar orders of magnitude and ranges [177]. For this purpose, we define the aggregate objective as follows:

$$Min\ \alpha U + \beta \sum_{i=1}^{N} \frac{\phi_{i2}}{C_2} + \gamma k \sum_{i=1}^{N} \frac{y_i P_i}{P_{RT} N_{RT} + P_{NeRT} N_{NeRT}} \tag{4.3.22}$$

where α, β and γ are the preference weights of each objective; k is a scaling factor; P_{RT} and P_{NeRT} are penalty metrics respectively defined for real-time and near real-time functions (see Table 4.2); and N_{RT} and N_{NeRT} are number of real-time and near real-time functions respectively. The optimization program of the `To-CUQV` solution is formulated as follows:

Minimize $\boldsymbol{\alpha U + \beta \sum_{i=1}^{N} \frac{\phi_{i2}}{C_2} + \gamma k \sum_{i=1}^{N} \frac{y_i P_i}{P_{RT} N_{RT} + P_{NeRT} N_{NeRT}}}$

Subject to:

$\sum_{j=1}^{2} x_{ij} = 1, \ \ \forall i \in [1, N]$

$x_{ij} \leq \phi_{ij} \leq C_j x_{ij}, \ \ \forall i \in [1, N], \ \ \forall j \in [1, 2]$

$\sum_{i=1}^{N} \phi_{i2} \leq C_2, \ \ \text{and} \ \ \sum_{i=1}^{N} \phi_{i1} \leq C_1 U, \ \ \forall i \in [1, N]$

$\mathcal{R}_{VNF_{ij}} \leq T_i, \ \ \forall i \in [1, N], \ \ \forall j \in [1, 2]$

$My_i \geq x_{i2} \mathcal{R}_{VNF_{ij}} + 2x_{i2} \mathcal{D}_{Network} - Max\_Delay_i \ \ \text{and}$

$M(1 - y_i) \geq -(x_{i2} \mathcal{R}_{VNF_{ij}} + 2x_{i2} \mathcal{D}_{Network} - Max\_Delay_i), \ \ \forall i \in [1, N]$

Problem 1: To-CUQV problem

4.3.3.2 Fixed QoS Violation Threshold (FQVT)

This solution is proposed for wireless carriers desiring to fix the QoS violation threshold in advance to guarantee a certain level of QoS. Therefore, we use the min-max approach to minimize the utilization rate of the edge cloudlet and the allocated resources while respecting the specified threshold of QoS violation. This threshold is defined as the maximum QoS penalty cost tolerated by the carrier denoted by P_{Th}. Its value is defined as a percentage of the maximum possible penalty cost. The optimization model of the FQVT solution is formulated as follows:

Minimize $\boldsymbol{\alpha U + \beta \sum_{i=1}^{N} \frac{\phi_{i2}}{C_2}}$

Subject to:

$\sum_{j=1}^{2} x_{ij} = 1, \ \ \forall i \in [1, N]$

$x_{ij} \leq \phi_{ij} \leq C_j x_{ij} \forall i \in [1, N], \ \ \forall j \in [1, 2]$

$\sum_{i=1}^{N} \phi_{i2} \leq C_2, \ \ \text{and} \ \ \sum_{i=1}^{N} \phi_{i1} \leq C_1 U, \ \ \forall i \in [1, N]$

$\mathcal{R}_{VNF_{ij}} \leq T_i, \ \ \forall i \in [1, N], \ \ \forall j \in [1, 2]$

$My_i \geq x_{i2} \mathcal{R}_{VNF_{ij}} + 2x_{i2} \mathcal{D}_{Network} - Max\_Delay_i \ \ \text{and}$

$M(1 - y_i) \geq -(x_{i2} \mathcal{R}_{VNF_{ij}} + 2x_{i2} \mathcal{D}_{Network} - Max\_Delay_i), \ \ \forall i \in [1, N]$

$\boldsymbol{\sum_{i=1}^{N} y_i P_i \leq P_{Th}}$

Problem 2: FQVT problem

4.3.3.3 Fixed Maximum Cloudlet Utilization level (FMCU)

A common approach for resource management in enterprise computing systems is to maintain the utilization level under a pre-defined upper bound in order to guarantee optimal performance and quality of service. Indeed, response time has small magnitude of change under low utilization rate, but it increases exponentially as the utilization reaches the maximum capacity [178]. From the other hand, wireless carriers may desire to limit the utilization of the cloudlet server under a sufficiently high level such that more capacity is available for hosting future functions. Thus, it is up to the wireless carrier to choose an appropriate utilization level according to his requirements. Similar to the previous solution, we apply the min-max approach. We denote by U_{max} the specified utilization level of the cloudlet. The optimization model, in this case, is formulated as follows:

$$\textbf{Minimize } \boldsymbol{\beta} \sum_{i=1}^{N} \frac{\phi_{i2}}{C_2} + \boldsymbol{\gamma k} \sum_{i=1}^{N} \frac{y_i P_i}{P_{RT} N_{RT} + P_{NeRT} N_{NeRT}}$$

Subject to:

$$\sum_{j=1}^{2} x_{ij} = 1, \quad \forall i \in [1, N]$$

$$x_{ij} \leq \phi_{ij} \leq C_j x_{ij} \forall i \in [1, N], \quad \forall j \in [1, 2]$$

$$\sum_{i=1}^{N} \phi_{i2} \leq C_2, \quad \text{and} \quad \sum_{i=1}^{N} \phi_{i1} \leq C_1 \boldsymbol{U_{max}}, \quad \forall i \in [1, N]$$

$$\mathcal{R}_{VNF_{ij}} \leq T_i, \quad \forall i \in [1, N], \quad \forall j \in [1, 2]$$

$$My_i \geq x_{i2} \mathcal{R}_{VNF_{ij}} + 2x_{i2} \mathcal{D}_{Network} - Max\_Delay_i \quad \text{and}$$

$$M(1 - y_i) \geq -(x_{i2} \mathcal{R}_{VNF_{ij}} + 2x_{i2} \mathcal{D}_{Network} - Max\_Delay_i), \quad \forall i \in [1, N]$$

Problem 3: FMCU problem

4.4 Performance Evaluation

In this section, we evaluate the performance of our proposals. To do so, we present settings of the conducted simulations, the performance metrics that we have evaluated as well as the obtained results.

4.4.1 Simulation Settings

Physical Infrastructure

In our experimental setup, we consider a 2.10 GHz Intel Xeon Processor E5-2620 v4 with 8 CPU cores [179] for the edge cloudlet server and we consider a 2.50 GHz Intel Xeon Processor E7-8890 v3 with 18 CPU cores [180] for the cloud server. We assume that the network delay between the cloudlet and the cloud servers is 15 ms.

VNFs

As a practical use case, we target a carrier Wi-Fi network based on an edge-central cloud architecture described in the previous chapter. Thus, the range of values that we will consider in the following particularly apply to such networks. Nevertheless, as seen in section 4.3, our provisioning and placement models do not make any assumption about the particular type of wireless technology and can be as well applied to any kind of carrier wireless network including mobile networks such as LTE and LTE-Advanced.

We consider a set of VNFs ($N \in$ [5,12]) composed of 25% of real-time functions and 50% of near real-time functions. Response time requirements T_i are assigned to each VNF uniformly between 1 ms and 5 ms. The arrival rate of user requests (i.e., user packet flows) to the system λ is fixed to 1 request per second. We assume that a request is about 5 Mbits.

The mean service rate of requests μ_{VMM_1}, μ_{VMM_2}, μ_1 and μ_2 depend on the Processor Base Frequency (PBF) of the server (i.e., number of CPU cycles per second) and the average number of required cycles by a request. Thus, μ_{VMM_i} and μ_i are respectively fixed to $PBF_i/10^6$ and $PBF_i/4.10^6$, while assuming that the average number of CPU cycles required per request by the VMM and by a VNF's VM are respectively equal to $10^6\, cycles/request$ and $4.10^6\, cycles/request$. The average number of processing cycles required per bit is estimated to be 0.25 for the VMM and 0.8 for a VNF's VM [181].

The probability p_0 is a decreasing function of the cloudlet capacity (i.e., CPU cores) and an increasing function of the total number of VNFs. Thus we define the probability p_0 as (see Figure 4.4):

$$p_0 = \exp(-\frac{C_1 + 1}{N}) \tag{4.4.23}$$

In addition, we assume that $p_0 = q_0$ and $q_s = \frac{1}{N+1}$.

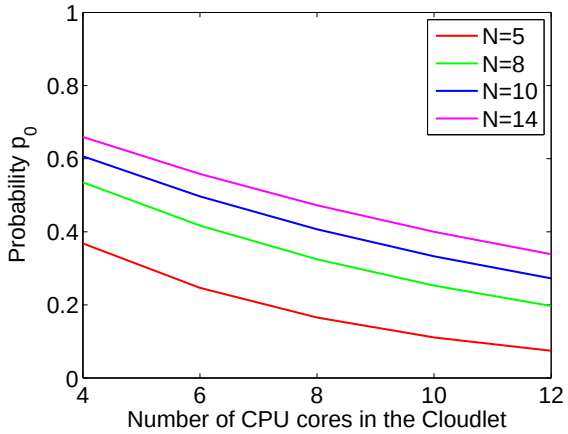


Figure 4.4: The probability p_0 function of the cloudlet capacity and total number of VNFs

The Simulator

To evaluate the proposed solutions, we have developed a simulator tool based on Matlab, CPLEX and YALMIP [182]. The results are obtained over many simulation instances (100) for each scenario and are calculated with a confidence interval of 95%.

4.4.2 The Baseline Approach

Since previous proposals on VM placement and VNF placement are not directly applicable to the studied scenario (see section 4.2.2), we developed a baseline VNF placement and provisioning algorithm similar to the approach used in hybrid clouds. Starting with VNFs having high priority level, the baseline algorithm first selects the cloudlet, if it is not-saturated, to place a VNF instance with adequate processing resources. Otherwise, the VNF instance will be placed on the cloud.

4.4.3 Performance Metrics

The main performance metrics used to evaluate our proposals and to compare them with the baseline approach are:

- ***The cloudlet utilization rate***

One of our goals is not to overload the cloudlet. Thus, the utilization rate of processing resource capacity after VNF placement represents an important performance metric. This is calculated as follows:

$$U = \sum_{i=1}^{N} \frac{\phi_{i1}}{C_1} \tag{4.4.24}$$

- ***The number of QoS violation***

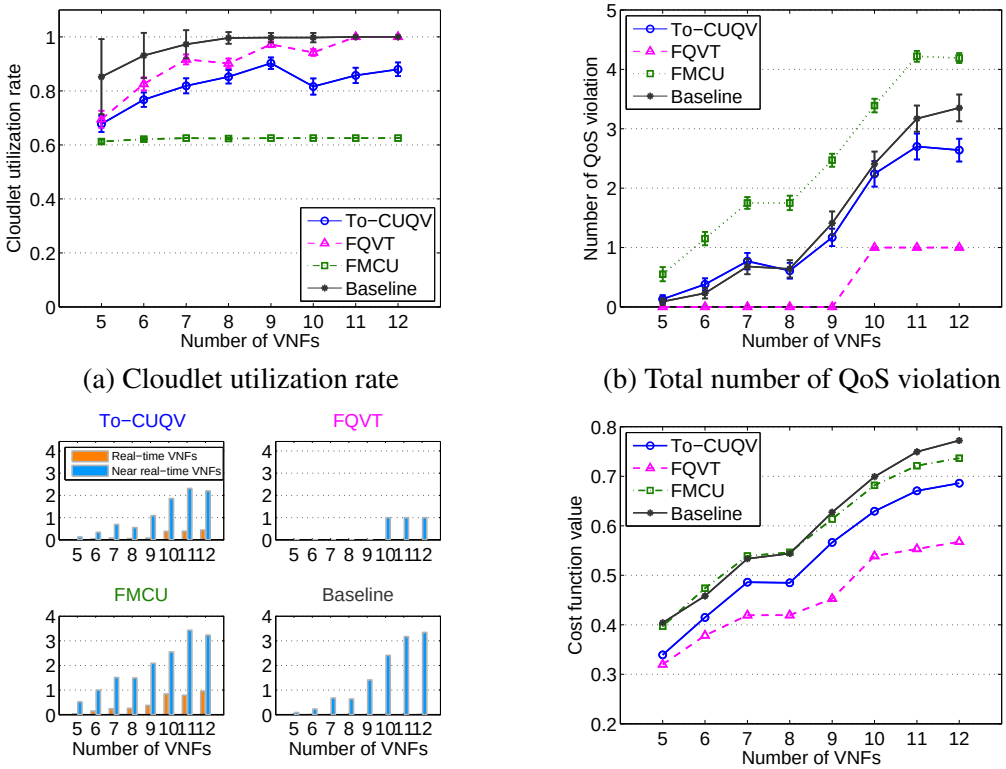
This metric represents the number of VNFs whose maximum delay, defined in their SLAs and representing the end-user perceived latency, has been exceeded. This metric only concerns VNFs placed on the distant cloud since in our scenario the Cloudlet is placed in the end-user premises.

- ***The cost function value***

This metric represents the global cost of VNF placement in terms of resource utilization and QoS violation. It is calculated using the cumulative objective function (Equation 4.3.22).

4.4.4 Simulation Results

In our experiments, we evaluate the different proposed solutions and we compare them to the baseline solution in terms of cloudlet utilization rate, occurrence of QoS vioaltion and cost. Moreover, we study the impact of the number of VNFs and the cloudlet computing capacities on all these



(a) Cloudlet utilization rate

(b) Total number of QoS violation

(c) Number of QoS violation per VNF class

(d) Cost function value

Figure 4.5: Comparison of the proposed solutions and the Baseline as a function of the number of VNFs

performance metrics. In these experiments, we set $\alpha = \gamma = 0.45$. β is set to a low value (i.e., 0.1) since the objective of minimizing the allocated cloud resources is not conflicting with the other ones. We also set $P_{Th} = 10\%$ for the `FQVT` solution and $U_{max} = 70\%$ for the `FMCU` solution.

Effect of the number of VNFs: Figure 4.5 compares the obtained results of the different solutions while varying the number of VNFs.

In Figure 4.5-(a), we can observe that the cloudlet becomes saturated with the baseline approach when the number of VNFs is approximately more than the number of available CPU cores (i.e., 8). In addition, as the baseline places VNFs by order of priority in the cloudlet, we can observe in Figure 4.5-(c) that there is no QoS violation of real-time functions but, in return, this leads to the cloudlet saturation.

On the other hand, `To-CUQV` always finds an optimal trade-off between the cloudlet utilization rate and the number of QoS violation while keeping the cost function as low as possible. Further-

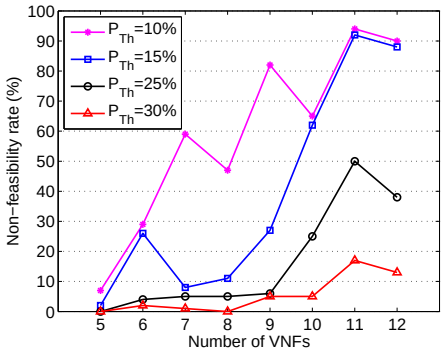


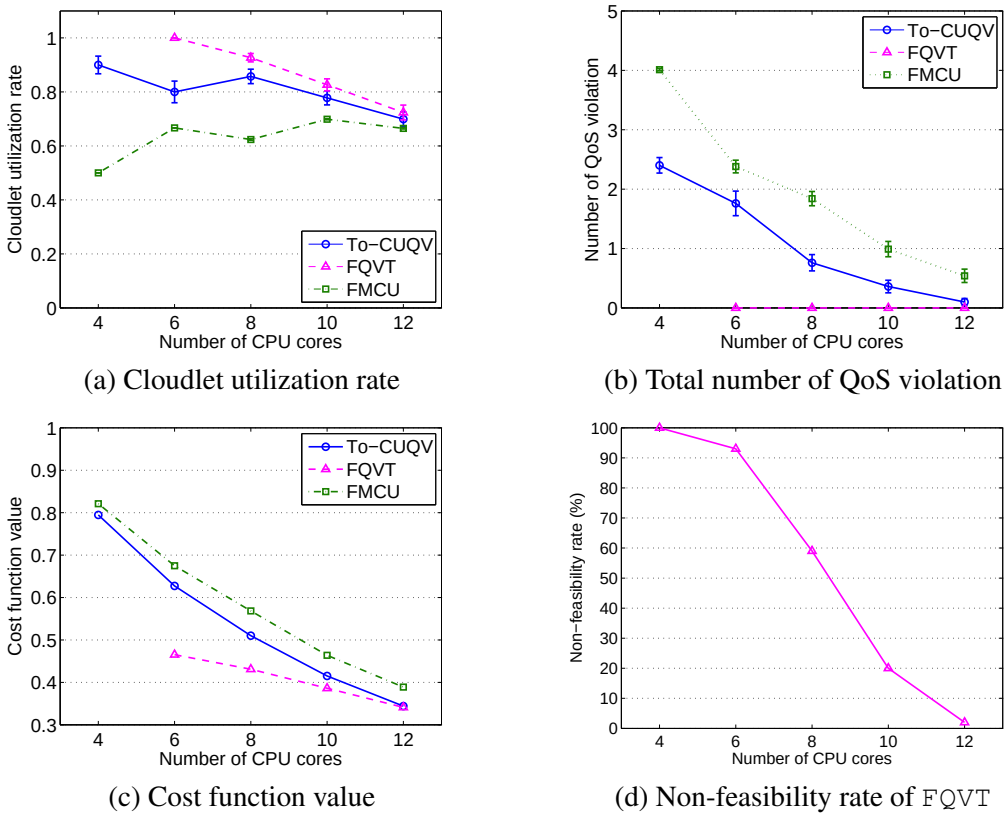
Figure 4.6: Non-feasibility rate of FQVT solution

more, compared to the baseline, To−CUQV reduces the total number of QoS violation particularly for $N > 8$ (see Figure 4.5-(b)). The reason is that To−CUQV may place real-time functions with heavy resource-consumption in the cloud when this enables to place additional near real-time functions in the cloudlet. Thus, To−CUQV does not only consider per-VNF QoS level but also the overall QoS level.

FMCU represents the least cloudlet utilization rate and respects the fixed maximum value (i.e., 70%). However, it generates the highest number of QoS violation (see Figure 4.5-(b) and (c)) and A high cost value (see Figure 4.5-(d)).

In Figure 4.5-(b) and (c), we can observe that FQVT ensures no QoS violation for $N \leq 9$. So, the cost shown in Figure 4.5-(c) for this interval corresponds only to the cost of resource utilization. It is clear that FQVT has the least cost and number of QoS violation. However, this solution is not always feasible. Indeed, the QoS violation threshold constraint can not always be satisfied. Figure 4.6 shows the non-feasibility rate in our experiments for $P_{Th} = 10\%$ as well as for other values (15%, 25% and 30%). We note that this rate is higher for low value of P_{Th} and it remarkably increases when $N > 8$. Only for $P_{Th} = 25\%$ and $P_{Th} = 30\%$, the solution becomes nearly 100% feasible for $N \leq 8$.

Effect of Cloudlet computing capacities: As we have noticed in the previous results, performance metrics degrade more remarkably when the number of VNFs becomes superior to the number of available CPU cores on the cloudlet. Thus, we fix the number of VNF to 8 and we vary the number of CPU cores to see more clearly the effect of this parameter on performances. As expected, Figure 4.7-(a), (b) and (c) show that performance metrics are improved when the number of CPU cores increases. We notice that FQVT solution is not feasible at all with 4 CPU cores. This means that it is impossible to have a QoS violation threshold of 10% of penalty cost with only 4 CPU cores. Figure 4.7-(d) depicts how the non-feasibility rate decreases as the number of CPU



(a) Cloudlet utilization rate

(b) Total number of QoS violation

(c) Cost function value

(d) Non-feasibility rate of FQVT

Figure 4.7: Comparison of the proposed solutions as a function of the number of CPU cores

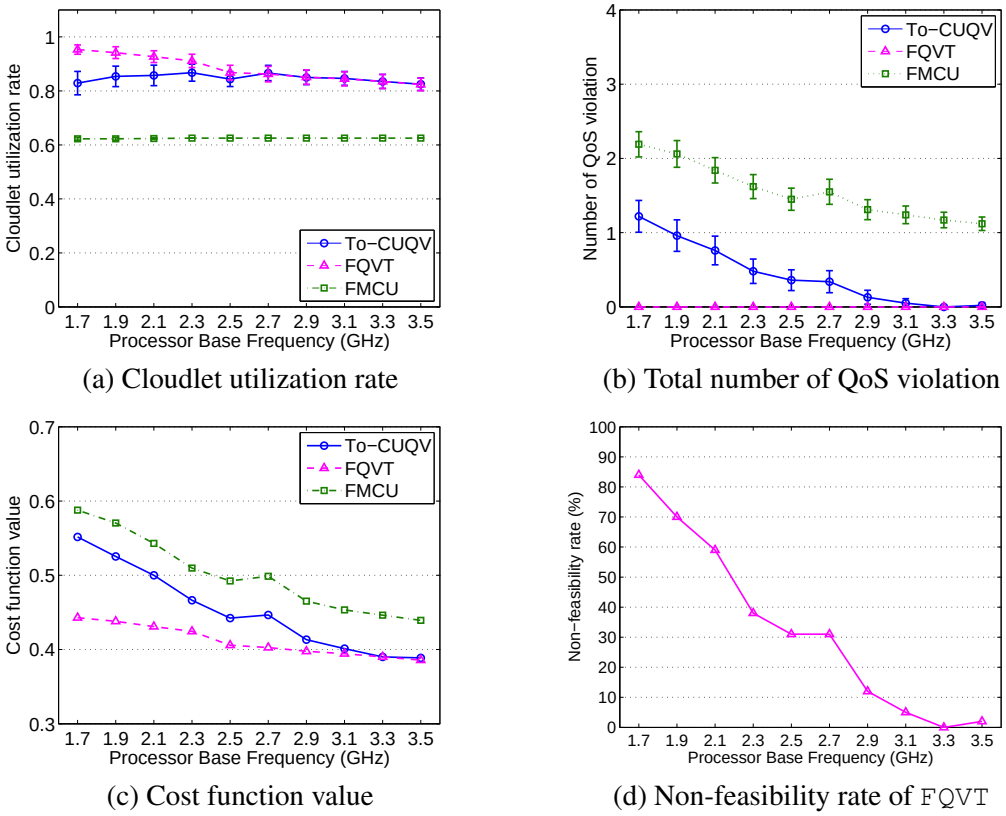
cores increases and it becomes almost 0 for 12 CPU cores.

On the other hand, we vary the PBF of the cloudlet server (see Figure 4.8). We can observe that the PBF has much clearer impact on the QoS level than on the cloudlet utilization rate. In fact, processors with a high PBF improve QoS level. Regarding FQVT solution, the effect of the PBF appears on its feasibility rate. Indeed, high PBF values provide better feasibility rate.

To summarize, this part represents a deep analysis and detailed results of our system performances using different parameters and solutions. These results can help carriers to choose the appropriate solution and dimension their system according to their needs and requirements.

4.5 Conclusion

The edge-central carrier cloud architecture along with NFV technology represent a promising solution for wireless operators to address future 5G challenges. In this chapter, we presented strategies



(a) Cloudlet utilization rate

(b) Total number of QoS violation

(c) Cost function value

(d) Non-feasibility rate of FQVT

Figure 4.8: Comparison of the proposed solutions as a function of the Processor Base Frequency

that enable QoS-driven VNF placement and provisioning in such environment using performance and QoS models. Simulation experiments and performance analysis show promising results. Indeed, our approaches are able to achieve a fair trade-off between two conflicting objectives, namely i) the optimization of resource utilization, especially in the capacity-constrained cloudlet system, and ii) the minimization of SLA violations. This trade-off depends on carrier requirements in terms of resource utilization and QoS level. In addition, an acceptable level of overall QoS is ensured. Finally, we showed how our system performances depend on the number of deployed VNFs and the cloudlet computing capacities.

Chapter 5

General Conclusion

Contents

5.1 Summary of Contributions

It is undeniable that Wi-Fi is gaining a lot of momentum in today's networks as well as in future networks. Hence, new carrier-grade requirements are emerging to support future user expectations and provide high-performance Wi-Fi networks. In this context, we investigated several problems surrounding the design and optimization of carrier-grade next-generation Wi-Fi networks.

More specifically, we addressed, in the first part of our thesis, emerging user experience requirements dealing with service discovery and access in carrier Wi-Fi networks. Our objective was to provide a personalized and seamless access to venue-based services offered through public Wi-Fi networks. Therefore, we proposed lightweight service discovery mechanisms prior to Wi-Fi association and defined unique and global identification and description scheme of these services. Through deep analysis and comparison with existing solutions, we showed that our solution has more advantages in terms of transparency, energy efficiency, user satisfaction, throughput and ease of deployment.

In the second part of our works, we addressed the challenge of conceiving and implementing a flexible carrier Wi-Fi architecture which promotes innovation, reduces network operation costs, and supports evolving contexts and service needs. Therefore, we proposed a novel architecture for

carrier-managed Wi-Fi networks that leverages Network Function Virtualization and Edge Computing concepts. This architecture is based on i) lightweight WTPs, ii) a WLAN Cloudlet that consolidates MAC-layer functions, network functions and value-added services, and iii) a centralized cloud platform to offer more scalability and control. This provides more agility and adaptability to integrate new services with minimum cost and offers more scalability to support increasing demand. This also decreases access latency by placing network functions and certain services close to end-users. The feasibility of our solution were proved through a proof-of-concept prototype and good performances were achieved.

In the final part, we addressed VNF management and deployment issues in the proposed architecture, in particular, and in wireless carrier edge-central network architecture, in general. More specifically, we proposed placement and provisioning strategies of VNFs in such architectures taking into account QoS requirements. Our objective was to optimize resource utilization, to prevent cloudlet overload and congestion, and to avoid violation of QoS requirements. For this purpose, queuing and QoS models along with optimization techniques were used to efficiently allocate resources and place VNFs. Simulation results showed that our proposed solutions are able to achieve a fair trade-off between two conflicting objectives, namely the optimization of resource utilization and the minimization of SLA violations. Moreover, an acceptable level of overall QoS is ensured.

5.2 Future Work

The problem of VNF management and orchestration remains a complicated task. Taking into account other parameters and constraints of the virtualized system seems to be important to perform more efficient VNF placement. Hence, our optimization VNF placement strategies may be enhanced by considering other resource dimensions such as memory and communication traffic requirements between VNFs. By considering network resources requirements, in particular, the VNF placement algorithm can place high-communicating VNFs within the same physical host with the aim of reducing wide-area communication costs and, potentially, service response times.

Furthermore, in order to accommodate dynamic workloads and network traffic load fluctuations, it is necessary to dynamically readjust VNF placement and resource provisioning. For this purpose, a set of mechanisms, such as VM live migration and horizontal / vertical scaling, may be enabled to ensure load balancing between the edge cloudlet and the central cloud and to resolve eventual bottlenecks in the cloudlet. Thus, an online VNF placement algorithm with low time complexity should be supported.

Last but not least, an interesting future direction is to implement the whole Wi-Fi carrier system, including i) the radio access network which supports the proposed pre-association service discovery mechanisms, ii) the WLAN edge cloudlet, iii) the central carrier cloud, and finally iv) the control and management system. The latter has the role of managing the edge-central cloud

infrastructure, orchestrating the allocation of resources needed by VNFs according to the defined algorithms, managing the VNF's lifecycle (e.g., instantiate VNF, scale VNF, update / upgrade VNF, and terminate VNF), and monitoring real-time network performances. For this, OpenContrail [183] and OpenStack [184] are examples of new solutions that could be useful.

5.3 Publications

This section summarizes the publications that have resulted from the work undertaken in this thesis.

- "Labels for common venue-based services," IETF Draft submitted as an individual submission, July 2014.
- "Personalized and seamless Wi-Fi access to Venue-Based Services," in the IEEE International Conference and Workshop on the Network of the Future-NOF, Paris, December 3-5, 2014.
- "Towards software-based carrier Wi-Fi architecture for a wider range of services," in the IEEE Global Communications Conference Workshops-GLOBECOM Wkshps, San Diego, December 6-10, 2015.
- "Cloudlet-and NFV-based carrierWi-Fi architecture for a wider range of services," in the Annals of Telecommunications, March 2016.
- "QoS-aware VNF Placement Optimization in Edge-Central Carrier Cloud Architecture," in the IEEE Global Communications Conference-GLOBECOM, Washington, December 4-8, 2016.
- "Analytical Models for QoS-driven VNF Placement and Provisioning in Wireless Carrier Cloud," submitted to the ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems-MSWIM, Malta, November 13-17, 2016.
- "Survey on Virtual Machine Placement in Virtualized Environments".

5.4 WBA Projects

It is worth mentioning that I have participated to the following WBA projects:

- **The NGH trial**

This trial aims to deliver a next generation Wi-Fi experience based on a seamless and secure access to public Wi-Fi access points on home and roaming partners' networks around the world. Furthermore, it helps drive the industry toward adoption of NGH networks. The trial involves end-to-end

inter-operator testing of the NGH requirements in a “real world" production environment. (See Appendix A for more details).

- **Wi-Fi Software Defined Networking (SDN) & Network Function Virtualisation (NFV) Guidelines**

The objective of this project is to develop a white paper [185] that examines the usage of SDN and NFV in the deployment of Wi-Fi network. More specifically, it provides use cases, an analysis of the state of the art, definition requirements, architectures alternatives, identification of gaps and challenges that Wi-Fi industry has to address to develop Wi-Fi SDN & NFV based solutions for Wi-Fi networks.

List of Figures

List of Tables

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020," White Paper, Cisco, February 2016.

[2] Wi-Fi Alliance, "Wi-Fi device shipments to surpass 15 billion by end of 2016." [Online]. Available: http://www.wi-fi.org/news-events/newsroom/wi-fi-device-shipments-to-surpass-15-billion-by-end-of-2016

[3] *Hotspot 2.0, Release 1*, Wi-Fi Alliance Technical Specification, Version 1.0.0, 2012.

[4] *Hotspot 2.0, Release 2*, Wi-Fi Alliance Technical Specification, Version 1.0.0, 2014.

[5] GSM Association, "IR.51 - IMS Profile for Voice, Video and SMS over Wi-Fi Version 3.0," March 2016.

[6] "Carrier Wi-Fi Guidelines," White Paper, Wireless Broadband Allliance, February 2014.

[7] Meteor Network. [Online]. Available: http://www.meteornetworks.com/

[8] F. Ben Jemaa and M. Pariente, "Personalized and seamless Wi-Fi access to Venue-Based Services," in *2014 International Conference and Workshop on the Network of the Future (NOF)*. IEEE, 2014, pp. 1–6.

[9] F. Ben Jemaa, G. Pujolle, and M. Pariente, "Labels for common venue-based services," Internet Draft, 2014. [Online]. Available: https://www.ietf.org/archive/id/draft-benjemaa-vbs-urn-00.txt

[10] F. Ben Jemaa, G. Pujolle, and M. Pariente, "Towards software-based carrier Wi-Fi architecture for a wider range of services," in *2015 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2015, pp. 1–6.

[11] F. Ben Jemaa, G. Pujolle, and M. Pariente, "Cloudlet-and NFV-based carrier Wi-Fi architecture for a wider range of services," *Annals of Telecommunications*, pp. 1–8, 2016.

[12] F. Ben Jemaa, G. Pujolle, and M. Pariente, "QoS-aware VNF placement optimization in edge-central carrier cloud architecture," in *2016 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2016.

[13] *IEEE Standard for Information technology–Telecommunications and information exchange between systems Local and metropolitan area networks–Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications- Amendment 9: Interworking with External Networks*, Std., 2011.

[14] Wi-Fi Alliance, "Wi-Fi Aware." [Online]. Available: http://www.wi-fi.org/discover-wi-fi/wi-fi-aware

[15] *Neighbor Awareness Networking*, Wi-Fi Alliance Technical Specification, Version 1.0, 2015.

[16] The Appache Software Foundation, "Apache River project (Jini Technology)." [Online]. Available: http://river.apache.org/

[17] UPnP Forum, "UPnP Device Architecture 2.0," 2015. [Online]. Available: http://upnp.org/specs/arch/UPnP-arch-DeviceArchitecture-v2.0.pdf

[18] E. Guttman, C. Perkins, J. Veizades, and M. Day, "Service location protocol, version 2," Internet Requests for Comments, RFC 2608, June 1999. [Online]. Available: https://www.ietf.org/rfc/rfc2608.txt

[19] The Salutation Consortium, "Salutation Architecture: Overview," White Paper, 1998. [Online]. Available: http://salutation.org/wp-content/uploads/2012/05/originalwp.pdf

[20] S. Cheshire and M. Krochmal, "Dns-based service discovery," Internet Requests for Comments, RFC 6763, February 2013. [Online]. Available: https://tools.ietf.org/html/rfc6763

[21] "Jini Discovery & Join Specification Version 3.0." [Online]. Available: https://river.apache.org/doc/specs/html/discovery-spec.html#19702

[22] Y. Y. Goland, T. Cai, P. Leach, Y. Gu, and S. Albright, "Simple service discovery protocol/1.0 operating without an arbiter," Internet Draft, October 1999. [Online]. Available: https://tools.ietf.org/html/draft-cai-ssdp-v1-03#section-2

[23] J.-S. Tsai, C. Liu, H. Elgebaly, and J. P. Kardach, "Service discovery architecture and method for wireless networks," U.S. Patent 7 403 512 B2, Jul. 22, 2008.

[24] D. Stephenson, E. Torres, J. Salowey, C. Ersoy, and N. Cam-Winget, "Pre-association mechanism to provide detailed description of wireless services," U.S. Patent 20 140 122 242 A1, May 1, 2014.

[25] V. Gupta and N. Canpolat, "Broadcast/multicast based network discovery," U.S. Patent 20 090 245 133 A1, Oct. 1, 2009.

[26] T. W. Kuehnel, A. A. Hassan, C. Huitema, D. Jones, S. Guven, S. J. Chan, S. R. Gatta, and Y. Lu, "Mechanism to convey discovery information in a wireless network," U.S. Patent 8 559 350 B2, Oct. 15, 2013.

[27] P. J. Leach, M. Mealling, and R. Salz, "A universally unique identifier (uuid) urn namespace," Internet Requests for Comments, RFC 4122, July 2005. [Online]. Available: https://tools.ietf.org/html/rfc4122

[28] "Status of Project IEEE 802.11aq - Pre-Association Discovery (PAD)." [Online]. Available: http://www.ieee802.org/11/Reports/tgaq_update.htm

[29] F. Yergeau, "UTF-8, a transformation format of ISO 10646," Internet Requests for Comments, RFC 3629, November 2003. [Online]. Available: https://tools.ietf.org/html/rfc3629

[30] H. Schulzrinne, "A uniform resource name (urn) for emergency and other well-known services," Internet Requests for Comments, RFC 5031, January 2008. [Online]. Available: https://tools.ietf.org/html/rfc5031

[31] L. Yang, P. Zerfos, and E. Sadot, "Architecture taxonomy for control and provisioning of wireless access points (capwap)," Internet Requests for Comments, RFC 4118, June 2005.

[32] "The Evolution of the Enterprise-Class Wireless LAN Access Point," Infonetics Research, January 2004. [Online]. Available: http://www.nsaservices.com/pdf/airespace/WP_Enterprise_Class_WLAN_AP.pdf

[33] "802.11 WLAN architecture, Best practices," White Paper, 3e Technologies International, February 2005. [Online]. Available: http://www.ultra-3eti.com/assets/1/7/WirelessArchitectureBestPractices.pdf

[34] "Distributed Intelligence: The Future of Wireless Networking Architecture," White Paper, Motorola, April 2011. [Online]. Available: http://www.csc.villanova.edu/~nadi/csc8580/S13/DistributedIntelligence.pdf

[35] "Meraki hosted architecture," White Paper, Meraki, Feb 2011. [Online]. Available: http://www.voyager.net.uk/wp-content/uploads/downloads/2014/02/WP-Meraki-Hosted-Architecture.pdf

[36] "The next logical evolution in WLAN architecture," Technical Brief, Motorola, 2014.

[37] "Virtualized architecture enables choice, efficiency, and agility for enterprise mobility," White Paper, Meru, 2012.

[38] J. Vestin, P. Dely, A. Kassler, N. Bayer, H. Einsiedler, and C. Peylo, "CloudMAC: towards software defined WLANs," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 16, no. 4, pp. 42–45, 2013.

[39] P. Salvador, S. Paris, C. Pisa, P. Patras, Y. Grunenberger, X. Perez-Costa, and J. Gozdecki, "A modular, flexible and virtualizable framework for IEEE 802.11," in *2012 Future Network & Mobile Summit (FutureNetw)*. IEEE, 2012, pp. 1–8.

[40] L. Xia, S. Kumar, X. Yang, P. Gopalakrishnan, Y. Liu, S. Schoenberg, and X. Guo, "Virtual wifi: bring virtualization from wired to wireless," in *ACM SIGPLAN Notices*, vol. 46, no. 7. ACM, 2011, pp. 181–192.

[41] O. Braham and G. Pujolle, "Virtual wireless network urbanization," in *2011 International Conference on the Network of the Future (NOF)*. IEEE, 2011, pp. 31–34.

[42] G. Aljabari and E. Eren, "Virtualization of wireless LAN infrastructures," in *2011 IEEE 6th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, vol. 2. IEEE, 2011, pp. 837–841.

[43] T. Hamaguchi, T. Komata, T. Nagai, and H. Shigeno, "A framework of better deployment for WLAN access point using virtualization technique," in *2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. IEEE, 2010, pp. 968–973.

[44] T. Nagai and H. Shigeno, "A framework of AP aggregation using virtualization for high density WLANs," in *2011 Third International Conference on Intelligent Networking and Collaborative Systems (INCoS)*. IEEE, 2011, pp. 350–355.

[45] "Network Functions Virtualisation: An Introduction, Benefits, Enablers, Challenges & Call for Action," White Paper, ETSI, October 2012. [Online]. Available: https://portal.etsi.org/nfv/nfv_white_paper.pdf

[46] Network Functions Virtualisation. [Online]. Available: http://www.etsi.org/technologies-clusters/technologies/nfv

[47] Network Function Virtualization Research Group. [Online]. Available: https://irtf.org/nfvrg

[48] Network Functions Virtualization Forum. [Online]. Available: http://www.atis.org/01_committ_forums/NFV/index.asp

[49] OPNFV. [Online]. Available: https://www.opnfv.org/

[50] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*. ACM, 2012, pp. 13–16.

[51] OpenFog Consortium. [Online]. Available: https://www.openfogconsortium.org/

[52] S. Davy, J. Famaey, J. Serrat, J. L. Gorricho, A. Miron, M. Dramitinos, P. M. Neves, S. Latré, and E. Goshen, "Challenges to support edge-as-a-service," *IEEE Communications Magazine*, vol. 52, no. 1, pp. 132–139, 2014.

[53] A. Manzalini, R. Minerva, F. Callegati, W. Cerroni, and A. Campi, "Clouds of virtual machines in edge networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 63–70, 2013.

[54] Q. Li, H. Niu, A. Papathanassiou, and G. Wu, "Edge cloud and underlay networks: Empowering 5g cell-less wireless architecture," in *Proceedings of European Wireless Conference*. VDE, 2014, pp. 1–6.

[55] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *IEEE pervasive Computing*, vol. 8, no. 4, pp. 14–23, 2009.

[56] K. A. Khan, Q. Wang, C. Grecos, C. Luo, and X. Wang, "Meshcloud: Integrated cloudlet and wireless mesh network for real-time applications," in *2013 IEEE 20th International Conference on Electronics, Circuits, and Systems (ICECS)*. IEEE, 2013, pp. 317–320.

[57] M. Felemban, S. Basalamah, and A. Ghafoor, "A distributed cloud architecture for mobile multimedia services," *IEEE Network*, vol. 27, no. 5, pp. 20–27, 2013.

[58] *IEEE Standard for Information technology–Telecommunications and information exchange between systems Local and metropolitan area networks–Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, Std., 2012.

[59] W. El-Hajj and H. Alazemi, "Optimal frequency assignment for ieee 802.11 wireless networks," *Wireless Communications and Mobile Computing*, vol. 9, no. 1, pp. 131–141, 2009.

[60] A. Farsi, N. Achir, and K. Boussetta, "Wlan planning: Separate and joint optimization of both access point placement and channel assignment," *Ann. Telecommun*, vol. 70, no. 5-6, pp. 263–274, 2015.

[61] "Network Functions Virtualisation (NFV): Use cases," ETSI Group Specification, October 2013. [Online]. Available: http://www.etsi.org/deliver/etsi_gs/nfv/001_099/001/01.01.01_60/gs_nfv001v010101p.pdf

[62] M. Bernaschi, F. Cacace, G. Iannello, M. Vellucci, and L. Vollero, "Opencapwap: An open source capwap implementation for the management and configuration of wifi hot-spots," *Computer Networks*, vol. 53, no. 2, pp. 217–230, 2009.

[63] OpenCAPWAP code source. [Online]. Available: https://github.com/vollero/openCAPWAP/tree/elena.ago

[64] P. Calhoun, M. Montemurro, and D. Stanley, "Control and provisioning of wireless access points (CAPWAP) protocol binding for IEEE 802.11," Internet Requests for Comments, RFC 5416, March 2009.

[65] pfSense. [Online]. Available: https://www.pfsense.org/

[66] Hostapd. [Online]. Available: http://w1.fi/hostapd/

[67] J. Sahoo, S. Mohapatra, and R. Lath, "Virtualization: A survey on concepts, taxonomy and associated security issues," *2010 Second International Conference on Computer and Network Technology (ICCNT)*, pp. 222–226, 2010.

[68] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *Journal of Internet Services and Applications*, vol. 1, no. 1, pp. 7–18, 2010.

[69] B. P. Rimal, E. Choi, and I. Lumb, "A taxonomy and survey of cloud computing systems," in *Fifth International Joint Conference on INC, IMS and IDC (NCM'09)*. IEEE, 2009, pp. 44–51.

[70] A. J. Ferrer, F. HernáNdez, J. Tordsson, E. Elmroth, A. Ali-Eldin, C. Zsigri, R. Sirvent, J. Guitart, R. M. Badia, K. Djemame *et al.*, "OPTIMIS: A holistic approach to cloud service provisioning," *Future Generation Computer Systems*, vol. 28, no. 1, pp. 66–77, 2012.

[71] M. Silvano and T. Paolo, *Knapsack problems: algorithms and computer implementations*. John Wiley and Sons New York, 1990.

[72] A. Verma, P. Ahuja, and A. Neogi, "pMapper: power and migration cost aware application placement in virtualized systems," in *Middleware 2008*. Springer, 2008, pp. 243–264.

[73] N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing sla violations," in *2017 10th IFIP/IEEE International Symposium on Integrated Network Management*. IEEE, 2007, pp. 119–128.

[74] J. Dong, X. Jin, H. Wang, Y. Li, P. Zhang, and S. Cheng, "Energy-saving virtual machine placement in cloud data centers," in *2013 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*. IEEE, 2013, pp. 618–624.

[75] J. Levine and F. Ducatelle, "Ant colony optimization and local search for bin packing and cutting stock problems," *Journal of the Operational Research Society*, vol. 55, no. 7, pp. 705–716, 2004.

[76] B. Li, J. Li, J. Huai, T. Wo, Q. Li, and L. Zhong, "Enacloud: An energy-saving application live placement approach for cloud computing environments," in *2009 IEEE International Conference on Cloud Computing (CLOUD'09)*. IEEE, 2009, pp. 17–24.

[77] B. Brugger, K. F. Doerner, R. F. Hartl, and M. Reimann, "Antpacking–an ant colony optimization approach for the one-dimensional bin packing problem," in *Evolutionary Computation in Combinatorial Optimization*. Springer, 2004, pp. 41–50.

[78] A. Beloglazov and R. Buyya, "Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers," in *Proceedings of the 8th International Workshop on Middleware for Grids, Clouds and e-Science*. ACM, 2010, p. 4.

[79] M. Y. Lim, F. Rawson, T. Bletsch, and V. W. Freeh, "Padd: Power aware domain distribution," in *2009 29th IEEE International Conference on Distributed Computing Systems (ICDCS'09)*. IEEE, 2009, pp. 239–247.

[80] A. Khosravi, S. K. Garg, and R. Buyya, "Energy and carbon-efficient placement of virtual machines in distributed cloud data centers," in *European Conference on Parallel Processing*. Springer, 2013, pp. 317–328.

[81] L. T. Kou and G. Markowsky, "Multidimensional bin packing algorithms," *IBM Journal of Research and development*, vol. 21, no. 5, pp. 443–448, 1977.

[82] C. Chekuri and S. Khanna, "On Multi-Dimensional Packing Problems." in *SODA*, vol. 99. Citeseer, 1999, pp. 185–194.

[83] E. Feller, L. Rilling, and C. Morin, "Energy-aware ant colony based workload placement in clouds," in *Proceedings of the 2011 IEEE/ACM 12th International Conference on Grid Computing*. IEEE Computer Society, 2011, pp. 26–33.

[84] M. Stillwell, D. Schanzenbach, F. Vivien, and H. Casanova, "Resource allocation algorithms for virtualized service hosting platforms," *Journal of Parallel and Distributed Computing*, vol. 70, no. 9, pp. 962–974, 2010.

[85] S. Srikantaiah, A. Kansal, and F. Zhao, "Energy aware consolidation for cloud computing," in *Proceedings of the 2008 conference on Power aware computing and systems*, vol. 10. San Diego, California, 2008.

[86] Y. Gao, H. Guan, Z. Qi, Y. Hou, and L. Liu, "A multi-objective ant colony system algorithm for virtual machine placement in cloud computing," *Journal of Computer and System Sciences*, vol. 79, no. 8, pp. 1230–1242, 2013.

[87] T. C. Ferreto, M. A. Netto, R. N. Calheiros, and C. A. De Rose, "Server consolidation with migration control for virtualized data centers," *Future Generation Computer Systems*, vol. 27, no. 8, pp. 1027–1034, 2011.

[88] G. Khanna, K. Beaty, G. Kar, and A. Kochut, "Application performance management in virtualized server environments," in *10th IEEE/IFIP Network Operations and Management Symposium (NOMS 2006)*. IEEE, 2006, pp. 373–381.

[89] U. Bellur and C. S. Rao, "Optimal placement algorithms for virtual machines," 2010. [Online]. Available: http://arxiv.org/pdf/1011.5064v1.pdf

[90] D. Ye and J. Chen, "Non-cooperative games on multidimensional resource allocation," *Future Generation Computer Systems*, vol. 29, no. 6, pp. 1345–1352, 2013.

[91] J. Tordsson, R. S. Montero, R. Moreno-Vozmediano, and I. M. Llorente, "Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers," *Future Generation Computer Systems*, vol. 28, no. 2, pp. 358–367, 2012.

[92] W. Shi and B. Hong, "Towards profitable virtual machine placement in the data center," in *2011 Fourth IEEE International Conference on Utility and Cloud Computing (UCC)*. IEEE, 2011, pp. 138–145.

[93] J. A. Aroca, A. F. Anta, M. A. Mosteiro, C. Thraves, and L. Wang, "Power-efficient Assignment of Virtual Machines to Physical Machines," in *Adaptive Resource Management and Scheduling for Cloud Computing*. Springer, 2014, pp. 71–88.

[94] D. Breitgand and A. Epstein, "SLA-aware placement of multi-virtual machine elastic services in compute clouds," in *2011 IFIP/IEEE International Symposium on Integrated Network Management (IM)*. IEEE, 2011, pp. 161–168.

[95] H. Goudarzi and M. Pedram, "Energy-efficient virtual machine replication and placement in a cloud computing system," in *2012 IEEE 5th International Conference on Cloud Computing (CLOUD)*. IEEE, 2012, pp. 750–757.

[96] D. Breitgand, A. Marashini, and J. Tordsson, "Policy-driven service placement optimization in federated clouds," *IBM Research Division, Tech. Rep*, 2011.

[97] R. E. Burkard, *Quadratic assignment problems*. Springer, 2013.

[98] E. M. Loiola, N. M. M. de Abreu, P. O. Boaventura-Netto, P. Hahn, and T. Querido, "A survey for the quadratic assignment problem," *European Journal of Operational Research*, vol. 176, no. 2, pp. 657–690, 2007.

[99] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," in *2010 Proceedings IEEE INFOCOM*. IEEE, 2010, pp. 1–9.

[100] D. M. Freimuth, X. Meng, V. Pappas, and L. Zhang, "Placement of virtual machines based on server cost and network cost," U.S. Patent US8 478 878 B2, Jul. 2, 2013.

[101] O. Biran, A. Corradi, M. Fanelli, L. Foschini, A. Nus, D. Raz, and E. Silvera, "A stable network-aware vm placement for cloud systems," in *Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2012)*. IEEE Computer Society, 2012, pp. 498–506.

[102] W. Fang, X. Liang, S. Li, L. Chiaraviglio, and N. Xiong, "VMPlanner: Optimizing virtual machine placement and traffic flow routing to reduce network power costs in cloud data centers," *Computer Networks*, vol. 57, no. 1, pp. 179–196, 2013.

[103] J.-k. DONG, H.-b. WANG, Y.-y. LI, and S.-d. CHENG, "Virtual machine placement optimizing to improve network performance in cloud data centers," *The Journal of China Universities of Posts and Telecommunications*, vol. 21, no. 3, pp. 62–70, 2014.

[104] V. Mann, A. Kumar, P. Dutta, and S. Kalyanaraman, "VMFlow: leveraging VM mobility to reduce network power costs in data centers," in *NETWORKING 2011*. Springer, 2011, pp. 198–211.

[105] D. Jayasinghe, C. Pu, T. Eilam, M. Steinder, I. Whally, and E. Snible, "Improving performance and availability of services hosted on iaas clouds with structural constraint-aware virtual machine placement," in *2011 IEEE International Conference on Services Computing (SCC)*. IEEE, 2011, pp. 72–79.

[106] N. Bansal, K.-W. Lee, V. Nagarajan, and M. Zafer, "Minimum congestion mapping in a cloud," in *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing*. ACM, 2011, pp. 267–276.

[107] Q. Zhang, Q. Zhu, M. F. Zhani, R. Boutaba, and J. L. Hellerstein, "Dynamic service placement in geographically distributed clouds," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 12, pp. 762–772, 2013.

[108] I. Bedhiaf, R. Ali, and O. Cherkaoui, "On the problem of mapping virtual machines to physical machines for delay sensitive services," in *2012 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2012, pp. 2628–2633.

[109] V. Mann, A. Kumar, P. Dutta, and S. Kalyanaraman, "VMFlow: leveraging VM mobility to reduce network power costs in data centers," in *International Conference on Research in Networking*. Springer, 2011, pp. 198–211.

[110] M. Alicherry and T. Lakshman, "Network aware resource allocation in distributed clouds," in *2012 proceedings IEEE Infocom*. IEEE, 2012, pp. 963–971.

[111] W. Fang, X. Liang, S. Li, L. Chiaraviglio, and N. Xiong, "Vmplanner: Optimizing virtual machine placement and traffic flow routing to reduce network power costs in cloud data centers," *Computer Networks*, vol. 57, no. 1, pp. 179–196, 2013.

[112] J. Chen, K. Chiew, D. Ye, L. Zhu, and W. Chen, "Aaga: Affinity-aware grouping for allocation of virtual machines," in *2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA)*. IEEE, 2013, pp. 235–242.

[113] W. Li, J. Tordsson, and E. Elmroth, "Modeling for dynamic cloud scheduling via migration of virtual machines," in *2011 IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE, 2011, pp. 163–171.

[114] M. Stillwell, D. Schanzenbach, F. Vivien, and H. Casanova, "Resource allocation algorithms for virtualized service hosting platforms," *Journal of Parallel and distributed Computing*, vol. 70, no. 9, pp. 962–974, 2010.

[115] N. Bansal, K.-W. Lee, V. Nagarajan, and M. Zafer, "Minimum congestion mapping in a cloud," in *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing*. ACM, 2011, pp. 267–276.

[116] D. Breitgand, A. Marashini, and J. Tordsson, "Policy-driven service placement optimization in federated clouds," *IBM Research Division, Tech. Rep*, vol. 9, pp. 11–15, 2011.

[117] T. C. Ferreto, M. A. Netto, R. N. Calheiros, and C. A. De Rose, "Server consolidation with migration control for virtualized data centers," *Future Generation Computer Systems*, vol. 27, no. 8, pp. 1027–1034, 2011.

[118] Q. Zhang, Q. Zhu, M. F. Zhani, R. Boutaba, and J. L. Hellerstein, "Dynamic service placement in geographically distributed clouds," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 12, pp. 762–772, 2013.

[119] D. Ye and J. Chen, "Non-cooperative games on multidimensional resource allocation," *Future Generation Computer Systems*, vol. 29, no. 6, pp. 1345–1352, 2013.

[120] L. Mashayekhy, M. M. Nejad, and D. Grosu, "Cloud federations in the sky: Formation game and mechanism," *IEEE Transactions on Cloud Computing*, vol. 3, no. 1, pp. 14–27, 2015.

[121] M. Guazzone, C. Anglano, and M. Sereno, "A game-theoretic approach to coalition formation in green cloud federations," in *2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*. IEEE, 2014, pp. 618–625.

[122] X. Xu, H. Yu, and X. Cong, "A qos-constrained resource allocation game in federated cloud," in *2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*. IEEE, 2013, pp. 268–275.

[123] L. Mashayekhy and D. Grosu, "A coalitional game-based mechanism for forming cloud federations," in *2012 IEEE Fifth International Conference on Utility and Cloud Computing (UCC)*. IEEE, 2012, pp. 223–227.

[124] M. M. Hassan, B. Song, and E.-N. Huh, "Distributed resource allocation games in horizontal dynamic cloud federation platform," in *2011 IEEE 13th International Conference on High Performance Computing and Communications (HPCC)*. IEEE, 2011, pp. 822–827.

[125] A. Verma, P. Ahuja, and A. Neogi, "pmapper: power and migration cost aware application placement in virtualized systems," in *ACM/IFIP/USENIX International Conference on Distributed Systems Platforms and Open Distributed Processing*. Springer, 2008, pp. 243–264.

[126] W. Iqbal, M. N. Dailey, and D. Carrera, "Sla-driven dynamic resource management for multi-tier web applications in a cloud," in *2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid)*. IEEE, 2010, pp. 832–837.

[127] J. Tordsson, R. S. Montero, R. Moreno-Vozmediano, and I. M. Llorente, "Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers," *Future Generation Computer Systems*, vol. 28, no. 2, pp. 358–367, 2012.

[128] E. Feller, L. Rilling, and C. Morin, "Energy-aware ant colony based workload placement in clouds," in *Proceedings of the 2011 IEEE/ACM 12th International Conference on Grid Computing*. IEEE Computer Society, 2011, pp. 26–33.

[129] Y. Gao, H. Guan, Z. Qi, Y. Hou, and L. Liu, "A multi-objective ant colony system algorithm for virtual machine placement in cloud computing," *Journal of Computer and System Sciences*, vol. 79, no. 8, pp. 1230–1242, 2013.

[130] A. Beloglazov and R. Buyya, "Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers," in *Proceedings of the 8th International Workshop on Middleware for Grids, Clouds and e-Science*, vol. 4. ACM, 2010.

[131] J. A. Aroca, A. F. Anta, M. A. Mosteiro, C. Thraves, and L. Wang, "Power-efficient assignment of virtual machines to physical machines," *Future Generation Computer Systems*, vol. 54, pp. 82–94, 2016.

[132] H. Goudarzi and M. Pedram, "Energy-efficient virtual machine replication and placement in a cloud computing system," in *2012 IEEE 5th International Conference on Cloud Computing (CLOUD)*. IEEE, 2012, pp. 750–757.

[133] H. N. Van, F. D. Tran, and J.-M. Menaud, "Performance and power management for cloud infrastructures," in *2010 IEEE 3rd international Conference on Cloud Computing*. IEEE, 2010, pp. 329–336.

[134] Y. Li, Y. Wang, B. Yin, and L. Guan, "An energy efficient resource management method in virtualized cloud environment," in *Network Operations and Management Symposium (APNOMS), 2012 14th Asia-Pacific*. IEEE, 2012, pp. 1–8.

[135] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future generation computer systems*, vol. 28, no. 5, pp. 755–768, 2012.

[136] M. Sharifi, H. Salimi, and M. Najafzadeh, "Power-efficient distributed scheduling of virtual machines using workload-aware consolidation techniques," *The Journal of Supercomputing*, vol. 61, no. 1, pp. 46–66, 2012.

[137] J. Dong, X. Jin, H. Wang, Y. Li, P. Zhang, and S. Cheng, "Energy-saving virtual machine placement in cloud data centers," in *13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*. IEEE, 2013, pp. 618–624.

[138] B. Wadhwa and A. Verma, "Energy and carbon efficient vm placement and migration technique for green cloud datacenters," in *2014 Seventh International Conference on Contemporary Computing (IC3)*. IEEE, 2014, pp. 189–193.

[139] R. Ranjana and J. Raja, "A survey on power aware virtual machine placement strategies in a cloud data center," in *2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE)*. IEEE, 2013, pp. 747–752.

[140] J. Sekhar, G. Jeba, and S. Durga, "A survey on energy efficient server consolidation through vm live migration," *International Journal of Advances in Engineering & Technology*, vol. 5, no. 1, pp. 515–525, 2012.

[141] W. Voorsluys, J. Broberg, S. Venugopal, and R. Buyya, "Cost of virtual machine live migration in clouds: A performance evaluation," in *IEEE International Conference on Cloud Computing*. Springer, 2009, pp. 254–265.

[142] G. Khanna, K. Beaty, G. Kar, and A. Kochut, "Application performance management in virtualized server environments," in *2006 IEEE/IFIP Network Operations and Management Symposium NOMS 2006*. IEEE, 2006, pp. 373–381.

[143] A. Amokrane, M. F. Zhani, R. Langar, R. Boutaba, and G. Pujolle, "Greenhead: Virtual data center embedding across distributed infrastructures," *IEEE Transactions on Cloud Computing*, vol. 1, no. 1, pp. 36–49, 2013.

[144] D. Jayasinghe, C. Pu, T. Eilam, M. Steinder, I. Whally, and E. Snible, "Improving performance and availability of services hosted on iaas clouds with structural constraint-aware virtual machine placement," in *2011 IEEE International Conference on Services Computing (SCC)*. IEEE, 2011, pp. 72–79.

[145] W. Shi and B. Hong, "Towards profitable virtual machine placement in the data center," in *2011 Fourth IEEE International Conference on Utility and Cloud Computing (UCC)*. IEEE, 2011, pp. 138–145.

[146] M. R. Gary and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*. WH Freeman and Company, New York, 1979.

[147] S. Sahni and T. Gonzalez, "P-complete approximation problems," *Journal of the ACM (JACM)*, vol. 23, no. 3, pp. 555–565, 1976.

[148] J. Puchinger and G. R. Raidl, "Combining metaheuristics and exact algorithms in combinatorial optimization: A survey and classification," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 2005, pp. 41–53.

[149] Free Software Foundation, Inc. GNU Linear Programming Kit. [Online]. Available: https://www.gnu.org/software/glpk/

[150] IBM. CPLEX Optimizer. [Online]. Available: https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/

[151] Gurobi Optimization, Inc. Gurobi . [Online]. Available: http://www.gurobi.com/

[152] L. T. Kou and G. Markowsky, "Multidimensional bin packing algorithms," *IBM Journal of Research and development*, vol. 21, no. 5, pp. 443–448, 1977.

[153] I. H. Osman and G. Laporte, "Metaheuristics: A bibliography," *Annals of Operations research*, vol. 63, no. 5, pp. 511–623, 1996.

[154] J. Xu and J. A. Fortes, "Multi-objective virtual machine placement in virtualized data center environments," in *2010 IEEE/ACM Int'l Conference on Green Computing and Communications (GreenCom) & Int'l Conference on Cyber, Physical and Social Computing (CPSCom)*. IEEE, 2010, pp. 179–188.

[155] S. Mehraghdam, M. Keller, and H. Karl, "Specifying and placing chains of virtual network functions," in *2014 IEEE 3rd International Conference on Cloud Networking (CloudNet)*. IEEE, 2014, pp. 7–13.

[156] S. Son, G. Jung, and S. C. Jun, "An SLA-based cloud computing that facilitates resource allocation in the distributed data centers of a cloud provider," *Journal of Supercomputing*, vol. 64, no. 2, pp. 606–637, 2013.

[157] J. Tordsson, R. S. Montero, R. Moreno-Vozmediano, and I. M. Llorente, "Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers," *Future Generation Computer Systems*, vol. 28, no. 2, pp. 358–367, 2012.

[158] Q. Zhang, Q. Zhu, M. F. Zhani, R. Boutaba, and J. L. Hellerstein, "Dynamic service placement in geographically distributed clouds," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 12, pp. 762–772, 2013.

[159] M. Alicherry and T. Lakshman, "Network Aware Resource Allocation in Distributed Clouds," in *2012 Proceedings IEEE INFOCOM*, 2012, pp. 963–971.

[160] R. Van Den Bossche, K. Vanmechelen, and J. Broeckhove, "Cost-optimal scheduling in hybrid IaaS clouds for deadline constrained workloads," *Proceedings - 2010 IEEE 3rd International Conference on Cloud Computing, CLOUD 2010*, pp. 228–235, 2010.

[161] J. Altmann and M. M. Kashef, "Cost model based service placement in federated hybrid clouds," *Future Generation Computer Systems*, vol. 41, pp. 79–90, 2014.

[162] R. N. Calheiros and R. Buyya, "Cost-effective provisioning and scheduling of deadline-constrained applications in hybrid clouds," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7651 LNCS, pp. 171–184, 2012.

[163] M. Malawski, K. Figiela, and J. Nabrzyski, "Cost minimization for computational applications on hybrid cloud infrastructures," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1786–1794, 2013.

[164] X. Qiu, W. L. Yeow, C. Wu, and F. C. M. Lau, "Cost-minimizing preemptive scheduling of mapreduce workloads on hybrid clouds," in *IEEE International Workshop on Quality of Service, IWQoS*, 2013, pp. 213–218.

[165] X. Zuo, G. Zhang, and W. Tan, "Self-adaptive learning pso-based deadline constrained task scheduling for hybrid iaas cloud," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 564–573, 2014.

[166] F. Bari, S. R. Chowdhury, R. Ahmed, and R. Boutaba, "On Orchestrating Virtual Network Functions," in *11th International Conference on Network and Service Management (CNSM)*, no. Section III. IEEE, 2015, pp. 50—-56.

[167] B. Addis, D. Belabed, M. Bouet, and S. Secci, "Virtual Network Functions Placement and Routing Optimization," in *IEEE 4th International Conference on Cloud Networking (CloudNet)*, 2015, pp. 171–177.

[168] A. Baumgartner, V. S. Reddy, and T. Bauschert, "Mobile core network virtualization: A model for combined virtual core network function placement and topology optimization," in *1st IEEE Conference on Network Softwarization (NETSOFT)*. IEEE, 2015, pp. 1—-9.

[169] T. Taleb, M. Bagaa, and A. Ksentini, "User mobility-aware Virtual Network Function placement for Virtual 5G Network Infrastructure," in *IEEE International Conference on Communications (ICC)*, 2015, pp. 3879–3884.

[170] R. Riggio, A. Bradai, T. Rasheed, J. Schulz-Zander, S. Kuklinski, and T. Ahmed, "Virtual Network Functions Orchestration in Wireless Networks," in *11th International Conference on Network and Service Management (CNSM)*. IEEE, 2015, pp. 108—-116.

[171] Y. Dong, X. Yang, J. Li, G. Liao, K. Tian, and H. Guan, "High performance network virtualization with sr-iov," *Journal of Parallel and Distributed Computing*, vol. 72, no. 11, pp. 1471–1480, 2012.

[172] E. Gelenbe, G. Pujolle, and J. Nelson, *Introduction to queueing networks*. Wiley Chichester, 1998, vol. 2.

[173] J. Labetoulle, G. Pujolle, and C. Soula, "Stationary distributions of flows in jackson networks," *Mathematics of operations research*, vol. 6, no. 2, pp. 173–185, 1981.

[174] "White paper: Network functions virtualization - challenges and solutions," Alcatel-Lucent, Tech. Rep., June 2013.

[175] G. L. Nemhauser and L. A. Wolsey, "Integer and combinatorial optimization john wiley & sons," *New York*, 1988.

[176] C.-L. Hwang and A. S. M. Masud, *Multiple Objective Decision Making, Methods and Applications: a state-of-the-art survey*. Springer, 1979.

[177] R. T. Marler and J. S. Arora, "The weighted sum method for multi-objective optimization: New insights," *Structural and Multidisciplinary Optimization*, vol. 41, no. 6, pp. 853–862, 2010.

[178] M. Pedram and I. Hwang, "Power and performance modeling in a virtualized server system," *Proceedings of the International Conference on Parallel Processing Workshops*, pp. 520–526, 2010.

[179] "Intel xeon processor e5 v4 family," (Accessed: April 2016). [Online]. Available: http://ark.intel.com/products/family/91287/Intel-Xeon-Processor-E5-v4-Family#@Server

[180] "Intel xeon processor e7 v3 family," (Accessed: April 2016). [Online]. Available: http://ark.intel.com/products/family/78585/Intel-Xeon-Processor-E7-v3-Family#@Server

[181] D. Raumer, F. Wohlfart, D. Scholz, P. Emmerich, and G. Carle, "Performance exploration of software-based packet processing systems," in *Leistungs-, Zuverlässigkeits-und Verlässlichkeitsbewertung von Kommunikationsnetzen und verteilten Systemen, 8. GI/ITG-Workshop MMBnet*, 2015.

[182] J. Löfberg, "Yalmip: A toolbox for modeling and optimization in matlab," in *Computer Aided Control Systems Design, 2004 IEEE International Symposium on*. IEEE, 2004, pp. 284–289.

[183] OpenContrail. [Online]. Available: http://www.opencontrail.org/

[184] OpenStack. [Online]. Available: https://www.openstack.org/

[185] "Software Defined Networking (SDN) and Network Function Virtualisation (NFV) for Wi-Fi network ," White Paper, WBA, Jan 2016. [Online]. Available: http://www.wballiance.com/resources/wba-white-papers/

Appendix A

The NGH Trial

The Next Generation Hotspot (NGH) trial is a WBA project launched since 2011 and conducted through four phases. It aims to deliver a next generation Wi-Fi experience based on a seamless and secure access to public Wi-Fi access points on home and roaming partners' networks around the world. Furthermore, it helps drive the industry toward adoption of NGH networks. The trial involves end-to-end inter-operator testing of the NGH requirements in a "real world" production environment. The trial is driven by a strong participation of a large community of industry players including mobile and Wi-Fi operators (e.g., AT&T, BT, NTT Docomo, Boingo, Fon, Meteor Network), infrastructure vendors (e.g., Cisco, Ruckus, Huawei, Ericson), device vendors (e.g., Intel, Samsung, Mediatek), and hub providers (BSG Wireless and Syniverse Technologies).

The main objective of this trial is to test the key functionalities introduced by the new Wi-Fi Alliance Hotspot 2.0 technology [4, 5] that aims to simplify, automate and secure access to public Wi-Fi networks and bring cellular-like end-user experience to Wi-Fi authentication and roaming. The industry certification program of Hotspot 2.0 technical specification launched by the Wi-Fi Alliance is known as Passpoint. The major features defined by Passpoint certification include:

- *Improved user experience*: Wi-Fi devices discover and select Passpoint-enabled networks in the background, without any active user intervention. Indeed, devices are authenticated automatically using Extensible Authentication Protocols (EAP) based on credentials stored beforehand on the device such as Subscriber Identity Module (SIM), a username and password, or certificate credentials.

- *Secure access*: All connections are secured with WPA2-Enterprise, which provides a level of security comparable to that of cellular networks.

- *Immediate account provisioning (on-line signup)*: A new user is allowed to have an immediate account at the point of access using a common provisioning methodology across vendors.

- *Operator policy*: Passpoint enables mechanisms to support operator-specific subscriber policies such as network selection and handover between Wi-Fi network and 3G/4G network policies.

Besides these features, the NGH trial also aims to test roaming capabilities and interoperability across networks.

Meteor Network is involved in this project as a Wi-Fi operator providing Passpoint-enabled network using Ruckus infrastructure (i.e., access points and Wi-Fi controller) interconnected to AT&T network, the roaming partner, through BSG wireless hubs. To carry out the tests, we are using a Samsung G900-F device with an experimental firmware supporting Passpoint and an AT&T SIM card. In this trial, Meteor Network represents the visited network operator while AT&T represents the home network provider (See Figure A.1).

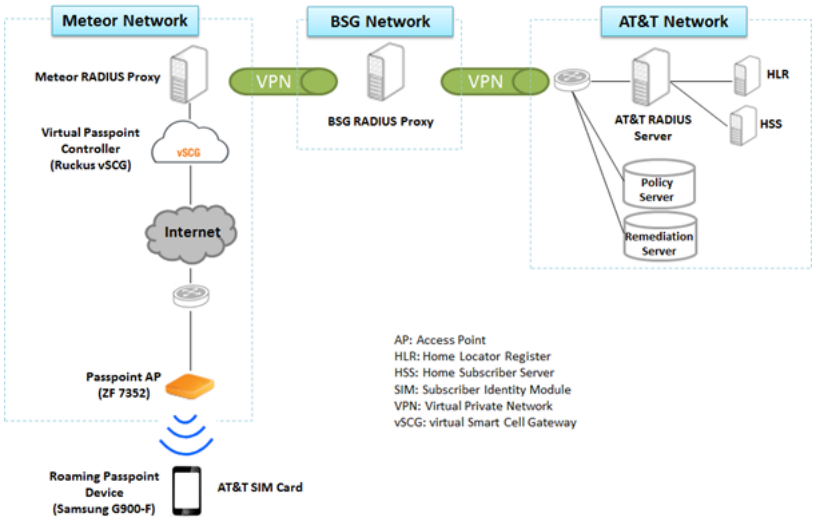


Figure A.1: Test Architecture

As a particular use case, we aim to validate automatic roaming to hotspots of visited operator using SIM credentials. For this purpose, a number of test cases are performed addressing different aspects:

- Network selection
- EAP-SIM authentication
- Radius Authentication, Authorization and Accounting (AAA)

Acronyms

Notation	Description
AAA	Authentication, Authorization and Accounting
AC	Access Controller
ACO	Ant Colony Optimization
ANDSF	Access Network Discovery and Selection Function
ANQP	Access Network Query Protocol
AP	Access Point
ATIS	Alliance for Telecommunications Industry Solutions
BF	Best-Fit
BFD	Best-Fit Decreasing
BPP	Bin-Packing Problem
BSS	Business Support Systems
CDN	Content Delivery Network
COP	Combinatorial Optimization Problem
COTS	Commercial Off The Shelf
CPU	Central Processing Unit
DNS-SD	DNS-based Service Discovery
DPI	Deep Packet Inspection
ETSI	European Telecommunications Standards Institute
FF	First-Fit
FFD	First-Fit Decreasing
FMCU	Fixed Maximum Cloudlet Utilization level

Notation	Description
FQVT	Fixed QoS Violation Threshold
GA	Genetic Algorithm
GAP	Generalized Assignment Problem
GAS	Generic Advertisement Service
GI	General Independent
GLPK	GNU Linear Programming Kit
GQAP	Generalized Quadratic Assignment Problem
HetNets	Heterogeneous Networks
HEW	High Efficiency WLAN
HSS	Home Subscriber Server
HTTP	Hypertext Transfer Protocol
I/O	Input/Output
ID	Identifier
IE	Information Element
IETF	Internet Engineering Task Force
IMS	IP Multimedia Subsystem
IoT	Internet of Things
IP	Infrastructure Provider
ISG	Industry Specification Group
ISP	Internet Service Provider
IT	Information Technology
LoST	Location-to-Service Translation
LP	Linear Programming
MANO	Management and Orchestration
MCFP	Multi-Commodity Flow Problem
MD-BPP	Multi-Dimension Bin Packing Problem
MILP	Mixed Integer Linear Program
MIMO	Multiple Input Multiple Output
MME	Mobility Management Entity
MODM	Multiple Objective Decision Making
MR-GAP	Multi-Resource Generalized Assignment Problem
MSAP	Mobility Services Advertisement Protocol
NFV	Network Function Virtualisation
NFVI	NFV Infrastructure
NFVRG	NFV Research Group
NGH	Next Generation Hotspot

Notation	**Description**
NP	Non-deterministic Polynomial-time
OFDMA	Orthogonal Frequency Division Multiple Access
OPNFV	Open Network Function Virtualization
OSS	Operations Support Systems
OTT	Over-The-Top
PBF	Processor Base Frequency
PDN-GW	Packet Data Network Gateway
PM	Physical Machine
QAP	Quadratic Assignment Problem
QoS	Quality of Service
RAN	Radio Access Network
RF	Radio Frequency
RTT	Round Trip Time
S-GW	Serving Gateway
SD	Service Discovery
SDN	Software Defined Networking
SDP	Service Discovery Protocol
SIP	Session Initiation Protocol
SLA	Service Level Agreement
SLP	Service Location Protocol
SoftAP	Software AP
SP	Service Provider
SSDP	Simple Service Discovery Protocol
TCP	Transmission Control Protocol
To-CUQV	Trade-off between Cloudlet Utilization and QoS Violation
UDP	User Datagram Protocol
UE	User Equipment
UPnP	Universal Plug and Play
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
UUID	Universally Unique Identifier
VBPP	Vector Bin Packing Problem
VBS	Venue Based Service

Notation	**Description**
VM	Virtual Machine
VMM	Virtual Machine Monitor
VMP	Virtual Machine Placement
VNF	Virtualized Network Function
VNF FG	Virtual Network Function Forwarding Graph
VNI	Visual Networking Index
VoIP	Voice over IP
VoLTE	Voice over LTE
VoWiFi	Voice over Wi-Fi
WBA	Wireless Broadband Alliance
WFA	Wi-Fi Alliance
Wi-Fi	Wireless Fidelity
WLAN	Wireless Local Area Network
WTP	Wireless Termination Point
XML	eXtensible Markup Language