FEDERAL UNIVERSITY OF PARÁ
INSTITUTE OF TECHNOLOGY
POSTGRADUATE PROGRAM IN ELECTRICAL ENGINEERING

LUCAS DE SOUSA PACHECO

MOBILITY AND CLOUD MANAGEMENT IN WIRELESS HETEROGENEOUS 5G NETWORKS

DM: 22/2020

BELÉM

2020

LUCAS DE SOUSA PACHECO

MOBILITY AND CLOUD MANAGEMENT IN WIRELESS HETEROGENEOUS 5G NETWORKS

Dissertation submitted to the Judging Committee at the Federal University of Pará as part of the requirements for obtaining a Master's Degree in Electrical Engineering in the area of Applied Computing.

Advisor: Dr. Eduardo Coelho Cerqueira

Co-advisor: Dr. Denis Lima do Rosário

BELÉM

2020

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"MOBILITY AND CLOUD MANAGEMENT IN WIRELESS HETEROGENEOUS 5G NETWORKS"

AUTOR: **LUCAS DE SOUSA PACHECO**

DISSERTAÇÃO DE MESTRADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM: 30/06/2020

BANCA EXAMINADORA:

Eduardo Cerqueira

Prof. Dr. Eduardo Coelho Cerqueira
(Orientador – PPGEE/UFPA)

Denis Lima do Rosário

Prof. Dr. Denis Lima do Rosário
(Coorientador – PPGEE/UFPA)

Diego Lisboa Cardoso

Prof. Dr. Diego Lisboa Cardoso
(Avaliador Interno – PPGEE/UFPA)

Helder May Nunes da Silva Oliveira

Prof. Dr. Helder May Nunes da Silva Oliveira
(Avaliador Externo ao Programa – ICEN/UFPA)

Thais Lira Tavares dos Santos

Prof.ª Dr.ª Thais Lira Tavares dos Santos
(Avaliadora Externa ao Programa – GERCOM/UFPA)

VISTO:

Prof.ª Dr.ª Maria Emília de Lima Tostes
(Coordenadora do PPGEE/ITEC/UFPA)

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

D278m de Sousa Pacheco, Lucas

MOBILITY AND CLOUD MANAGEMENT IN WIRELESS HETEROGENEOUS 5G NETWORKS / Lucas de Sousa Pacheco. — 2020.

77 f. : il. color.

Orientador(a): Prof. Dr. Eduardo Coelho Cerqueira
Coorientador(a): Prof. Dr. Denis Lima do Rosário

Dissertação (Mestrado) - Programa de Pós-Graduação em Engenharia Elétrica, Instituto de Tecnologia, Universidade Federal do Pará, Belém, 2020.

1. Gerenciamento de Mobilidade. 2. Redes Veiculares. 3. Migração de serviço. 4. Handover. I. Título.

CDD 384

I'd like to dedicate this work to my family.

Acknowledgements

Hereby I thank all the people that helped me during these last few years, especially my parents Cilene and José, who made it all possible through much support in many ways. A special thanks to all the rest of my family, in particular, my grandparents for all the love and support. I'd also like to express my gratitude to professors Eduardo Cerqueira and Denis Rosário, who challenged me to improve and learn more, while also believing in my potential. Love and affection are all I have to express for my friends, who helped me much during these years.

Since the beginning of this course, many great people have earned my deep consideration and respect, like my lab co-workers and many others. Although many difficulties were faced on this trajectory, all of you have made it all worth it. Thanks to all the excellent teachers that have coached me as well on a search for knowledge and self-improvement. Also, Dinah and Bilbo, who are the loveliest (sometimes) pets I've ever met. Nevertheless, to all the people that I haven't mentioned, my sincere thanks as well.

"I'm being quoted to introduce something, but I have no idea what it is and certainly don't endorse it."
Randall Munroe

Resumo

Resumo da Dissertação apresentada à UFPA como parte dos requisitos necessários para obtenção do grau de Mestre em Engenharia Elétrica.

MOBILITY AND CLOUD MANAGEMENT IN WIRELESS HETEROGENEOUS 5G NETWORKS

Orientador: Eduardo Coelho Cerqueira
Coorientador: Denis Lima do Rosário
Palavras-Chave: Gerenciamento de Mobilidade; Redes Veiculares; Qualidade da experiência; QoE; Migração de serviço.

O ramo de gerência de mobilidade de redes é responsável pelos protocolos e ações tomadas pela rede para garantircontinuidade dos serviços consumidos por usuários móveis. Nesta dissertação é analisado como as redes de próxima geração abrirão caminho para a distribuição de vídeo em redes veiculares (VANETs), compostas por uma infraestrutura heterogênea ultradensa, unindo tecnologias de comunicação sem fio existentes para obter maior eficiência espectral. É apresentado um algoritmo de *handover* chamado HoVe. Baseado em vários critérios para distribuição de vídeo em VANETs 5G ultradensas. Resultados de simulação mostram a eficiência do HoVe em fornecer vídeos com qualidade 19% superior a algoritmos do estado-da-arte, melhorando a taxa de entrega de pacotes em pelo menos 30%. Este trabalho estuda um caso particular de VANETs que se beneficia da computação na borda da rede, o caso de Veículos Autonômos Conectados, ou CAVs. A computação de borda e em névoa são soluções emergentes para processamento remoto de dados para veículos autônomos. Este trabalho propõe o algoritmo MOSAIC para migração de serviço e gerenciamento de recursos para comunicação entre camadas e entre camadas na computação de borda e em névoa. Resultados da simulação mostram a eficiência do algoritmo proposto com melhor desempenho de ate 50% em termos de latência e cinco vezes menos falhas de migração.

Abstract

Abstract of Dissertation presented to UFPA as a partial fulfillment of the requirements for the degree of Master in Electrical Engineering.

MOBILITY AND CLOUD MANAGEMENT IN WIRELESS HETEROGENEOUS 5G NETWORKS

Advisor: Eduardo Coelho Cerqueira
Co-advisor: Denis Lima do Rosário
Key words: Mobility Management; Vehicular Networks; Handover; Quality of Experience; QoE; Service Migration.

The network mobility management branch is responsible for the protocols and actions taken by the network to ensure connectivity and the continuity of services consumed by mobile users. In this dissertation we analyse how next-generation networks pave the way for the distribution of video in vehicular networks (VANETs), composed by an heterogeneous ultra-dense infrastructure, joining existing wireless communication technologies to obtain greater spectral efficiency. A handover algorithm called HoVe is presented. Based on various criteria for video distribution on ultra-dense 5G VANETs. The simulation results show HoVe's efficiency in providing videos with 19% higher quality than state-of-the-art algorithms, improving the package delivery rate by at least 30%. This work studies a particular case of VANETs that benefits from computing at the edge of the network, the case of Connected Autonomous Vehicles, or CAVs. Edge and mist computing are emerging solutions for remote data processing for autonomous vehicles, offering greater computational power, as well as the low latency required by autonomous driving. This work proposes the MOSAIC algorithm for service migration and resource management for communication between layers and between layers in edge and fog computing. Simulation results show the efficiency of the proposed algorithm with a better performance of up to 50% in terms of latency and five times less migration failures.

Contents

List of Abbreviations

3GPP 3rd Generation Partnership Project

AHP Analytycal Hierarchical Process

AP Access Point

ARIMA Auto Regressive Integrated Moving Average

CAV Connected Autonomous Vehicle

CI Consistency Index

CPU Central Processing Unit

CR Consustency Ratio

DSRC Dedicated short-range communications

ELECTRE Elimination EtChoix Traduisant la REalité

FAHP Fuzzy Analytycal Hierarchical Process

GA Genetic Algorithm

GB GigaByte

GH GigaHertz

GPS Global Positioning System

IEEE Institute of Electrical and Electronics Engineers

IP Internet Protocol

ITS Intelligent Transportation System

ITU telecommunications and Information Communication Technology

IVC Inter Vehicle Communication

KF Kalman Filter

LCD Liquid Crystal Display

LiDAR Light RADAR

LTE Long Term Evolution

MADM Multiple Attribute Decision Mechanism

MANET Mobile Ad-Hoc Networks

MH MegaHetz

MIMO Multiple Input Multiple Output

MME Mobility Management Entity

MOS Mean Opinion Score

MPEG Moving Picture Experts Group

MSE Mean Squared Error

NC Non-Cooperative

NS Network Simulator

OBU On-board Units

PBGT Power Budget

PC Personal Computer

PDR Packet Delivery Ratio

PGBT Power Budget

PSNR Peak Signal to Noise Ratio

RAM Random Access Memory

RMSE Root Mean Square Error

RSRP Reference Signal Receive Power

RSRQ Reference Signal Received Quality

RSSI Received Signal Strength Indicator

SDN Software Defined Network

SINR Signal to Interference Plus Noise Ratio

SSIM Structural Simmilarity

TMS Traffic Management System

TTT Time-to-Trigger

UDN Ultra-Dense Network

UE User Equipment

URLLC Ultra-Reliable Low-Latency Communication

V2I Vehicle-to-Infrastructure

V2V Vehicle-to-Vehicle

V2X Vehicle-to-Everything

VANET Vehicular Ad-hoc Network

VCM VANET Connection Manager

VECC Vehicular Edge Cloud Computing

VF Video Flow

VM Virtual Machine

VQM Video Quality Measurement

VQMT Video Quality Measurement Tool

WAVE Wireless Access in Vehicular Environments

eNB eNodeB

MME Mobility Management Entity

TOPSIS Technique for Order Preference by Similarity to Ideal Solution

List of Figures

List of Tables

CHAPTER 1

Introduction

This chapter introduces the main concepts and challenges of mobility management in heterogeneous wireless networks. Outlines the main research lines and provides context for the subsequent chapters.

1.1 Mobility Management

The great popularization of mobile services and devices brings with itself the necessity and opportunity of ubiquitous connectivity. Such devices must support mobile and cloud-based services anywhere and at any time. Users expect excellent service provisioning and coverage, and an unsatisfactory user experience translates into a monetary loss for operators and providers. Therefore, protocols and strategies for network management must be carefully designed to support the integration of new technologies and scenarios, such as next-generation networks, ultra-dense scenarios networks, and heterogeneous users and services [8]. Next-generation wireless networks will pave the way for extensive use of high demanding applications such as video-based services for mobile users, anytime and anywhere [48], including real-time distribution of advertisement or entertainment videos over vehicular networks (VANETs). Supporting these multimedia applications will be one of the critical issues for the success of future networks.

Next-generation communications will not only rely on new access technologies, such as Massive MIMO and Millimeter Wave, but they will also take advantage of existing communication infrastructures, such as LTE and WiFi, to provide ubiquitous and efficient communication [59]. In this sense, 5G networks will be composed of denser heterogeneous radio deployment compared to 4G systems, increasing the available throughput at the edge of the network. Denser networks consist of the increased presence if macrocells, microcells, small cells, relays, and other communication solutions per unit area. It serves to achieve both higher spectral efficiency and higher spectrum reuse rates for mobile networks [23, 14, 31]. The deployed smaller cells offload the traffic that would otherwise be entirely directed at macrocells, and enable the communication of all kinds of devices in highly dense, ubiquitous, and heterogeneous environments, having a high impact not only from a business standpoint but also from a social one, bringing people together and bringing enormous benefits to their lives [46].

One of the main challenges presented by current and future networking scenarios is the great mobility associated with the users and the highly heterogeneous nature of such networks [69]. These networks consist of several integrated technologies, sometimes made by different manufacturers and with different protocols. Not only the network equipment are more heterogeneous than ever, but user devices as well. The boundaries of what is considered a user device have broadened dramatically over the past few years, and now includes wearable devices, sensors, smartphones, and even vehicles. This introduces several particularities in the network management, as the mobility patterns of each node may differ significantly, *i.e.* a vehicular device follows a more predictable and trajectory and with a higher velocity than a device with a pedestrian, which is more unpredictable in comparison.

These challenges constitute mobility management problems and must be treated differently in network management. Mobility management consists of the strategies, protocols, and algorithms employed by the network and service providers to ensure that users have a seamless experience even while moving from one network to another. A seamless experience, while being a broad term, is a hugely important factor in defining the user's Quality of Experience (QoE). It can be said that proper mobility management is undetectable to the end-users, as their services continue to be consumed without interruptions. However, each user must be considered carefully. Current networking technologies are often reactive, performing handovers and migrations after the users leave a particular coverage area. To fully guarantee service continuity in real-world scenarios, a predictive approach may achieve significant improvements, as shown in subsequent chapters.

Thus, delivering a seamless experience to increasingly demanding users is a vital network requirement. The mobility management strategies applied in the network have the purpose of performing the necessary operations to guarantee continuity for services and connectivity for each user. This requires a wide range of operations, including handover, session transfers, mobile IP management, service migrations, and many more.

The present dissertation tackles two relevant research problems concerning mobility management in wireless networks; described as follows:

1. In wireless networks, one of the main consequences of user mobility is the number of disconnections and transfer events, namely handovers. Traditionally handovers happen when a user crosses the coverage area from one wireless cell into the other. However, handovers are disrupting events, which may cause packet losses and disconnections. The number of handovers and the patterns in which they happen depends significantly on cell deployment, on the user mobility, and the handover

algorithm employed by the network. More handovers tend to happen if more cells are present in the network or if the user's velocity is high. Thus novel handover strategies may be necessary in modern scenarios, where the objective may not be simply to provide users with the highest SINR at all times but to offer the best possible Quality of Experience (QoE) for end-users.

2. Another challenge in wireless and vehicular networks arises from providing services to users in a ubiquitous manner. In future vehicular scenarios, the presence of Connected Autonomous Vehicles will not only be expected but also play an essential role, leading to a more than 500 billion dollar market by 2026 [25]. Autonomous vehicles rely on the vast amount of sensors which they have equipped. However, the consensus in the literature is that autonomous vehicles must be connected to access additional context information and processing [22]. Offloading computation to remote servers is not a trivial task in this scenario. Only servers geographically close to the vehicles are suitable for the offloading, and the mobility of cars constitutes yet another challenge for the network.

We will now review the main technical challenges concerning these two research problems:

1.2 Research problem #1: Handover in Ultra-Dense Networks

The first of these challenges arise from the denser network deployment trend, also known as network densification. Densification aims to provide users with high SINR while maintaining cell usage low most of the time. In a dense network, cells are deployed in higher numbers and generally with lower transmission power, as not to compromise in interference levels. This means that the ratio of cells per user is higher than in traditional networks, thus achieving higher throughput for each user and relatively good coverage.

One of the primary services in such networks is multimedia content consumption, such as video streaming. Video streaming is already an important market driver, as it is mostly shared from entertainers and marketers, and is set to achieve an even greater relevancy [64]. Such videos must be shared with a decent Quality of Experience (QoE) and Quality of Service (QoS) support to be delivered with satisfactory quality levels for end-users [9]. One of the critical issues for the future and success of video distribution over 5G VANETs is the capacity to support efficient mobility management algorithms. This is a consequence of vehicles moving through different areas, and consequently, switching between different networks [47]. Since vehicles switch networks more than other users, like pedestrians, their mobility management must be optimized, as handovers for vehicular devices are more frequent than pedestrian users, resulting in excessive signaling overhead and reduced performance for these users. However, the highly-dense and heterogeneous nature of 5G networks, while enabling higher data rates, also causes more frequent disconnections for mobile users.

Taking into consideration QoE and QoS parameters in mobility management can be a viable option to achieve a minimum quality required for video transmissions. Still, they may not be enough in dense scenarios [4]. The network can take advantage of the trajectory of vehicular nodes, since it is somewhat predictable. Estimating a vehicle's future geographic position, even in the short term, can significantly improve network decisions [74]. In this sense, Traffic Management Systems (TMSs) can be integrated into the QoE-awareness handover process to improve the decision-making process [21].

5G communications have the mission to be responsive, fast, and power-efficient. It aims to support efficient mobility and resource management schemes to increase the Quality of Experience (QoE) while optimizing the usage of high demanded wireless/radio resources [39]. However, the increased number of heterogeneous cells makes mobility management a challenging task for VANETs, since vehicles, especially in urban environments, frequently switch among different heterogeneous networks, *i.e.,* vehicles travel leaving an area of a cell to enter another one very often [47]. Many handovers result in excessive signaling overhead, disconnection, and ping-pong effect, *i.e.,* a vehicle disconnects from a cell and afterward connects again to another one moments later [7]. These issues increase the packets/video frames losses, leading to a poor QoE for video applications in such a VANET scenario [27].

Skipping unnecessary handovers is beneficial to the network and also to the user's experience [23]. A skipping-based handover consists of avoiding consecutive handovers to maintain the QoE as high as possible. This means reducing the handover frequency by sacrificing some of the best cell connectivity associations [10]. Hence, this allows keeping a longer service duration with the serving cell with at least a minimum service quality level, while reducing signaling overhead and zapping delay. For instance, a handover decision based on Received Signal Strength Indication, RSSI, would benefit to perform a handover every time the received SINR is not the highest possible, thus mitigating the ping-pong effect [23]. Skipping-based handover schemes are often associated with mobility prediction to maximize the connection duration without compromising the network/application performance [11, 6]. This is achieved by giving priority to cells with the highest probability that the user remains connected for more time [45]. However, skipping-based handover schemes alone are not enough to deliver videos with QoE support. A handover decision based on a mobility prediction coupled with QoE and QoS parameters improve video delivery over VANETs by avoiding ping-pong handovers and improving network resources usage [9].

In this dissertation, we propose a multi-criteria skipping-based handover algorithm for video distribution over ultra-dense VANETs, called HoVe. It guarantees seamless handovers in an ultra-dense VANETs scenario to deliver videos with high QoE by taking into account mobility prediction, QoS, QoE, and radio parameters. HoVe supports an Analytic Hierarchy Process (AHP) to assign different degrees of importance for each criterion. HoVe considers proactive Ping-Pong avoidance for handover decision, by skipping handovers when QoE and QoS are acceptable and stable. The implementation of HoVe is available for downloading on Github[1].

We tested two mobility prediction techniques with HoVe, namely AutoRegressive Integrated Moving Average (ARIMA) and Kalman Filter (KF). ARIMA provided a higher accuracy for mobility prediction based on a real-world vehicular dataset analysis compared with KF. Therefore, we chose ARIMA to be considered as a mobility prediction technique used by HoVe. Simulation results showed that the HoVe algorithm delivered videos with QoE 14% better than state-of-the-art algorithms in ultra-dense VANET scenarios. For instance, the Mean Opinion Score (MOS) results showed an improvement of 30% in subjective evaluations, while ping-pong handover was kept at a low 2% rate. The main contributions of this work are summarized as follows: (i) a skipping-based handover algorithm that maximizes connection time to a serving cell; (ii) a multi-criteria decision-making technique for handover decisions in an ultra-dense VANET scenario; and (iii) simulation results to show the performance of HoVe to deliver videos with QoE support in ultra-dense VANET scenarios compared to existing handover algorithms.

1.3 Research problem #2: Connected Autonomous Vehicles and Service Migration

The next problem investigated by this work arises from the extensive research on the autonomous vehicles field. Such a field will constitute a transportation revolution on itself, and the network challenges presented to bring up these concepts are also equally extensive. Mobility is an inherent factor in such environments, where the constituting nodes may yield high velocities.

Connected Autonomous Vehicles, also known as CAVs and autonomous driving technology, will revolutionize transportation systems and bring immeasurable benefits to our society [38, 22]. CAVs extend and rely on the notion of connected vehicles, where vehicles will provide and consume new services, such as infotainment, safety, and offloading from other cars and remote servers [22, 5]. However, CAV is still in a preliminary stage, and there are many challenges to be addressed, mainly due to the high mobility of vehicles and the dynamic usage of 5G radio resources. Therefore, CAV needs new mobility/resource-aware approaches to reduce latency for the applications and improve network/computing resources usage. [5].

CAVs are equipped with a wide range of sensors and actuators that collect a significant amount of data from the environment, extract context information, and perform driving or service decisions. CAVs must process a large amount of collected with very tight latency requirements. Some works even suggest that up to 2GB of collected data must be processed per vehicle every second [37]. This constitutes a great challenge for the local processing of these data since the vehicle's onboard units (OBU) have limited

[1]https://github.com/lsiddd/hove

computation capabilities [73]. This calls for the offloading of these data to remote servers. However, the latency requirements of autonomous driving applications cannot be met with traditional computing paradigms, such as Cloud Computing, to guarantee proper safety for the vehicle [37]. However, the vast amount of data constitutes a challenge for the processing and extraction of context information [73]. The vehicles in the network are expected to generate immense amounts of data, and local processing of all this data to promptly make decisions can be a difficult task.

Edge computing will be a significant part of 5G networks, as mobile users' computing will happen directly at Cells and Access Points. In this sense, edge computing can enhance reliability, perform the latency-sensitive computations, validate, and offload decision-making in CAVs [65]. Edge-enabled environments can offer high bandwidth and low latency, which will be essential in autonomous vehicle scenarios. Being a highly distributed architecture, edge computing can access relevant context information shared between the vehicles (and servers) in their coverage area [40]. However, the computing power of such approaches is still inferior to traditional cloud computing and must be carefully managed.

Since the edge computing paradigm is geographically distributed, the services being consumed are susceptible to the high mobility of vehicles [19]. Traditional mobility management mainly consists of vehicle switching cells as the serving cells' signal strength decreases and will no longer be an efficient solution. As vehicles move to different areas in the city, the services being consumed in an edge architecture can be disrupted, or be located many hops away from the user, reducing the Quality of Service (QoS) of CAV applications. In the era of the autonomous vehicle, an efficient mobility management scheme must be considered to guarantee an acceptable QoS for CAVs [15].

Service migration is an outstanding solution to keep the services as close as possible to the CAV, assuring the minimal QoS requirements for CAV applications. It is also an essential factor in modern mobility management, as we increasingly rely on cloud-based services. Service migration consists of transferring the service running on a virtual machine or container to a new server, or just transferring the user session if the base files for the applications are already present in the target server. Service migration is almost always a mobility-aware task. Services cannot be transferred reliably without mobility information [16]. A straightforward strategy to keep services close to vehicles is to perform migration after every handover to the new nearest edge server. However, this strategy has a poor performance in highly dynamic CAV scenarios, which is expected in 5G and 6G systems. A migration based on QoS, in which servers with poor QoS performance are avoided, may solve some of these issues, but it still may cause service disruptions in mobile environments [33].

Migrations can mitigate the problem of keeping services close to the CAVs; however, transferring virtual machines and containers is a resource-heavy operation, with considerable cost to the network. In this context, one of the strategies employed is the pre-migration of services, where the migration occurs before it becomes necessary, reducing the chance of service disruptions and disconnections. The resources at each edge server should also be taken into account, as migration to servers with few resources can compromise the QoS needed for autonomous vehicle applications and impact the entire vehicle's decisions.

In this context, in this dissertation, we also propose Mobility-based Service Migration in CAV, called MOSAIC. A service migration and resource manager for Vehicular Edge Computing tackles the challenge of offloading the computation from CAVs with the smallest possible latency, ensuring that the vehicle will be served with the necessary resources. MOSAIC considers QoS, resources, and mobility information for decision making, and also MOSAIC uses a Multiple-Criteria Decision-Making algorithm to decide the best possible edge server for each vehicle, and handle the necessary migrations as cars move through the scenario. MOSAIC improves the latency in up to 90% and the throughput in up to 25% compared to state-of-the-art algorithms. Migration failures were non-existent in MOSAIC

1.4 Text Organization

An introduction of the motivations for mobility management research has been presented in this chapter, and subsequent chapters build on the notions laid out to present then the two solutions developed in this dissertation. The remaining of the text is organized as follows:

- Chapter 2 lays out the main concepts used and needed in this work. Going from the basic notions and operations in network mobility management to the handover process in ultra-dense networks, concepts, and functioning of Connected Autonomous Vehicles Scenarios, and finally, service migrations in edge-enabled networks.
- Chapter 3 looks into recent works in the state-of-the-art to find the most relevant ones in the handover, service migration, ultra-dense, and vehicular networks, and edge computing. We hope to show possible directions and opportunities to improve the state-of-the-art.
- Chapter 4 presents the concepts in HoVe, a QoE- and mobility-aware handover algorithm for vehicular scenarios. We fully describe the algorithm functioning and perform an experimental evaluation of the algorithm compared to state-of-the-art works.
- Chapter 5 presents the service migration algorithm MOSAIC. MOSAIC relies on a mobility prediction scheme to perform proactive service migrations in Connected Autonomous Vehicles Scenarios. We also present experimental results to show the efficiency of MOSAIC in realistic network scenarios.
- Chapter 6 summarises the conclusions of this work and highlights the advances to the state-of-the-art in both research problems presented, as well as a list of the publications made in the course of this dissertation.

CHAPTER 2

Basic Concepts

This chapter lays out the theoretical grounds on which this work is constructed. We dive into the elements of heterogeneous wireless networks and how mobile users interact with each other and with the network while consuming services. First, we take a look at how such networks are constructed, and then we analyze how user mobility affects the network from a management point of view. Later on, we discuss the case of edge-enabled networks and Connected Autonomous Vehicles, as well as the challenges imposed in the network to guarantee their functioning.

2.1 Mobility Management

Mobility management refers to the techniques, protocols, and algorithms used to ensure services can always reach mobile users, usually in cellular networks. It is one of the primary functions that allow mobile networks in general to work. Mobility management has been a crucial factor from the very first generations of mobile networks. With the advent of next-generation networks, with dense cell deployments, the techniques once used must be reevaluated and redesigned.

Some of the key operations for mobility management are:

- *Location Update*: this is the primary feature of mobility management. It consists of informing the network when a user moves to another area.
- *Roaming*: it consists of allowing users to start data sessions outside from the operator's coverage area, using a visited operator network. In this work, we only tackle intra-operator mobility management to limit the complexity of the solutions.

- *Initial Cell Selection*: it is the first step in establishing a mobile connection. The User Equipment (UE) must acquire information from the cell about its identifier, signal quality, and other information. Then the mobile node may proceed to synchronize to a cell's frequency and time slots. It is usually performed simply, selecting the cells with the greatest measured SINR.

- *Handover*: refers to the process of a mobile node changing from one cell's network to another. This process may be started by the network by the mobile node and is generally triggered by the relative received signal power. The handover algorithm in operation is crucial to understand various network behaviors and must be carefully crafted. The handover process discussed more in-depth in Section 2.4.

- *Service Migration*: The notion of service migration is rather new in the context of mobile networks, as the services being consumed by users were usually kept in distant centralize locations away from the edge of the network and its users. The handover process was enough to provide service continuity for mobile nodes; however, as services may now be located close to the wireless cells themselves, user mobility raises the necessity of migrating the services to follow user mobility patterns. Discussed in more depth in Section 2.8.

2.2 Vehicular Ad-Hoc Networks

VANETs are a subset of the broader definition of Mobile Ad-Hoc Networks (MANETs). The integration of wireless communication, like WiFi and LTE, into vehicles has raised the interest of the scientific community ever since as long as the 1980s. As users expect to be connected anywhere and at any time, VANETs are a crucial technology to provide services to users in traffic. Some of the most relevant ones can be defined as safety, transport efficiency, and entertainment applications [28].

The market potential of VANETs has long been recognized by stakeholders as well as the possible improvement to the general quality of life for transportation users. The ability to communicate and access services and applications from within vehicles will redefine the ways we live, work, and entertain ourselves. [56]. VANETs can be considered the cornerstone on which Intelligent Transportation Systems will be built. Integrating the ability to communicate with other vehicles and fixed infrastructure vehicles can significantly enhance the lives of drivers and passengers. Their highly dynamic nature constitutes a challenging environment where data has to be transferred with very low latency and at high rates [56].

Common VANET applications are:

- Safety Applications: Several safety applications can benefit the communication between vehicles. Some of the research trends are collision avoidance and warnings, SOS service, emergency response, and support for authorities [58]. Such applications often require timely communications and decision-making, which brings the opportunity to deploy them at the edge of the network.

- Transportation Efficiency: by allowing vehicles to communicate with each other, VANETs allow the existence of intelligent communication systems. Some of the possible applications and benefits from this notion are traffic management, pollution reduction, platooning, lane changing/merging, bird's eye view, and many more [58]. The benefits of a transportation point-of-view are immeasurable!

- Entertainment Applications: Although VANETs were primarily considered for safety applications, in recent years, entertainment purposes have been considered, and user interest in these applications is expected to become a market driver in the future [56]. Users are already able to consume and produce multimedia content from their vehicles. Advertisers can provide location-based suggestions and other use cases. However, all this communication must be made with the help of QoE-aware mechanisms.

2.3 Ultra-Dense Networks

The throughput requirements of modern applications grow at steady rates, as shown by trends and forecasts [26], especially when considering next-generation scenarios. Several bottlenecks must be addressed to keep providing users with sufficient resources, one of which is the capacity of the network edge. Increasing individual cell capacity is an expensive and challenging task from a research and development point of view, so increasing the number of cells, constituting the notion of Heterogeneous Ultra-Dense Networks (H-UDN), is one of the critical enablers for these requirements.

The concept of H-UDN can be defined as a denser deployment of lower power cells and access points to bring them closer to a large number of geographically distributed end-users. The denser deployment is made in contrast to more centralized cell deployments, as found in previous generations of cellular communication. It results in a higher SINR and throughput to end-users, in general. This enables a higher spectral efficiency for the cells and allows a greater frequency reuse level [23].

It is hard to define H-UDN quantitatively Some authors go as far as defining it as networks containing more than 10^3 cells per kilometer squared, with inter-site distances in the order of meters. However, mobile users and their data demands are not uniformly distributed, and the deployment strategies should follow the hot spot patterns of users.

2.3.1 Mobility Management in Ultra-Dense Networks

While the benefits of UDN for stationary users are undeniable, as we have discussed, the more significant presence of small cells is an excellent challenge for mobile users, from a management point of view. As users move through the scenario, they cross the coverage area of many cells, creating overhead as the network must decide when handover events occur, possibly for a significant number of users and way too frequently. To make things worse, each handover event generates several signaling messages which may occupy significant control channel resources.

2.4 The Handover Process

Handover is a critical mechanism in mobile networks. It is the process in which after a user leaves the coverage area of a cell to a new one, the disconnection with the previous cells happens, and a new connection is formed to the new one. A handover is a costly operation for the network since its execution requires a series of operations to transfer the user's session seamlessly to the new cell, and redirect any data flows being consumed seamlessly. Figure 1 depicts the typical messages exchanged between the participants in the handover process to complete a single successful handover. Figure 1 shows the messages typically exchanged in the handover process. The main actors involved in the process are the User Device (UE), the source and target eNodeBs (eNB), the Mobility Management Entity (MME), which oversees this process, and the network gateway. The handover measurement is composed of the Measurement Control request, and the Measurement Report issued by the user, the decision phase of the algorithm is covered by the Handover (HO) Decision step, and the subsequent steps comprise the Handover Execution Phase.

Figure 2 depicts the typical behavior of the coverage areas of wireless cells, represented by red dots. The colored area around each cell represents its coverage area. This pattern is known as a Voronoi Tesselation [13]. Each Voronoi polygon is obtained as the set of points closer to that specific cell than to any other cell. with an increased number of cells the coverage area of each cell becomes very reduced in an area (and with reduced transmission power), making the users traverse through more coverage areas and perform more handovers.

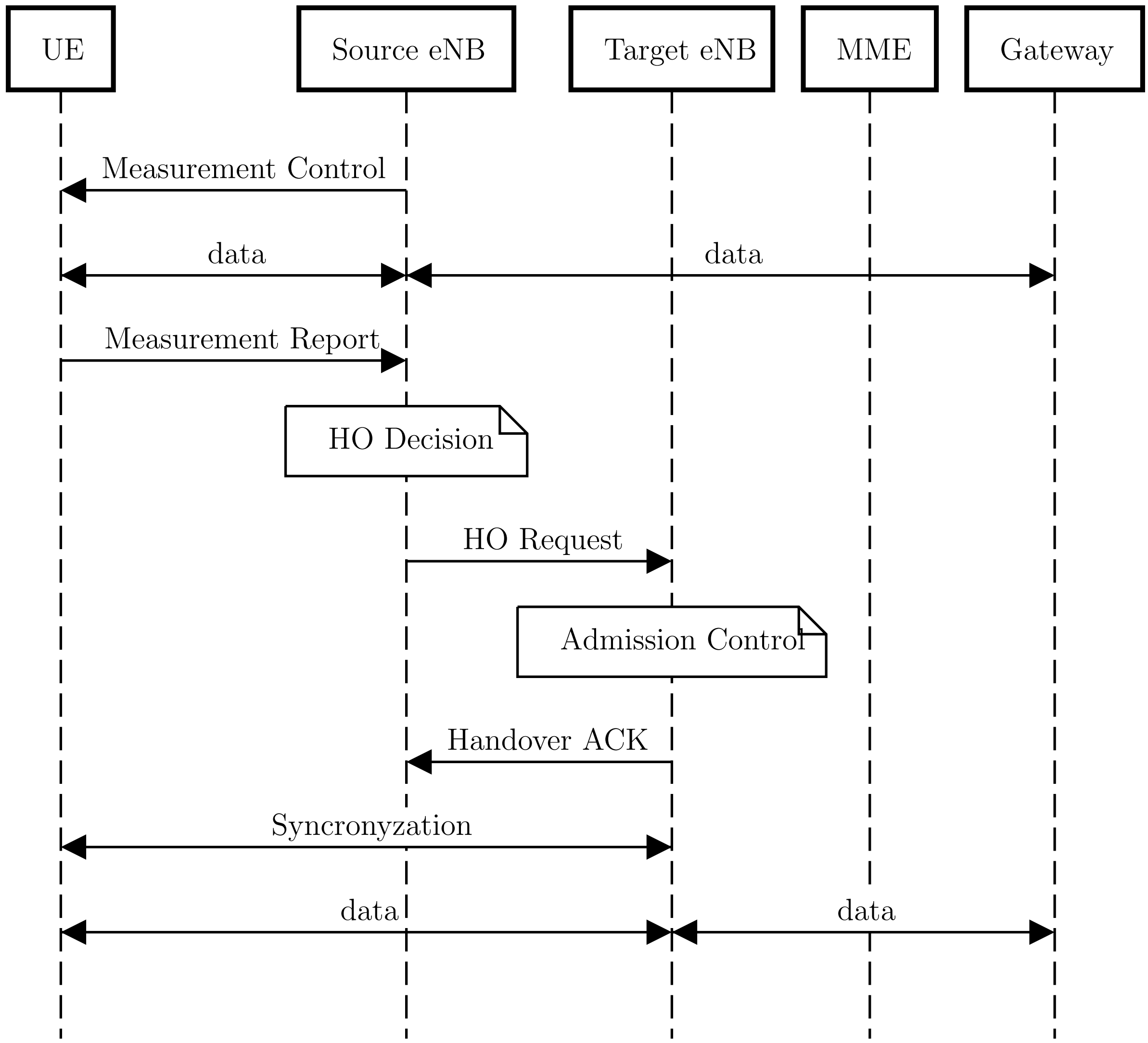


Figure 1: Mobile Network Handover Procedure Sequence

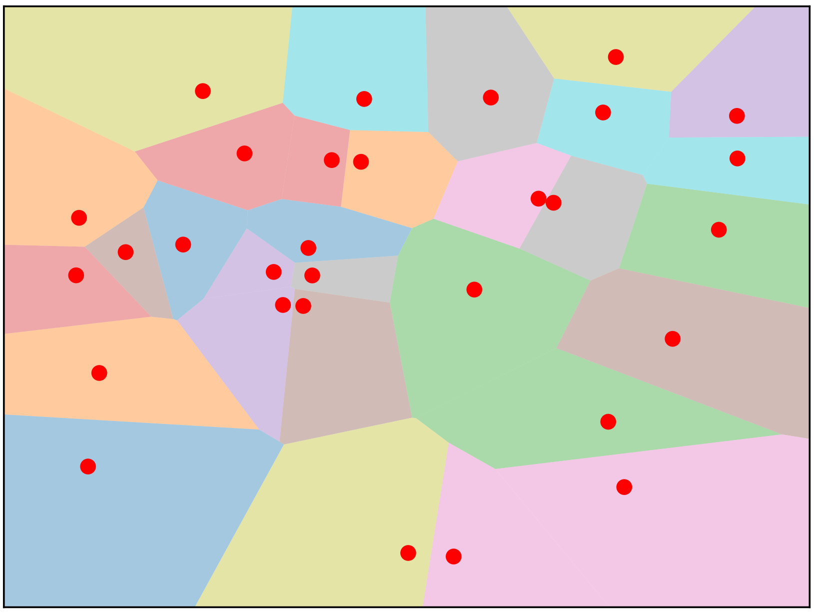


Figure 2: Voronoi Tesselation Cells Corerage Area Pattern

The handover procedure can be defined in three distinct phases: measurement, decision, and execution. We now proceed to discuss each phase individually.

2.4.1 Measurement Phase

In mobile networks, user devices must collect specific metrics periodically, such as SINR, signal strength from serving and neighbor cells, GPS coordinates in some cases, and others. These metrics sometimes are translated into network events, such as the LTE A2, A3, and A4 events later discussed. The network algorithm requests from the user devices the necessary metrics and events periodically, maintaining control over individual user metrics. This communication happens in the forms of measurement control messages, and measurement report messages, and marks the first steps in the handover algorithm loop.

2.4.2 Decision Phase

The decision phase of the handover algorithm comprehends the algorithm's business logic, in which the earlier requested measurement reports are fed into some decision mechanism to find the most appropriate cell for the user device to connect to. Typical decision phases only use signal quality/strength as parameters, often deciding that the best possible cells are the highest numerical value for these quantities. However, it is possible to apply multiple-criteria decision-making mechanisms to achieve more reliable decisions.

This phase needs to consider each neighbor cell included in the measurement reports, or on a neighbors lookup table, to decide the best one. If the algorithm finds the best cell for the user, not being the current serving cell, a handover procedure must be initiated.

2.4.3 Execution Phase

After the handover decision has been made, the handover manager needs to inform all network elements of its decision. A handover request is made by the user's current cell to the target for the session to be transferred. If the handover is acknowledged and confirmed, the link between the user and the current cell may be broken and a new one made with the target cell. The order in which the disconnection and connection events occur characterize the type of handover, as described as follows.

- Hard handover: corresponds to a break-before-make type of handover, where the connection with the serving cell is terminated before a new one is made. This approach reduces complexity for the user, who does not need to be connected to both cells at the same time. Still, the entire process needs to be made in a concise time window for the transfer to be seamless, *i.e.,* for it not to affect the user's experience.

- Soft handover: in the soft handover scheme, a make-before-break approach is used. In other words, the user's session if first transferred to the target cell, and only then the connection with the previous one is broken. Although it seems to be a more reliable method in the case of a handover failure, the user's device must remain connected to both cells for a short time window.

2.4.4 Handover Algorithms

2.4.4.1 Strongest Cell Handover Algorithm

The Strongest Cell, Power Budget (PBGT) is the default handover algorithm in many wireless networks for its simplicity and robustness. The only necessary input for the Handover Manager to make a decision is to receive the power budget from the serving cell and all of the candidate cells measured by the user.

The algorithm is based on two adjustable parameters: hysteresis and time-to-trigger. A hysteresis parameter is considered the minimum difference between the neighbor cells and the serving cells' signal strength, and the time-to-trigger parameter is used, defined as the minimum amount of time that the hysteresis condition needs to be valid for the handover to be made

Figure 3 exemplifies how the algorithm performs its decision in the context of an LTE network. The Reference Signal Received Power (RSRP) stands for the power received by the user. Even though the power received by the target cell is higher than the source cells since the second 37, only in the second 40, the hysteresis condition is met, and the time-to-trigger count is started.

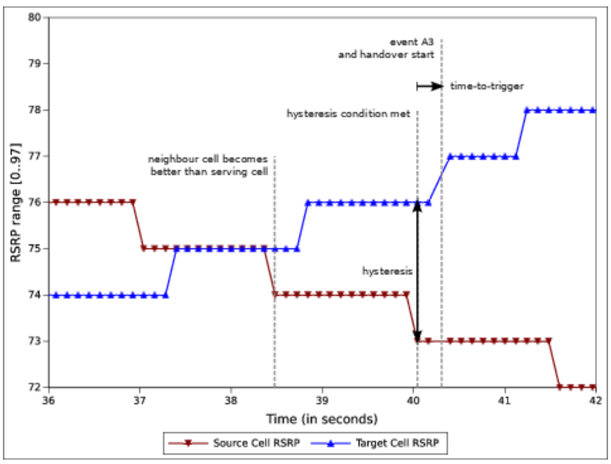


Figure 3: Strongest Cell Algorithm (nsnam.org)

2.4.4.2 RSSI-based Handover Algorithm

Another handover algorithm widely used in wireless networks is the RSSI-based, in which not only the raw power budget is considered, but also the received signal's quality in terms of noise and interference.

The user device measures the RSSI perceived from the current cell and serves one periodically on this algorithm. The RSSI-based algorithm uses two thresholds in its decision, for both current and candidate cells.

For the current cell, a lower threshold is considered, if the RSSI measured for this cell finds itself below this threshold, the handover is considered, however, it is only performed if at the RSSI for at least one neighboring cell is a predefined offset above the current cell's. These thresholds intend to avoid unnecessary handover events, which may degrade the user's experience.

2.5 Multi-Criteria Decision-Making Techniques

When balancing options in systems that can choose among multiple candidates in a certain context, such as cell for a handover algorithm. The problem for the decision scheme arises when each candidate in the decision process possesses many attributes that add different aspects to the decision process. Finding an optimal solution in such context is a challenging task and involves searching a multi-dimensional attribute space, possibly adding computational cost. These problems constitute Multi-Criteria Decision-Making, MCDM, problems, to which several approaches to find a solution, or a group of solutions, have been proposed. The main methods described in the state-of-the-art are: AHP, ELECTRE, TOPSIS, AND Grey Theory. We describe hereafter each method as defined by Aruldos et. al. [12]:

- AHP: The basic idea of AHP is to capture experts' knowledge of phenomena under study. Using the concepts of fuzzy set theory and hierarchical structure analysis a systematic approach is followed for alternative selection and justification problem. Decision-makers usually find that it is more confident to give interval judgments than fixed value judgments [12].

- TOPSIS: The TOPSIS method assumes that each criterion has a tendency of monotonically increasing or decreasing utility which leads to easily define the positive and the negative ideal solutions. To evaluate the relative closeness of the alternatives to the ideal solution Euclidean distance approach is proposed. A series of comparisons of these relative distances will provide the preference order of the alternatives [12].

- ELECTRE (Elimination EtChoix Traduisant la REalite) is one of the MCDM methods and this method allows decision makers to select the best choice with utmost advantage and least conflict in the function of various criteria [12].

- Grey Theory: has a high mathematical analysis of the systems which are partly known and partly unknown and is defined as “insufficient data” and “weak knowledge”. When the decision-making process is not obvious Grey Theory examines the international analysis, there exist a great number of input data and it is distinct and insufficient [12].

Based on the given examples of MCDM methods, the authors choose to rely on AHP, as both problems tackled in this text can be described to some extent by expert’s opinion concerning the metrics being optimized.

2.6 Video Distribution with QoE Support

Video streaming is one of the most widely used applications over wireless networks and is expected to account for more than 80 % of all IP traffic by 2022 [1]. The ability to deliver multimedia content with acceptable quality has become crucial for network operators to keep customers satisfied. To execute such a task, the network must be aware of the quality currently delivered to make the necessary optimizations.

Measuring the quality of video transmission is a non-trivial task, as objective QoS parameters may not be enough to estimate the user’s satisfaction. When able to estimate the user’s satisfaction and the variables that interfere with such, network operators can adapt and optimize the QoE delivered.

While QoE may refer to any type of application experienced by the user, QoE of video transmissions has received a great deal of attention with the growth of video demand all around the world. With that in mind, it is necessary to distinguish between objective and subjective Quality of Experience assessment.

2.6.1 Objective QoE Assessment

Objective QoE assessment is the simplest type of QoE computation, as it involves measurable quantities, like signal-to-noise-ratio, number of interruptions, and other variables. Some techniques compare the original and received sequences to obtain the exact level of degradation, such as PSNR, SSIM, and VQM.

- PSNR: Peak signal-to-noise-ratio, in essence, the ratio is defined as the maximum power of the signal being considered over the power of the noise. The PSNR metric is very correlated to the Mean Squared Error (MSE). As the MSE decreases, PSNR tends to infinite.

- SSIM: The Structural Similarity Index is a metric that receives two images of the same size and outputs a value from 0 to 1, where one means that the two inputted images are identical. From a video perspective, the metric computes the SSIM of every corresponding pair of frames from the original and transmitted sequences and outputs a mean of all values.

- VQM: The Video Quality Measurement metric is based on Discrete Cosine Transform and attempts to weigh how the human eye perceives contrast and luminance in its formula. The metric ranges from 0 to 4, where 0 means a lossless transmission. Just like for the SSIM metric, the VQM is calculated frame by frame.

2.6.2 Subjective QoE Assessment

Subjective QoE analysis is based on the fact that since humans will be consuming the video, their impressions should be enough to attest to its quality. While it is a more trustworthy system in terms of human perception, it is more challenging to scale and automate. One of the most popular subjective QoE assessment techniques is the Mean Opinion Score (MOS).

The Mean Opinion Score is a very simplistic QoE assessment technique, but useful in the sense that it takes in the actual analysis from humans. The model standardized by the International Telecommunications Union (ITU) consists of a scale from 1 to 5, where values close to 1 mean an unfortunate quality video sequence and values closer to 5 mean sequences with good, or acceptable, quality.

While it accurately depicts the human perception of the quality of the transmitted video stream, it has the disadvantage of requiring one or many people to watch the sequence and then grade it, now allowing adjustments mid-transmission.

2.6.3 Hybrid QoE Assessment

Hybrid QoE techniques are a type of mix between Subjective and Objective techniques. These can be defined as an attempt to approximate the subjective evaluation process and produce results similar to the ones a human being would.

Hybrid, or pseudo-subjective techniques, address the limitations of both objective and subjective techniques, as they do not require human interaction, but approximate a human evaluation on time. Several hybrid QoE relies on machine learning techniques to achieve similar results to the ones of humans.

The technique used in this work to perform QoE evaluation will be discussed in Chapter 4 is the pMOS and will be discussed in Chapter 4. The pMOS tool is a random forest trained to react to frame loss rates on a video streaming and output a MOS evaluation. The evaluations are based on the MOS grades given by human volunteers of videos with different loss rates. Using so, a low complexity evaluation can be made during the runtime of the transmission, so network parameters may be adjusted to improve QoE.

2.7 Cloud, Fog and Edge computing

This section introduces the key differences between the main edge-enabled network tiers: cloud, fog, and edge computing, as well as how each model can be applied to facilitate decision-making with acceptable QoS. The main concepts, as defined by Yousefpour *et. al.* [70], are as follows.

2.7.1 The Cloud Model

The cloud computing model refers to remote services being executed in a possibly distant location in large data centers with a high capacity of computing and storage. This model is typical for a small set of data centers to provide users worldwide through the internet.

Cloud data centers offer huge pools of ubiquitous processing and storage capabilities for scalable workloads. One of the consequences of this model is that services may be served a large number of hops away from its clients, increasing the significant latency to reach the servers and get the required responses. Such latency associated with cloud servers makes it unfeasible.

The most common cost approach is a pay-as-you-go model, which allows users to be charged only for the resources they actively use. The types of services offered by cloud providers are as follows:

1. IaaS: Infrastructure as a Service means that the client can contract and configure certain hardware specifications, such as the size of memory, CPU, as well as networking configurations for the server.
2. PaaS: Platform as a Service allows users to use to deploy applications without having to configure the server infrastructure themselves.
3. SaaS: In the Software as a Service paradigm, the software is licensed for users in a platform- and infrastructure-agnostic way. This means that software can be accessed on-demand through a network with little to no concern on the clients' maintenance.

2.7.2 The Fog Model

The IoT-to-Cloud path usually contains a gap in which computing and storage are traditionally not provisioned for end devices. This introduces latency in the applications which need to be routed to distant data centers. The fog computing paradigm is introduced to close this gap and provide efficient cloud-like resources with acceptable QoS, introduction computing, and storage capabilities between the end devices and cloud servers.

The main characteristic of fog computing is a reduced distance from servers and services to the end-users, and consequently reduced the number of hops in the routes from server to the user. Servers in a fog computing environment are horizontally distributed, supporting multiple domains interconnected. Fog servers are characterized mainly by the types of applications they can support. Fog servers can support a range of latency-sensitive applications that cloud services cannot. This results in decreased latency for applications, enabling a more extensive range of applications to be executed remotely, such as autonomous navigation, and augmented reality [62].

One of the critical differences between the fog model and the cloud model is the geographical distribution of the server in the scenario. While cloud servers are all centralized in a distant location, fog servers cover “fog zones”, *i.e.*, they are geographically distributed so that there is always a fog server in the proximity of users. This makes fog services more sensitive to user mobility, as a user can traverse from a fog zone to another, possibly changing networks as it travels. This means that a server that was once optimal may no longer suit its application requirements as a user moves.

However, since users are closer to the services, it is easier to deploy context-aware services in the fog than in the cloud. The fog may be aware, for example, of network conditions for users, as well as mobility patterns and even social patterns.

2.7.3 The Edge Model

The edge computing paradigm if close to the one of fog computing, in a sense that it aims to bring servers and services closer to the end-users, however, in the edge paradigm this is taken a step further, delivering services to the edge of the network, in base stations and access points, in other words, the edge servers must be one hop away from the IoT devices they serve.

Edge servers, just like in the cloud or the fog, must provide storage and computation capabilities for the end devices. Because edge servers are highly close to end-users, they support applications with much stricter requirements, such as applications that require Ultra-Reliable Low-Latency Communication (URLLC), mission-critical applications, and such.

Compared to the fog and the cloud, edge computing consists of relatively small server pools, with limited resources. However, it tends to serve a smaller number of users, resulting in higher service availability, since users do not have to wait as much for their requests to be allocated.

2.8 Service Migration in Edge-enabled Environments

The tremendous geographical distribution present in the Fog and Edge computing paradigms may constitute a problem when applied to mobile scenarios, especially as in mobile and vehicular networks, due to the high velocities and mobility involved in these. Because edge nodes are often found in access points and base stations, as an end device, such as a vehicle traverses through the scenario, causing network changes at certain times, handover events change the topology of the network. If an edge service is being consumed by the mobile node, the QoS of the service may be compromised by the network change.

Service migration is also very challenging. When a user moves through several adjacent or over-lapped geographical areas, service migration should deal with: 1) whether the ongoing service should be migrated out of the current edge server that hosts this service; 2) if the answer is yes, then which edge server the service should be emigrated to; 3) how the service migration process should be carried out, considering the overhead and QoS requirements. This problem comes from the trade-off of migration cost (e.g., migration cost and transmission cost)in the service migration process and improvement of users' expectations on QoS that can be achieved after the migration (i.e., reducing the latency for users or network overhead). It is tough to obtain the optimal service migration because of the high uncertainty of user mobility and request patterns, as well as potential non-linearity of transmission and migration cost [67].

2.9 Connected Autonomous Vehicles

The term "Autonomoues Vehicles" is extensive, possibly referring to several levels of automation within the context of mobility. However, there is no doubt that autonomous vehicles and autonomous driving technology will be a cornerstone of modern transportation systems. They will reduce the number of circulating vehicles, crashes, human errors on the road, and pave the way for new services consumption in traffic [41]. The most basic definition of an autonomous car is a car that can drive with reduced, or nonexistent, human interference.

The levels of autonomous driving technologies are as follows:

0. **No Automation** – At Level 0 Autonomy, the driver performs all operating tasks like steering, braking, accelerating or slowing down, and so forth.

1. **Driver Assistance** – At this level, the vehicle can assist with some functions, but the driver still handles all accelerating, braking, and monitoring of the surrounding environment.

2. **Partial Automation** – The vehicle can assist with steering or acceleration functions and allow the driver to disengage from some of their tasks. The driver must always be ready to take control of the vehicle, and they are still responsible for most safety-critical functions and all monitoring of the environment.

3. **Conditional Automation** – The vehicle itself controls all monitoring of the environment (using sensors like LiDAR [66]). The driver's atten tion is still critical at

this level but can disengage from "safety-critical" functions like braking and leave it to the technology when conditions are safe.

4. **High Automation** – At Level 4, the autonomous driving system would first notify the driver when conditions are safe, and only then does the driver switch the vehicle into this mode. It cannot determine between more dynamic driving situations like traffic jams or a merge onto the highway.

5. **Complete Automation** – This level of autonomous driving requires absolutely no human attention. There is no need for pedals, brakes, or a steering wheel. The autonomous vehicle system controls all critical tasks, monitors the environment, and identifies unique driving conditions like traffic jams.

2.10 Service Offloading for CAV

Connected autonomous vehicles require ubiquitous connectivity for their correct functioning. Some of the essential operations they require are *(i)*: task offloading to edge and fog servers; and *(ii)*: context information sharing with the network and with other vehicles. This section will tackle the former. We will review some of the problems that arise from this necessity.

- Task Offloading: Task offloading is not exclusive to autonomous vehicles. To overcome the computational limitations of local devices, it has been a problem in other contexts, including mobile and Unmanned Aerial Vehicles. The case of autonomous vehicles has some particularities, especially the stringent requirements to be met.

- Latency Requirements: Autonomous vehicles are sensitive to latency and disconnections, since compromised QoS may put the passengers and people around the vehicle in danger. Thus, many authors agree that if the decisions for an autonomous vehicle must be taken within 10ms of the measurement [61]. Thus cloud offloading does not allow for these requirements to be met, leaving only edge and fog servers available.

- Resource Allocation: Edge servers do not have as many resources as cloud data centers. This means that a limited number of vehicles may use the same server in a given moment. If the demand is too high, some of the vehicles will have to use higher-level servers in the fog layer or above, leading to sub-optimal performance.

- Server choice and allocation: Server choice in edge scenarios must be prudent in this case. If the network allows a vehicle or a group of vehicles to behave greedily, taking themselves the best servers, some of the other vehicles may not find the requirements necessary to run their applications. Because of this, the network must decide the allocation process in a way that all vehicles have the minimum resources necessary.

- Migrations: As previously discussed, the high mobility in vehicular scenarios is challenging to the dead from an edge and fog perspective. This is because as the vehicles go from one network area to other services need to "follow" them through migrations to new servers, to stay close and with a short latency. It is not guaranteed that the migrations will happen promptly, or that the target server has enough resources to support the new vehicle, so the network must plan for migrations proactively to reduce failures and service interruptions.

- Virtualization: The offloading services in the servers must be isolated, such as services in cloud data centers. This is done through the creation of Virtual Machines and Containers. Virtual Machines (VMs), while being more replicable, are heavy and must carry the entire platform on which services run, while containers only need application-specific files. Migrations may be done transferring the VMs or containers through the backhaul to the new servers. Transferring VMs will induce a higher workload in the network than containers.

CHAPTER 3

Related Works

This Chapter showcases the main recent works in the state-of-the-art that tackle the research problems present in this dissertation. We define the main shortcomings of each work and present the best research directions folowed our solutions.

3.1 Research Problem #1: Handover in Ultra-Dense Networks

Gong *et.al* [26] proposed a Fuzzy Analytical Hierarchical Process (FAHP) algorithm to reduce failure and ping-pong probability in Heterogeneous Ultra-dense by defining a Time-To-Trigger (TTT) during handover execution. The proposed mechanism considers a two-tier heterogeneous network, in which the tier of the cells is given by a Poisson Point Process, which may result in unrealistic scenario deployment. In such a case, it may be better to follow 3GPP guidelines for a realistic scenario. Handover procedures from the macrocells to small cells and vice versa are implemented in a similar manner, which is a valid approach, and handover failures are implemented analytically. The authors use the remaining time in the cell coverage and the cell traffic load as a factor in their Fuzzy systems, which must be normalized since they are different measures with different ranges. Although it highlights the importance of a multi-parameter handover decision, the use of TTT can have undesired effects, such as link failures and delayed handovers [32]. Simulations are carried out to validate the solution compared to other two handover approaches: a more simplistic, more reliable cell algorithm, and a handover based on node movement. These are relatively simple solutions, and it is unclear how the proposed algorithm would work against more recent works in the state-of-the-art.

Another work built on top of a Fuzzy system is proposer by Silva *et.al.* [60], which proposes an adaptive TTT threshold for handover based on Fuzzy logic and user speed. Such a handover algorithm collects mobility parameters to predict user location for content dissemination, and not for handover purposes, showing that offloading from macrocells to Small Cells can be essential in a heterogeneous environment. The proposed solution evaluates a preliminary condition that states that any candidate cell must receive a received power at least higher than the current one. If this condition evaluates true, then the algorithm must define the most appropriate hysteresis value for the handover. A wrong choice in this value may cause the nodes to remain in worsening connections, or to facilitate ping-pong handovers. The Fuzzy parameters defined by the authors are: node velocity, received RSRP; and received RSRQ. These parameters are put through membership functions, and a set of rules which regulate the output—the defuzzification process than finds the corresponding hysteresis value for the combination of inputs. One of the main benefits of the proposed scheme is the reduced ping-pong rates in dense scenarios. Simulations in a very dense scenario, in an area of $1km^2$, with 200 small and two macrocells, show an improvement in the average number of handovers, handover failure ratio, and ping-pong handover. However, the scheme is not intended for multimedia traffic, so it does not consider QoE for decision making.

Another work that focuses on an adaptive Time-To-Trigger method for handover optimization is proposed by Liu *et. al.* [36], who study the problem created by dense deployment of small cells in the network. Frequently, users tend to be in the coverage areas of more than one cell at a time, challenging the mobility management in the network, increasing the frequency of handovers, and the number of ping-pong handovers. The authors propose a joint Fuzzy and TOPSIS model for cell and Time-To-Trigger value selection. The proposed model offers a handover scheme, which integrates both fuzzy logic and multiple attributes decision algorithms (MADM). The authors also propose a clustering approach to optimize the fuzzy membership function definitions. However, this work is not tuned to the particular challenges of vehicular networks, as well as for the distribution of multimedia content in the network.

Arshad *et.al.* [11] showed that handover introduces an overhead in the network and is, sometimes, redundant. Skipping some handovers can be beneficial for the network while maintaining a seamless QoS. However, that work offers small support for video transmission and may not suit the strict requirements involved. Demarchou *et.al.* [23] studied the challenge of reducing handover rates (*i.e.,* , handover skipping) in ultra-dense networks. That work considers the trajectory prediction in the skipping decision, but only assumes a simple model based on position and velocity. Xu *et. al.* [68] proposed a delay oriented cross-tier handover skipping to maximize the performance of low latency applications in ultra-dense networks. Their work derived an analytical expression for the adequate capacity of users during the handover execution and proposed a resource allocation scheme in Target Cells to reduce blocking probability. It does not employ predictive schemes, or mobility information into the decision, which may improve the decision quality and positively impact user QoE.

Medeiros *et. al.* [43] showed the importance of performing a multi-criteria handover decision to balance metrics from different layers, namely, radio measurements, QoS, and QoE. That work uses AHP to balance the metrics according to predefined importance levels assigned to each. Still, the algorithm presents high handover rates, which is harmful to QoE in dense scenarios. Sargento *et. al.* [57] proposed a connection manager for VANETs with heterogeneous technologies, VANET Connection Manager (VCM), which is based on an Analytical Hierarchic Process (AHP) that combines information from multiple sources (vehicle speed, GPS, heading, RSSI, and available technologies such as DSRC/WAVE, IEEE 802.11 and 4G Cellular), and decides what is the best connection available at all times, also trying to minimize the number of handovers. The AHP is optimized using interaction with a Genetic Algorithm (GA). This approach includes mobility prediction through the expected connectivity time but does not include QoE requirements. Zhang *et. al.* [72] proposed a classification of applications sensitive and insensitive based on user experience. A handover decision switches to a more energy-efficient network during idle timer and a high-performance network when predicted. Chen *et. al.* [17] proposed a QoE estimation to correlate QoS and QoE to improve user satisfaction, not focusing only on call blocking probability and handover dropping probability. However, video sharing requires more subjective metrics to describe QoE, such as MOS, which can be mimicked by machine learning algorithms and integrated into automated decisions.

Table 1 shows the presented related works organized in terms of three main characteristics: the techniques used by the algorithm, the presence of a QoE assessment toll, the knowledge of user mobility, and the used of a handover skipping strategy for ultra-dense networks.

Table 1: Summary of analyzed handover algorithms for ultra-dense VANET scenarios

Work Considered	Features			
	Technique Used	QoE-Aware	Mobility Prediction	Handover Skipping
Gong et al. [26]	Adaptive TTT			
Silva et al. [60]	Adaptive TTT			
Liu et al. [36]	Fuzzy Logic			
Arshad et al. [11]	Handover Skipping			✓
Demarchou et al. [23]	Handover Skipping		Assumed present	✓
Xu et al. [68]	Delay-Oriented Handover Skipping			✓
Medeiros et al. [43]	AHP	✓		
Sargento et al. [57]	AHP		Expected contact time	✓
Zhang et al. [72]	Q Learning	✓		
Chen et al. [17]	Q Learning	✓		
HoVe	Mobility Prediction + AHP	✓	✓	✓

3.2 Research problem #2: Connected Autonomous Vehicles and Service Migration

Allocating resources and performing computing tasks at the edge of the network is a prominent solution for low latency and high bandwidth requirements for applications and services in-vehicle environments. However, it brings with itself challenges concerning resources and mobility management. Different approaches have been suggested in recent years to solve these problems. One of the promising ones is the pre-migration of services to follow the user's mobility and keep low latency. This section reviews some of the state-of-the-art solutions to these challenges.

Some works study the requirements of vehicular applications in autonomous and connected vehicles, and the limitations of onboard units in performing the necessary computations, such as Li *et. al.*[34]. The work proposes a computation allocation framework for offloading of CAVs tasks from onboard units to Vehicular Edge Cloud Computing (VECC). The proposed solution proves itself with greater energy efficiency by allocating the minimum required resource blocks to each vehicle. However, this works does not tack some of the critical issues VECC carries, including mobility and resource management.

In real-world scenarios, the accuracy of the mobility prediction decreases the more extended the predicted period is. Yu *et. al.* [71] proposes an offline pre-migration of services in mobile edge computing. The work uses a mobility prediction scheme to minimize the average latency of the service in the long term. The algorithm, while finding an optimal solution, may have limitations in real-time, as the processing is assumed to be offline.

Other approaches assume mobility management in vehicular networks from a transportation point-of-view. Liao *et. al.* [35] proposes a vehicle-as-a-service approach to mobility management and computational tasks migration in edge-enable vehicular networks for path planning. This work takes advantage of the innate mobility of the vehicles to distribute computing capability through the network evenly. This work assumes that computing tasks can be done locally and does not consider the case of service migration. However, the results do not show the impact of the proposed scheme in a real or simulated vehicular scenario.

Chen *et. al.* [18] investigates the new paradigm of Cognitive Edge Computing. In this work, cognitive engines in edge servers can learn the computing and network resources available at the edge and solve the problem of communication bandwidth and delay through the fusion of computing, communication, and storage. The introduction of Cognition to the network operation showed that predicting user behavior can significantly improve the quality level perceived by the end-user. However, it is not best suited for the CAV use case, as it was designed for an application that is not as stringent in requirements.

Ouyang *et. al.* [49] tackles the problem of keeping services close to used in edge computing scenarios, where user mobility is unpredictable. The system does not need prior information about user mobility statistics, as it uses real-time optimization to reduce the problem's complexity. The solution aims to reduce overall migration costs but could be optimized for vehicular scenarios with specialized mobility models. Also, Gao *et. al.* [24] proposes a heuristic-based migration algorithm to serve users with varying deadlines, considering user-generated data and the contact patterns between the nodes. Despite employing mobility models in the decision, the proposed solution lacks in terms of QoS and radio resources support for the applications and services.

Table 2 summarizes the main characteristics of previous works in terms of support for mobility prediction, the use of pre-migration, and the nature of the computation (online/offline). Based on our analysis of the state-of-the-art, we conclude that the pre-migration of decision tasks from CAV to edge servers with support for mobility prediction, and an online decision scheme has not been done in the state-of-the-art, to the extent of our knowledge.

Table 2: Summary of Existing Works for Service Migration for Edge-Enabled CAV

Work	Pre-migration	Mobility Support	Online Computation
Yu et. al. [71]	✓	✓	
Liao et. al. [35]		✓	
Chen et al [18]	✓	✓	
Li et. al. [34]		✓	
Ouyang et. al.[49]	✓		✓
Gao et. al. [24]		✓	✓
MOSAIC	✓	✓	✓

CHAPTER 4

A Handover Algorithm for Video Distribution over Vehicular Networks

This section introduces the HoVe algorithm. Which provides handover with QoE support for video flows in 5G VANETs, considering Navigation History, QoE, and radio parameters for the handover decision. We consider a 5G scenario comprising Small Cells and Macro Cells and a Traffic Management System (TMS). HoVe relies on AHP to adjust the degree of importance of each parameter, as well as to compute the quality of each available network to select the best network for the vehicle to connect.

4.1 Algorithm Description

The handover process is performed in three distinct steps: measurement, decision, and execution. The first step consists of information gathering, where the algorithm collects important metrics for decision-making, *i.e.,* , radio resources, packet delivery ratio, QoE, and Vehicle Mobility. Afterward, this information is evaluated in the Decision step to choose the best network available. If the algorithm decides so, a handover is performed. HoVe uses a seamless handover process (make-before-break).

The decision phase occurs individually in each cell, where the handover manager entity receives measurements, performs the decision, and coordinates the handover execution. Moreover, each network component has information about the location of the network cells and can use it in the evaluation process.

Figure 4 illustrates the interactions between the vehicles, access points, and the Handover Manager. Mobile nodes continuously monitor packet flows to obtain current QoE levels; this information is then sent to the Handover Manager along with Radio

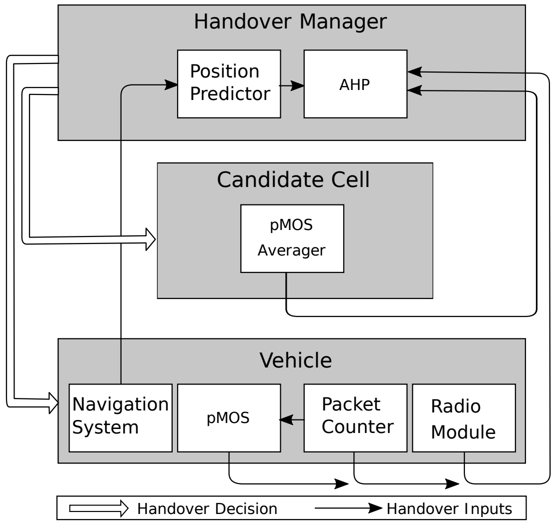


Figure 4: System Overview

measurements and the vehicle's coordinates. Navigation Information/routes of vehicles are used to predict the user's near future positions, and when all the inputs are available, the AHP algorithm is executed to evaluate all networks. If a handover is necessary, the current serving cell initiates the communication with the target cell and transfers the user. HoVe is compatible with traditional handover protocols and can be easily integrated.

Wireless links may last for very short amounts of time due to the highly mobile nature of vehicular networks. Therefore, the handover algorithm must choose a network that remains available for a longer time window according to a short-term position predictor.

4.1.1 Mobility Prediction Scenario

We consider both ARIMA and KF as use cases for the mobility prediction technique considered by Hove, but it can be any other position prediction scheme. Both ARIMA and KF can be used to predict the vehicle's future position $L(x_i, y_i, t+1)$ based on the current one $L(x_i, y_i, t)$. In this sense, Hove iterates the mobility prediction algorithm every time a new measurement arrives, where the intervals between measurements define the granularity of the filter. In our tests, we adopted the granularity of 1 second.

4.1.1.1 Autoregressive Integrated Moving Average

The Autoregressive Integrated Moving Average, ARIMA is a statistical model to analyze and forecast time series. It works by taking values of series and making them stationary if necessary. A stationary time series has no trend, and the amplitude of its variations around the mean is constant. In the ARIMA model, future values of series are assumed to be a linear combination of past values and past moving averages.

ARIMA is described as a 3-tuple (p, d, q), where p corresponds to the number of past measurements weighted in the estimation, d consists of the number of differencing series to make statistically stationary, and q corresponds to the number of past moving averages. The basic formulation of the model is given by Eq. 4.1. We denote past terms as y, past moving averages as ϵ, while θ and ϕ are individual weights for each term and will be trained by the model.

$$\begin{aligned} y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-1} + \phi_3 y_{t-3} + \cdots + \phi_p y_{t-p} \\ \epsilon_0 + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-1} + \theta_3 \epsilon_{t-3} + \cdots + \theta_q \epsilon_{t-q}. \end{aligned} \tag{4.1}$$

The number of past value terms and past moving averages depends on the studied series, where some series are mostly dependant on weighted past values and do not need any moving average terms. The model can be represented by the notation $ARIMA(5,1,0)$, which means we use five past terms, perform one differentiation, and consider no past moving averages. This configuration was found by means of a grid-search.

ARIMA is used to forecast a single-variable time series, and, thus, it has to be done a training step separately for the latitude and longitude measurements. The first step for the general ARIMA formulation is to define the differencing order, *i.e.,* , the number of times each term is subtracted from the next one, given by the parameter d, as shown in Eq. 4.2). The ARIMA model can be used for the vehicle mobility prediction $L(x_i, y_i, t+1)$. In this sense, the model must be trained for each vehicle separately and for each coordinate (*i.e.,* , latitude, and longitude).

$$y_t = \begin{cases} Y_t, & \text{if } d = 0 \\ (Y_t - Y_{t-1}), & \text{if } d = 1 \\ (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}), & \text{if } d = 2 \\ \text{and so on} \end{cases} \tag{4.2}$$

4.1.1.2 Kalman Filter

KF tries to estimate a state $x_t \in \mathbb{R}^n$ based on previous state x_{t-1}, *i.e.,* , the filter only needs the value of the previous state to estimate the next one. The state x in a KF is a vector containing a pair of vehicle geographic coordinates g_t, namely latitude and longitude, at a given moment t (*i.e.,* , $L(x_i, y_i, t)$). Explicitly, we model the process as in a stochastic difference equation shown in Eq. 4.3. We denote A as a $n \times n$ matrix that relates the previous state to the current one, and $w \in \mathbb{R}^n$ as noise estimation.

$$x_t = Ax_{t-1} + w_{t-1}. \tag{4.3}$$

The estimation considers a measurement given by Z_k, as shown in Eq. 4.4. It can be modeled in terms of the prediction with a correcting factor H and a noise v_k.

$$Z_k = HXk + v_k. \tag{4.4}$$

We define $\hat{x}_k^-$ as previous state, x_k as predicted state, and $\hat{x}_k$ as following state, where $\hat{x}_k^-$ and $\hat{x}_k$ are real values of the process. We want to estimate x_k based on the measurement Z_k. The previous and following errors are defined by e_k^- and e_k, respectively, as shown in Eqs. 4.5 and 4.6.

$$e_k^- = x_k - \hat{x}_k^-. \tag{4.5}$$

$$e_k = x_k - \hat{x}_k. \tag{4.6}$$

Also, the previous state covariance can be defined based on Eq. 4.7, and the following state covariance by Eq. 4.8 as the expected value of the error, times the error matrix transpose. The goal of the filter is to minimize the error covariance P_k.

$$P_k^- = E\left[e_k^- e_k^{-T}\right]. \tag{4.7}$$

$$P_k = E\left[e_k e_k^T\right]. \tag{4.8}$$

We express the following state as a linear combination of the previous state, and a correction term proportional to the difference between measurement and state value, as shown in Eq. 4.9, the value of $\hat{x}_k$ corresponds to the vector of predicted coordinates in the next measurement g_{t+1}.

$$\hat{x}_k = \hat{x}_k^- + K\left(z_k - H\hat{x}_k^-\right). \tag{4.9}$$

The matrix $K_{n\times m}$ is the gain, which should minimize the following error covariance. We can minimize the error by replacing Eq. 4.9 into Eq. 4.6 and, then, deriving the result. In this way, final formulas for computing the gain of the filter to be used in the estimation is given by Eqs. 4.10 and 4.11.

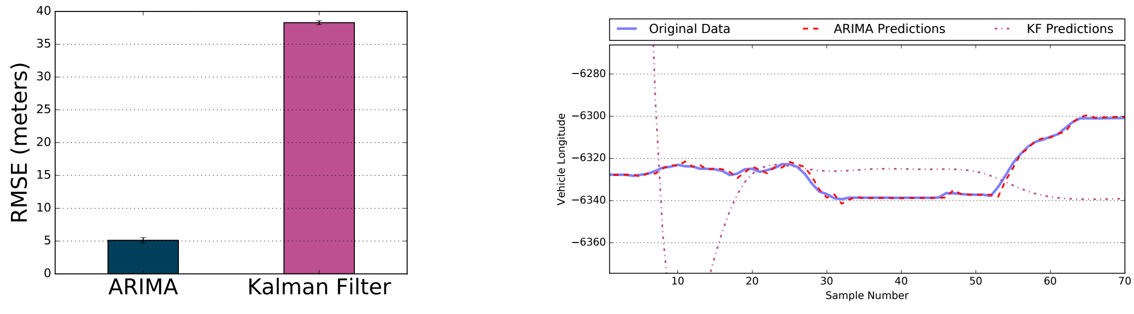
$$K_k = P_k^- H^T\left(HP_k^T H^T + R\right)^{-1}. \tag{4.10}$$

$$K_k = \frac{P_k^- H^T}{HP_k^- H^T + R}. \tag{4.11}$$

4.1.1.3 Mobility Prediction Accuracy

We tested the mobility prediction accuracy of KF and ARIMA in a real-world vehicular dataset to choose one of them as part of the handover algorithm. In this sense, we considered a vehicular mobility trace collected from approximately 500 taxis from San Francisco [54]. The dataset consists of GPS measurements of 500+ cabs in the San Francisco bay area over one month, generating more than 10 million samples. We consider ARIMA(5,1,0) in such a dataset, *i.e.,* , it means that we consider two past values, the series is direffenced twice to make it stationary, and one is moving average term. These parameters were found using a Grid Search estimator for better performance, which tests different configurations to find a local optimal configuration. We consider 60% of the data for training and the remaining 40% for tests.

Figure 5(a) shows the average Root-Mean-Square Deviation (RMSE) for the ARIMA and KF to predict each vehicle location in the dataset. By analyzing the results, we can observe that KF has an error 85.7% higher than the ARIMA. Vehicle movement may be irregular and non-linear for the most part, but KF is more accurate when the analyzed data has a linear nature due to its interactive nature. In this sense, KF needs time to adjust to mobility changes in parameters such as speed and direction, *i.e.,* , KF makes adjustments online. On the other hand, ARIMA can predict the mobility pattern with high accuracy after training and is very robust even with non-linear data. RMSE results can be explained using Figure 5(b), which shows the vehicle's longitude over time for a given vehicle. By analyzing the results, we can conclude that ARIMA predictions are much closer to the original data points. In contrast, KF predictions, in some cases, are very distant from the original data points. For instance, at sample 30, the vehicle turned (left or right), and ARIMA can predict such a vehicle mobility pattern, while the KF does not detect it.



(a) (b)

Figure 5: Mobility prediction results (a) RMSE and (b) Vehicle Longitude predictions for ARIMA and Kalman Filter applied for one vehicle of San Francisco Taxi Dataset

4.1.2 QoE Assesment

Hove uses pMOS, a low complexity QoE monitor presented by Medeiros *et. al.* [44]. Videos are typically composed of the frame types, Intra-coded picture (I), Predicted picture (P), and Bidirectional predicted picture (B), each with a different degree of importance when reconstruction the video sequence. The I-Frames carry all the information needed in a picture, as P-Frames and B-Frames only carry the bits of information that changes from the previous image to the current.

pMOS consists of a random forest that receives as input the loss rates for I-, P- and B-frames and outputs a Mean Score Opinion (MOS) value. pMOS was trained with a subjective analysis performed by human subjects for imitating human perception to frame losses. In this context, users consuming video content identify lost frames in the video and their respective types. The loss-ratio for each frame type is reported to HoVe and fed to the pMOS module. Furthermore, radio measurements are also traditionally reported to HoVe. SINR uses a signal quality metric and is also weighted in the handover decision.

The handover Manager finds the best network given the collected metrics, configuring a Multiple-Criteria Decision-Making problem. We chose AHP to balance the input metrics. AHP considers a pairwise comparison between the numerical values of each collected parameter and their relative degrees of importance, to adjust their weights of each parameter at runtime. The weights of the inputs must be defined when configuring the algorithm. High weight means more importance should be attached to this particular metric, and we define five importance levels, as shown in Table 5.

Table 3: Pairwise Context Importance

$c_{i,j}$	Definition
4	i is much more important than j
2	i is more important than j
1	i is as important as j
1/2	i is less important than j
1/4	i is much less important than j

The Handover Manager constructs for each vehicle a matrix to compare all pairs of metrics. We denote $c_{i,j}$ as how important the i^{th} element is compared with the j^{th} element. Also, $A = (C_{i,j})_{n\times n}$ represents the comparison matrix, where n denotes the number of elements to be compared, as shown in Eq. (5.1).

$$A = (C_{i,j})_{n\times n} = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \end{array} \begin{array}{c} \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\ \begin{pmatrix} c_{1,1} & c_{1,2} & c_{1,3} \\ c_{2,1} & c_{2,2} & c_{2,3} \\ c_{3,1} & c_{3,2} & c_{3,3} \end{pmatrix} \end{array} \tag{4.12}$$

To guarantee consistent QoE throughout a transmission, the pMOS metric has the highest priority compared to mobility and QoS and Signal. We define the trajectory parameter as the estimated distance between the vehicle and the access point in the short-term future. QoS and Signal parameters are combined into a single input for the algorithm.

$$I = \begin{array}{c c} & \begin{array}{ccc} QoE & Distance & QoS/Signal \end{array} \\ \begin{array}{c} QoE \\ Distance \\ QoS/Signal \end{array} & \begin{pmatrix} 1 & 2 & 4 \\ 1/2 & 1 & 2 \\ 1/4 & 1/2 & 1 \end{pmatrix} \end{array} \tag{4.13}$$

After the selection of the relative importance, the matrix is normalized by dividing each element by the sum of its column and finding the eigenvector for the matrix. For instance, in Eq. 4.13 we find the eigenvector $W = [0.57\ 0.28\ 0.14]$, meaning that that QoE will have a weight of 0.57, 0.28 for Distance and 0.14 for QoS/Signal.

The Handover Manager computes the score S_i for all available networks based on Eq. (4.14), where c_i represents the weight for a given metric, and P_j is the value for a given metric, *i.e.,* , QoE, QoS, and Link Duration, obtained in the handover measurement phase. Finally, the handover Manager selects the cell with the highest S_i value, which is the most suitable access point for the vehicle to connect at the moment and what the video.

$$S_i = \sum_{j=1}^{n} c_j \times P_j \tag{4.14}$$

Figure 6 details the steps involved in the HoVe execution. The current serving cell periodically requests measurement reports to the user, in our case, the vehicle. The vehicle evaluates the current QoE, in the case of video content being consumed, and sends it to its radio, QoE, and coordinates measurements to the current cell. If a handover is necessary, the Serving Cell requests the transfer to the Target Cell, and the channel allocation and synchronization can begin. When the process is complete, the user sessions will now be handled by its new serving cell.

4.2 Evaluation

4.2.1 Simulation Description and Metrics

HoVe was implemented and tested on the NS-3.27[1] simulator, where 33 simulations were conducted with different randomly generated seeds that were fed to its default pseudo-random number generator (MRG32k3a). Thus, it is possible to provide independent streams of random variables for each probabilistic model used. Results show the values with a confidence interval of 95%.

NS-3 implements the LTE protocol stack for communication between the mobile user with the radio base station. We consider simulation parameters presented by Tartarini et al. [63]. The scenario is chosen as a typical urban ultra-dense vehicular network

[1] http://www.nsnam.org/

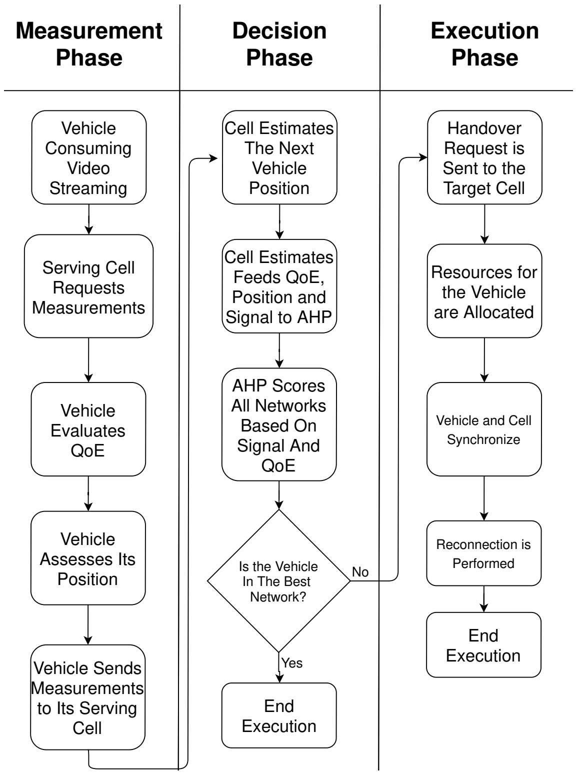


Figure 6: Execution Flowchart for HoVe

with two tiers, composed of two High Power eNodeBs (Macro Cells / LTE) and Low Power eNodeBs (Small Cells / WiFi) randomly distributed in the simulation scenario. In this scenario, nodes move in a grid topology at a 2D rectangular area of $4km^2$ (2000m $\times$ 2000m).

For the simulation of traffic and vehicle mobility, we employed the Simulation of Urban Mobility (SUMO)[2], which is an open-source traffic simulator to model and to manipulate objects in the grid scenario. SUMO allows us to reproduce the desired vehicle movements with a predefined path and speeds based on empirical data. We consider a scenario composed of vehicles at different speeds as expected in real cities (ranging between 10-70 km/h).

We considered video sequences with different motions and complexity levels, *i.e.,* Container, Mobile, and Highway, which are downloaded from a well-known Video-trace repository[3]. Even small differences in the videos' characteristics can influence the obtained QoE values [76]. These videos mainly have a duration of 10 seconds (except Highway with 20 seconds) and 300 frames each (except Highway with 600 frames), encoded with an H.264 codec ranging from 210 kbps (Highway) up to 230 kbps (Container), 30 fps and intermediate size (352 x 288 pixels). It should be noted that all the videos evaluated are streamed in a loop. The decoder uses a Frame-Copy method as error concealment, replacing each lost frame with the last received one to reduce frame loss and to maintain

[2]http://sumo.dlr.de
[3]http://media.xiph.org/video/derf/

video quality. The main simulation parameters can be seen in Table 6.

Table 4: Simulation Parameters

Parameter	Value
Nodes speed	$[10 - 70]\ km/h$
Number of UEs	60
Number of Macro Cells	2
Number of Small Cells	50
Macro Cell Transmission Power	46 dBm
Small Cell Transmission Power	23 dBm
Propagation Loss Model	Nakagami
Scenario Size	$2km \times 2km$
Network Topology	6×6 grid
Video Tested	Highway, Container and Mobile
Simulated Time	60 Seconds
Transmission Start Time	20 Seconds
Number of Simulations	33

The handover algorithms compared are implemented on the LTE-handover API present in the NS-3 Simulator, where all the relevant metrics can be accessed and evaluated for the decision and execution of the handover. NS-3 implements a hard handover mechanism (break-before-make), and the measurements and evaluations are performed periodically.

HoVe is tested against the SER [44] algorithm, and standard LTE handover mechanisms such as RSSI-based handover and Strongest Cells referred to as PBGT (Power Budget). SER is a QoE-aware handover algorithm for Heterogeneous Networks (HetNets); it has shown superior quality in the delivery of videos for mobile users. The Strongest Cell handover performs a signal strength-based decision, in which the handover is executed if a neighbor cell's received strength is superior to the serving cell's plus a hysteresis value. Such a difference is maintained throughout a previous set Time-To-Trigger [2]. Furthermore, the RSSI-Based Handover Algorithm uses LTE's events A2 and A4 to trigger the handover execution. Both solutions take into account solely radio measurements in the process. RSSI-based and Strongest Cells are present by default in the simulator. SER was implemented as described in the paper, where it is defined [44].

QoE metrics overcome the limitations of QoS metrics for video quality assessment since QoS metrics fail to capture subjective aspects of video content related to the human experience [9]. In this way, we consider Structural Similarity (SSIM) as the QoE metric to evaluate the end-users' video degradation. SSIM compares the variance between the original video and the original sequence concerning luminance, contrast, and structural similarity. SSIM values range from 0 to 1, as defined in Chapter 2.

4.2.2 Simulation Results

Figure 7 shows the average SSIM achieved by each algorithm tested in the form of a bar chart, with a confidence interval of 95%. We can see that HoVe was able to deliver videos with higher user experience than competing algorithms. Even SER, which considers QoE, wasn't able to adapt well to a denser scenario, causing decreased QoE to the users. We consider that a satisfactory user experience requires an SSIM of at least 0.8. HoVe shows an average of 0.92. Considering the lowest bound of HoVe's confidence interval with the other algorithm's highest bound, HoVe performs 18% better than SER, 19% better than the RSSI-based, and 19% better than PBGT.

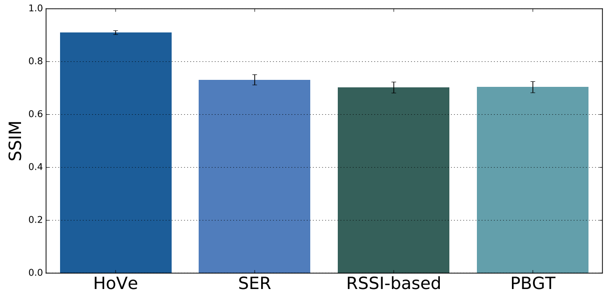


Figure 7: SSIM Obtained by Each Algorithm Over the Simulations

Figure 8 shows the average amount of handovers necessary to deliver a single video as the simulated time increases. We notice that the SER algorithm performs the highest amount of handovers, given that it has no constraints like a hysteresis or a Time-To-Trigger. The RSSI-based approach also has a higher number of handovers, as it is more sensitive to channel variations. At the same time, the Strongest Cell mechanism makes the least amount of handovers among the tested algorithms because it tends to connect to Macro Cells more frequently, which can be less effective in terms of bandwidth. We can see that after 20 seconds into the simulation, the rate at which handovers are performed stabilizes as the decision is optimized to maximize the duration of the link while maintaining acceptable QoE levels.

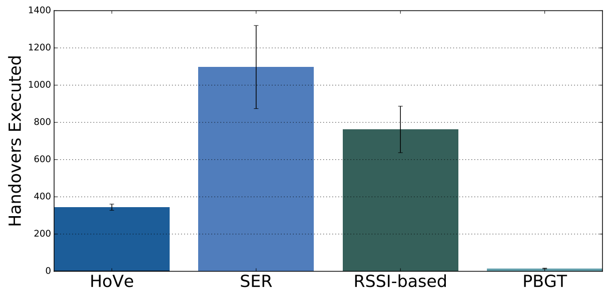


Figure 8: Average Number of Handovers in the Transmission of one Video

Figure 9 shows that proper QoS levels were also ensured with the use of HoVe, which maintains the PDR at around 80%, at least 30% more than any of the other algorithms. SER, RSSI-based, and PBGT maintain the PDR between 50% and 40%, due to more inefficient mobility management: a high number of handovers in the case of SER and RSSI-based, and keeping connected to an overloaded cell in the case of PBGT.

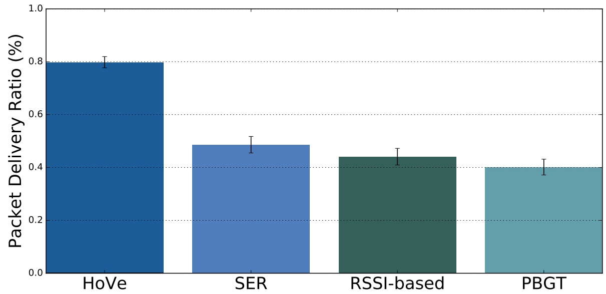


Figure 9: PDR Obtained by Each Algorithm

A random video was selected to illustrate the behavior of the perceived QoE at each moment of the transmission regarding SSIM, as shown in Figure 10. HoVe provides a consistently better SSIM score at each frame of the video throughout all of its duration. We can see three moments where the SSIM value for HoVe dropped, corresponding to instants where a handover was performed. SER delivered better QoE than RSSI-based and PBGT. However, the quality was volatile. We notice that the quality of the transmission dropped around frame #1241. This is because this is a more complex frame with a higher probability of being lost, compromising than the following frames in the GoP.

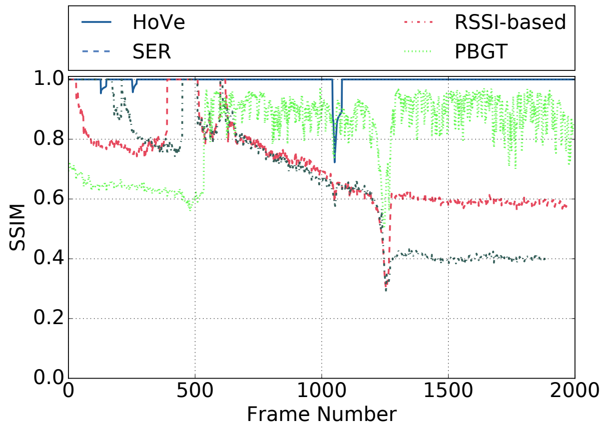


Figure 10: SSIM For Each Frame of Video number #42

Figure 11 shows the ping-pong handover rate by HoVe, SER, RSSI-based, and PGBT handover algorithms. It is essential to highlight that we consider a ping-pong handover as soon as a user leaves a cell and returns to it within a window of 4 seconds. By analyzing the results, we can conclude that HoVe keeps the ping-pong rate around 2%, which is an indication of a better decision policy that avoids such a phenomenon. As mentioned before, PBGT performs a smaller amount of handovers, and, consequently, has a smaller ping-pong probability within the considered window. On the other hand, NC-Skipping, SER, and SINR-based algorithms have higher ping-pongs, due to the fact they do not have a transparent barrier against it. Even with a skipping mechanism, these approaches are not coupled with a multiple criteria strategy and are then also susceptible to ping-pong.

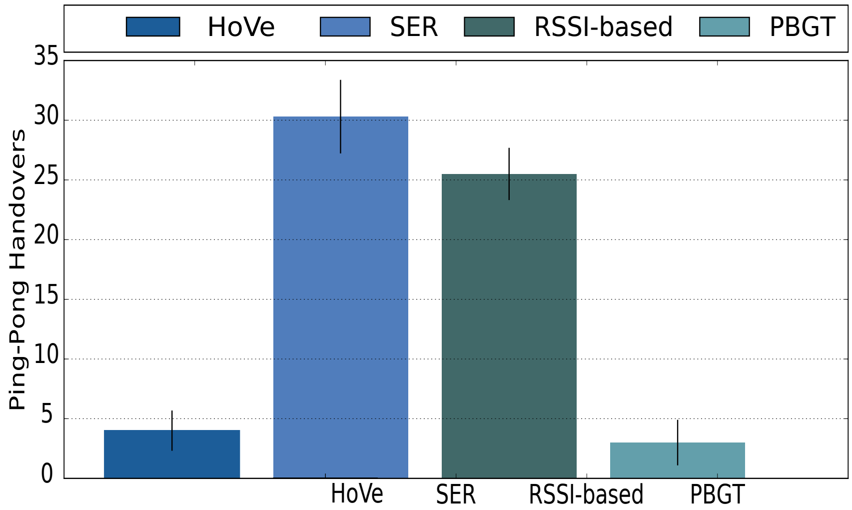


Figure 11: Ping-Pong Handover ratio by different handover algorithms

In Figure 12, random frames were selected from the videos transmitted under each algorithm. Figures 12(a), 12(f) and 12(k) show the frames from the original videos transmitted, alongside with the versions delivered to the end-users by each algorithm under the same scenario. We notice that frames are significantly closer to the original when HoVe is used compared to the other algorithms tested. The most accentuated degradation is perceived on the videos with the most motion, like Highway and Mobile, since it makes them more sensitive to frame losses, causing the most impact on the QoE to end-users.


(a) Highway - Original


(b) Highway - HoVe


(c) Highway - RSSI-based


(d) Highway - Strongest Cell


(e) Highway - SER


(f) Container - Original


(g) Container - HoVe

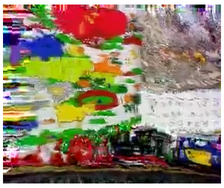

(h) Container - RSSI-based


(i) Container - Strongest Cell


(j) Container - SER


(k) Mobile - Original


(l) Mobile - HoVe


(m) Mobile - RSSI-based


(n) Mobile - Strongest Cell

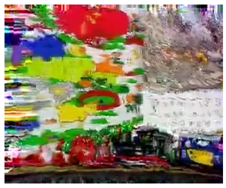

(o) Mobile - SER

Figure 12: Frames #111 from Highway, #213 from Container and #151 from Mobile

CHAPTER 5

A Migration Algorithm for Autonomous Vehicles

This Chapter details MOSAIC, a server tier-, computing resources-, QoS-, and migration time-oriented service migration and resource management algorithm for intra-tier and inter-tier communication in vehicular edge and fog computing. MOSAIC takes into account CAV's mobility information obtained using an Intelligent Transportation System; available computing resources in the fog and edge; delay in the candidate servers; and migration cost, in terms of time necessary for migration. The algorithm uses a multi-criteria decision-making scheme to select the best server, on edge or in a fog, to provide computing offloading for the CAV with low latency and high throughput.

5.1 MOSAIC Algorithm

MOSAIC assumes that service migrations in this scenario are preferably made in a live manner, *i.e.,* a service instance can only be interrupted after a copy of it has been transferred to the target edge server. This is achieved by performing live service pre-migrations based on user mobility data. MOSAIC assumes a scenario containing CAVs, cellular networks, and N-Tier fog and servers, such as presented by state-of-the-art works [37, 71].

5.1.1 Scenario Overview

The scenario is composed of a set of n CAV. Each CAV is assumed to have a radio transceiver to enable the communication between vehicles (V2V) and with an infrastructure (V2I). CAVs have a set of sensors that generate data with a rate of D; this data is later transmitted to the vehicle's serving cell through radio interfaces. In this sense, the scenario is composed of a set of cells, where each cell has an individual identity. We consider that each cell in the scenario is associated with an Edge server through a reliable point-to-point link. Closer to the network core, one tier above the edge servers in computing power, we consider a set of m fog servers. This is characterized by increased latency in comparison with the edge servers and increased computing capabilities. Additionally, we consider a centralized orchestrator in which MOSAIC is executed, such as the one presented in [29]. The orchestrator manages things such as the available resources in each edge server, schedules migrations, triggers the mobility management, and other tasks.

We assume an Intelligent Transportation System (ITS) aware of the vehicle's trajectories, both past trajectories, and planned routes. In this context, the orchestrator is aware of cell's coverage areas through a Voronoi Tessellation [13], and that handovers happen when a CAV leaves a coverage area for another, therefore it is possible to estimate when topology changes are going to happen.

Servers must perform computation tasks that are either part of a container, or a Virtual Machine (VM) instance. Servers can connect through a wired channel to perform the migration of Containers and VMs as expected in 5G and 6G networks. Also, edge and fog servers can communicate and perform data transfers between each other through the network.

Figure 13 show the systems' components in terms of the layers they belong to. We can define a set of five different layers that compose the system, namely, cloud layer, mobility management layer, fog layer, communication layer, and data source layer. Each layer is responsible for specific tasks: the data source layer performs sensing in general, the communication layer is responsible for wireless data transmissions, the fog layer provides scalability for the computing tasks, the cloud layer performs the management of the network, and the mobility management layer is an application layer connected to the system through the cloud layer [30].

5.1.2 Algorithm description

MOSAIC requests to the clients of each edge server, the latency and RTT values that they are experiencing, and a threshold based on the application are applied to define if the latency if low enough or not. To guarantee safety applications for CAV, a maximum latency of $10ms$ is necessary [61], thus this is the threshold applied by MOSAIC. The threshold can be adapted to different application requirements.

To execute a seamless pre-migration, our orchestrator can forecast imminent changes to network topology and recalculate the appropriate servers for each application under the new topology. For that, we consider that the orchestrator has the information about cell coverage areas and CAV trajectories, and can perform migrations proactively.

In current LTE networks, the network must manage session transfer and synchro-

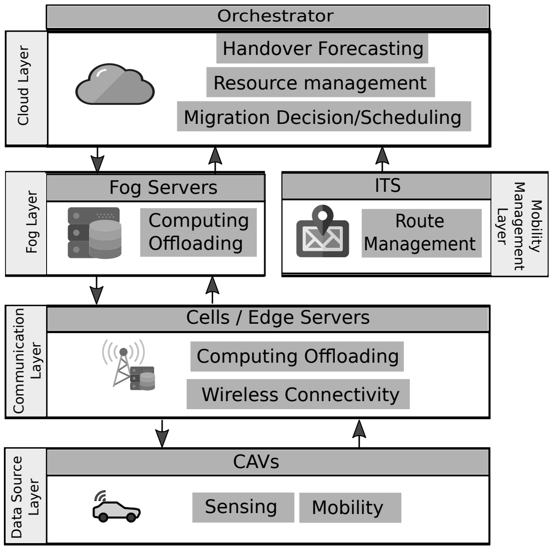


Figure 13: MOSAIC Scenario Components Overview

nization between the user and the new network. In an edge scenario, a service migration may also be necessary, as a handover means that the previous edge server may not be the best or closest.

As a general rule, every handover calls for a service migration in an edge scenario. However, cell coverage areas may overlap, providing multiple options of edge server and cells. It is also essential to check if the target edge server has enough resources for the application.

5.1.3 Migration Decision

For each server, MOSAIC keeps track of the following parameters at all times: tier of the server, meaning if it is a fog or an edge server; Available Resources free in the server; QoS, defined here as the average latency form the server to its connected CAVs, and estimated time to perform a migration to the server, based on backhaul link capacity, and the size of the migration. For the QoS and Migration Time parameters, which must be minimized, we consider the inverse of the numerical value of the parameter in the computation.

MOSAIC finds the best edge server given the collected metrics configuring a Multiple-Criteria Decision-Making problem. We chose AHP to balance the input metrics since it considers a pairwise comparison between the numerical values of each collected parameter and their relative degrees of importance, to adjust their weights of each parameter at runtime. The weights of the inputs must be defined when configuring the algorithm. High weight means more importance should be attached to this particular metric, and we define five importance levels, as shown in Table 5.

MOSAIC constructs for each vehicle a matrix to compare all pairs of metrics. We denote $c_{i,j}$ as how important the i^{th} element is compared with the j^{th} element. Also, $A =$

Table 5: Pairwise Context Importance

$c_{i,j}$	Definition
4	i is much more important than j
2	i is more important than j
1	i is as important as j
1/2	i is less important than j
1/4	i is much less important than j

$(C_{i,j})_{n\times n}$ represents the comparison matrix, where n denotes the number of elements to be compared, as shown in Eq. (5.1).

$$A = (C_{i,j})_{n\times n} = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_2 & c_3 & c_4 \end{array} \\ \begin{pmatrix} c_{1,1} & c_{1,2} & c_{1,3} & c_{1,4} \\ c_{2,1} & c_{2,2} & c_{2,3} & c_{2,4} \\ c_{3,1} & c_{3,2} & c_{3,3} & c_{3,4} \\ c_{4,1} & c_{4,2} & c_{4,3} & c_{4,4} \end{pmatrix} \end{array} \tag{5.1}$$

The metrics collected are placed in the matrix as Eq. 5.2 shows. Since the priority for MOSAIC is to keep a low latency for CAVs, the Tier parameter is the most important one, two times more important than the Resources, four times more important than the QoS, and also four times more important than the Migration Time. The rest of the comparisons are made to be coherent with these values, as shown in Eq. 5.2. We find the eigenvector of the matrix $(C_{i,j})$ by dividing each element by the sum of its column, obtaining the eigenvector [0.50 0.25 0.125 0.125], which means that Tier has weight 0.50, Resources have weight 0.25, 0.125 for the QoS, and 0.125 for the Migration Time. Note that the consistency ratio of the decision matrix is 0% (lower than 10% is acceptable). In this way, MOSAIC performs a product between the eigenvector and a vector that stores the measured values, obtaining the score for each available server.

$$A = (C_{i,j})_{nxn} = \begin{array}{c} \\ Tier \\ R \\ QoS \\ T \end{array} \begin{array}{c} \begin{array}{cccc} Tier & R & QoS & T \end{array} \\ \begin{pmatrix} 1 & 2 & 4 & 4 \\ 1/2 & 1 & 2 & 2 \\ 1/4 & 1/2 & 1 & 1 \\ 1/4 & 1/2 & 1 & 1 \end{pmatrix} \end{array} \rightarrow \tag{5.2}$$

$$\rightarrow [0.50\ 0.25\ 0.125\ 0.125]$$

Algorithm 1 presents how the migration process is triggered. The first decision of MOSAIC is whether migration is necessary of not. Migration may be necessary because of vehicle mobility, as it gets more distant from the server, and the latency increases, or for QoS reasons (a server is too loaded, for example).

Algorithm 2 shows how MOSAIC chooses the edge server to which the user sesison

Algorithm 1: MOSAIC monitor

```
while vehicle is connected do
    Consult mobility;
    if handover is eminent then
        MOSAIC_MIGRATION;
    Measure QoS;
    if QoS is below to the threshold then
        MOSAIC_MIGRATION;
```

Algorithm 2: MOSAIC_MIGRATION

```
Data: Number of Hops Acceptable
List available servers;
for Each available edge server do
    Get QoS for the server;
    Estimate the number of hops from user to server;
    Run AHP and give the server a score;
while Server has not been chosen do
    Get a server with the greatest score;
    Estimate migration time;
    if Migration can be done on time then
        Choose this server as target;
    else
        Remove this server from list;
```

will be migrated by the MOSAIC_ MIGRATION method. The essential characteristic of the decision is whether the target server can deliver the latency and computation requirements and, if so if the migration can be made promptly. MOSAIC assumes that each edge server can assess the link bandwidth from itself to other edge servers, and uses this bandwidth value to estimate the time it would take to migrate the user session to candidate edge servers. MOSAIC is relatively low complexity, as the AHP calculation is not an expensive operation. The complexity of the algorithm is proportional to the product of the number of CAVs and the number of Edge Servers.

As soon as MOSAIC detects that migration is necessary, the algorithm must evaluate all available servers in the vehicle routes (servers that would meet distance and latency requirements) regarding the server's resources and the cost to migrate the service to that specific server.

5.2 Evaluation

This section describes the evaluation methodology, including scenario description, simulation parameters, and metrics used to evaluate the performance of different migration algorithm for Edge-enabled CAV scenarios.

5.2.1 Scenario description and methodology

We implemented MOSAIC and other three existing migration algorithms by using NS-3.29[1] simulator, which implements the LTE protocol stack for V2I communication. In our scenario, we consider a 2km × 2km area comprising 20 ENodeBs randomly allocated, as well as 30, 60, and 90 CAV, respectively. The simulation considers the Nakagami path loss model, which can be very suitable for urban scenarios [55]. We conducted 33 simulations with different randomly generated seeds fed to the simulator's pseudo-random number generator (MRG32k3a). Results show the values with a confidence interval of 95%. Simulation parameters can be seen in Table 6.

For the simulation of traffic and vehicle mobility, we employed the BonnMotion[2], which is a topology generation and analysis tool. BonnMotion allows us to reproduce the desired vehicle movements with a predefined path and speeds based on empirical data. The CAVs move at different speeds as expected in real cities (ranging between 10-70 km/h).

Table 6: Main Simulation Parameters

Parameter	Value
Number of CAVs	[30, 60, 90]
Number of Cells	20
Average Speed of Vehicles	20 m/s
Mobilty Model	Manhattan Grid
Propagation Loss Model	Nakagami
Scenario Size	$2Km \times 2Km$
Downlink Frequency	2120 (MHz)
Uplink Frequency	1930 (MHz)
Duration of the simulation	120s
Size of VMs	10GB
Size of Containers	1GB

MOSAIC is tested against the other three service migration strategies. First, a scenario where no migrations happen at all; in this scenario, the edge server to which the CAV is initially connected keeps providing service until the end of the simulation. Then, a greedy strategy was also implemented, where the service is migrated to the new closest edge server after every handover. The last strategy, proposed by Li et al. [33], consists

[1]http://www.nsnam.org/
[2]https://sys.cs.uos.de/bonnmotion/

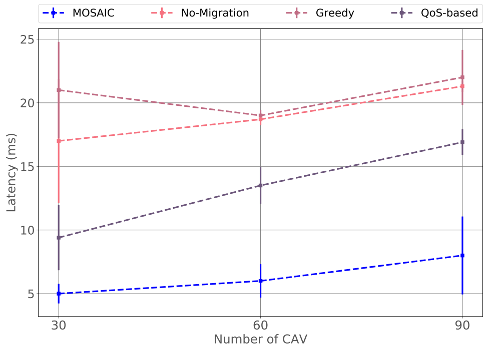


Figure 14: Average Latency for Different Number of Vehicles in a CAV scenario

of QoS-aware service migration, aiming to minimize latency. All services running in the network are randomly assigned to either a VM or a container, which influences the size of the task.

5.2.2 Results

We analyse the performance of MOSAIC compared to other state-of-the-art works from the perspective of some of the main metrics in service migration and CAV experiments: latency, probably the main metric in the scenario, providing the main requirements for the presence of CAVS; migrations attempted and failed, which have to do with the resources used in the network in consequence of the migrations executed; and monetary cost of the algorithm utilization, which is important to scale the solution in real-world scenarios.

Figure 2 shows the average latency each algorithm obtained in the 30, 60, and 90 CAV scenarios, respectively. We can see that only MOSAIC is able to maintain latency for its clients below 10ms in all of the scenarios. The QoS-based algorithm was able to achieve the minimum latency requirement only on the 30 CAV scenario. However, it was not able to keep up with the latency as more CAV entered the network and used up edge resources. We can see that the greedy algorithm has more unpredictable behavior due to trying to perform excessive migrations, resulting in a significant number of failures. The No-Migration approach resulted in steadily increasing latency as more CAVs were considered, as expected.

Live service migrations are a time-sensitive task, with strict deadlines. If migration is not finished in time (*i.e.,* the service is not in the edge server by the time the CAV has left its coverage area), the task is no longer useful. Migration failures take up backhaul resources and compromise QoS. Complementing the previous results, Figure 3 shows the network's average number of migration failures under each of the algorithms in the 30 CAV scenario. We can see that MOSAIC had the fewest migration failures out of all analyzed approaches (keeping it around 5 per simulation), except for No-Migration (which does not perform any migrations). This is a consequence of MOSAIC's careful resource management and mobility information. The QoS-based approach performed

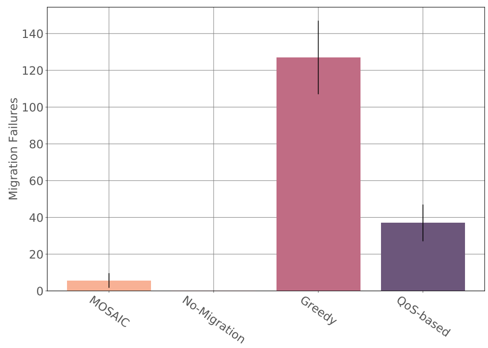


Figure 15: Migration Failures for Different Algorithms in a CAV scenario

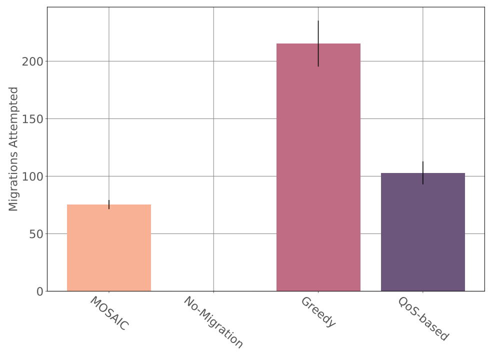


Figure 16: Number of Migrations Attempted by Each Algorithm

around 37 failed migrations, much fewer than the Greedy approach, in which around 120 migrations, on average, have failed per simulation, because this approach does not check for resources, or deadlines when performing migrations.

Figure 4 represents the average number of migrations attempted by each algorithm during the simulations. We can see a similar pattern of the one in Figure 3. We understand that MOSAIC had a failure rate of around 5%, the greedy approach had almost 52% of its migration attempts failing, and the QoS-based approach kept it around 27%. This behavior is explained by MOSAIC's mobility management module. When given the vehicle's trajectory information and resource information, the algorithm may be exact in its migration requests, avoiding unnecessary ones.

However, how MOSAIC does not perform any sort of cost management, potentially have higher monetary costs. Therefore, we use the AWS TCO Calculator[3] to calculate the financial cost of CPU time per hour. The CPU time cost proportionally decreases when renting a higher number of CPU cores in the same AWS region. We considered the deployment into two regions, corresponding to when the edge services are on the cell, which the CAV is directly connected to, and the case in which the service is more distant than the user. Costs for the tiers used are shown in Table 7.

Figure 5 shows the average monetary cost for the offloading of each vehicle to the edge under each of the algorithms tested, according to the values of Table 5. We can

[3]https://awstcocalculator.com/

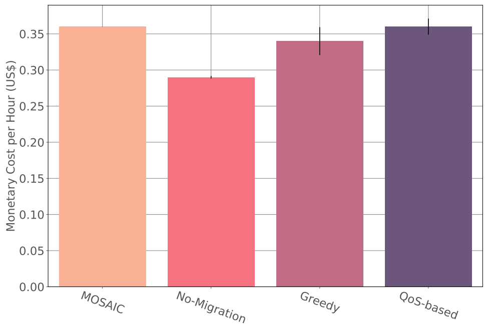


Figure 17: Monetary Cost per Vehicle per Hour Under Each Algorithm

see that MOSAIC, together with the QoS-based approach, have the highest monetary costs for their execution, since they, more often than not, keep their users in a tier 1 server. However, only MOSAIC is able to do so while maintaining acceptable QoS values. No-Migration has the lowest cost since most of its users have been allocated in a fog sever due to the lack of proper resource management. The greedy approach, however, had an inferior server utilization cost compared to MOSAIC. This is because although it always attempts to keep users in Edge servers, they fail, and users are allocated to cheaper fog servers.

Table 7: Monetary Cost of CPU Utilization

Node	CPU Cores	Memory	Storage	Cost/Hour
Tier 2	8	32 GB	2 TB	$0,22896
Tier 1	4	16 GB	1 TB	$0,42408

CHAPTER 6

Conclusions

As applied in previous generations of mobile communications, mobility management must be redesigned to support modern network deployment strategies. Such networks are increasingly dense and must support a wide range of services, such as real-time and low latency communications, with various requirements. In this context, user devices are more heterogeneous than ever and are incorporated into wearables, smartphones, tablets, and even vehicles.

In the case of vehicles, especially, the network must adapt to high mobility rates and velocities without disrupting the services being consumed by mobile users. Users in vehicles will consume a large amount of video-based content, which is expected to be delivered with QoE support. The first problem tackled in this dissertation is a handover algorithm for video transmission over vehicular networks with QoE support called HoVe. HoVe takes into account the estimated QoE value of the video flow, a user mobility prediction scheme, QoS, and signal intensity associated with wireless cells. Simulation results show an improvement of 18% in terms of QoE when HoVe is used, and an improvement of up to 30% in terms of packet delivery ratio.

However, vehicular scenarios bring a challenge in terms of computation as well. The heavy presence of Connected Autonomous Vehicles (CAVs) in future networks is expected, and In this dissertation, we propose a service migration algorithm for edge-enabled vehicular networks. In such a context, autonomous vehicles must offload computation to fog and edge servers to process vast amounts of data within tight latency requirements. We propose the MOSAIC, a migration decision algorithm for CAVs in edge-enabled scenarios capable of choosing the best edge servers to provide computation offloading for CAV. onsiders server tier, computing resources, QoS, and migration time offer low latency, and high throughput using predictively migrating CAVs offload instances to their future servers. Based on simulation results, we show that chieves superior performance in terms of throughput, latency, and migration failures compared to state-of-the-art algorithms.

6.1 Future Works

The mobility management field spans over several areas in computing and network management. We list some of the possible directions for research as follows:

1. Migration of heterogeneous services in smart city scenarios: smart cities are a great market driver and characterized by the heady presence of heterogeneous devices, networks, and services. Managing the location of said serviced in an edge-enabled smart city is a challenging and critical task.
2. Group mobility assessment: humans don't often move through a scenario randomly, but rather with underlying motivations that may be common to groups of different people. Analyzing group mobility patterns can improve network management and service provisioning.
3. Semantic trajectory searches: This research topic consists of finding complementary information to user mobility, beyond geographical coordinates. Things such as weather, health status, social media posts, and local events can be assessed and correlated to user mobility for complete information from a prediction and management point-of-view.

6.2 Published Works

The main results published in this dissertation were published in the following works:

1. [51] - **PACHECO, L**.; ROSÁRIO, D; CERQUEIRA, E; VILLAS, L. Service migration in edge computing environments for connected autonomous vehicles. *SBRC*, SBC, 2020.
2. [53] - **PACHECO, L**. OLIVEIRA, H; ROSÁRIO, D; CERQUEIRA, E; BRAUN, T. Service migration in edge computing environments for connected autonomous vehicles. *ISCC*, IEEE, 2020.
3. [50] - **PACHECO, L**.; MEDEIROS, I; SANTOS, H; OLIVEIRA, H.; ROSÁRIO, D; CERQUEIRA, E; NETO, A. A handover algorithm for video sharing over vehicular networks. *LADC*, IEEE, 2019.
4. [52] - **PACHECO, L**.; ROSÁRIO, D; CERQUEIRA, E; VILLAS, L; BRAUN, T. Service Migration in Heterogeneous Smart City Networks networks. *MSWIM*, ACM, 2019.

Complementary results were published in various other works, listed below:

4. [20] COSTA, A.; **PACHECO, L**; ROSARIO, D; CERQUEIRA, E; VILLAS, L; SARGENTO, S; LOUREIRO, A. Skipping-based handover algorithm for video distribution over ultra-dense vanet. *Computer Networks*, Elsevier, 2020.

5. [42] - MATNI, N.; MORAIS, J.; **PACHECO, L**; OLIVEIRA, H.; ROSARIO, D; CERQUEIRA, E; NETO, A Experimenting long-range wide area network in an e-health environment: Discussion and future directions. *IWCMC*, IEEE, 2020.

6. [75] - ZHAO, Z.; KARIMZEDEH, M; **PACHECO, L**; SANTOS, H; ROSÁRIO, D; CERQUEIRA, E; BRAUN, T. Edge-based Mobility Management with Transferable Reinforcement Learning Trajectory Prediction. *TNSM*, IEEE, 2020.

References

1 Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper, Mar 2017. 33

2 3GPP. Evolved universal terrestrial radio access (e-utra); radio resource control (rrc); protocol specification. *3GPP TS 36.331 V9.4.0 (2010-09), Evolved Universal Terrestrial Radio Access (E-UTRA);Radio Resource Control (RRC);Protocol specification(Release 9)*, 2011. 53

3

4 Rami Ahmad, Elankovan A Sundararajan, Nor E Othman, and Mahamod Ismail. Handover in lte-advanced wireless networks: state of art and survey of decision algorithm. *Telecommunication Systems*, 66(3):533–558, 2017. 21

5 Noura Aljeri and Azzedine Boukerche. Fog-enabled vehicular networks: A new challenge for mobility management. *Internet Technology Letters*, 2019. 22

6 Noura Aljeri and Azzedine Boukerche. Movement prediction models for vehicular networks: an empirical analysis. *Wireless Networks*, 25(4):1505–1518, 2019. 21

7 Noura Aljeri and Azzedine Boukerche. A two-tier machine learning-based handover management scheme for intelligent vehicular networks. *Ad Hoc Networks*, 94:101930, 2019. 21

8 Alaa A.R. Alsaeedy and Edwin K.P. Chong. Mobility Management for 5G IoT Devices: Improving Power Consumption with Lightweight Signaling Overhead. *IEEE Internet of Things Journal*, 6(5):8237–8247, 2019. 18

9 Antonios Argyriou, Konstantinos Poularakis, George Iosifidis, and Leandros Tassiulas. Video delivery in dense 5g cellular networks. *IEEE Netw*, 31:28–34, 2017. 20, 21, 53

10 Rabe Arshad, Hesham ElSawy, Sameh Sorour, Tareq Y Al-Naffouri, and Mohamed-Slim Alouini. Handover management in 5g and beyond: A topology aware skipping approach. *IEEE Access*, 4:9073–9081, 2016. 21

11 Rabe Arshad, Hesham ElSawy, Sameh Sorour, Tareq Y Al-Naffouri, and Mohamed-Slim Alouini. Velocity-aware handover management in two-tier cellular networks. *IEEE Transactions on Wireless Communications*, 16(3):1851–1867, 2017. 21, 41, 42

12 Martin Aruldoss, T Miranda Lakshmi, and V Prasanna Venkatesan. A survey on multi criteria decision making methods and its applications. *American Journal of Information Systems*, 1(1):31–43, 2013. 32, 33

13 François Baccelli and Bartłomiej Błaszczyszyn. On a coverage process ranging from the boolean model to the poisson–voronoi tessellation with applications to wireless communications. *Advances in Applied Probability*, 33(2):293–323, 2001. 28, 59

14 Hamzeh Beyranvand, Martin Lévesque, Martin Maier, Jawad A Salehi, Christos Verikoukis, and David Tipper. Toward 5g: Fiwi enhanced lte-a hetnets with reliable low-latency fiber backhaul sharing and wifi offloading. *IEEE/ACM Transactions on Networking*, 25(2):690–707, 2017. 19

15 Yuanguo Bi, Guangjie Han, Chuan Lin, Qingxu Deng, Lei Guo, and Fuliang Li. Mobility support for fog computing: An sdn approach. *IEEE Communications Magazine*, 56(5):53–59, 2018. 23

16 Luiz Fernando Bittencourt, Márcio Moraes Lopes, Ioan Petri, and Omer F Rana. Towards virtual machine migration in fog computing. In *10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*, pages 1–8. IEEE, 2015. 23

17 Jiamei Chen, Yao Wang, Yufeng Li, and Ershen Wang. Qoe-aware intelligent vertical handoff scheme over heterogeneous wireless access networks. *IEEE Access*, 6:38285–38293, 2018. 42

18 Min Chen, Wei Li, Giancarlo Fortino, Yixue Hao, Long Hu, and Iztok Humar. A dynamic service migration mechanism in edge cognitive computing. *ACM Transactions on Internet Technology*, 19(2):30, 2019. 43, 44

19 Yuh-Shyan Chen and Yi-Ting Tsai. A mobility management using follow-me cloud-cloudlet in fog-computing-based rans for smart cities. *Sensors*, 18(2):489, 2018. 23

20 Allan Costa, Lucas Pacheco, Denis Rosário, Eduardo Cerqueira, Leandro Villas, Suzana Sargento, and Antonio Loureiro. Skipping-based handover algorithm for video distribution over ultra-dense vanet. *Computer Networks*, 2020. 69

21 Joahannes Costa, Denis Rosário, Allan M de Souza, Leandro A Villas, and Eduardo Cerqueira. Data dissemination based on complex networks' metrics for distributed traffic management systems. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 01062–01067. IEEE, 2018. 21

22 Rodolfo W.L. Coutinho and Azzedine Boukerche. Guidelines for the Design of Vehicular Cloud Infrastructures for Connected Autonomous Vehicles. *IEEE Wireless Communications*, 26(4):6–11, 2019. 20, 22

23 Eleni Demarchou, Constantinos Psomas, and Ioannis Krikidis. Mobility Management in Ultra-Dense Networks: Handover Skipping Techniques. *IEEE Access*, 6:11921–11930, 2018. 19, 21, 27, 41, 42

24 Zhipeng Gao, Jie Meng, Qian Wang, and Yang Yang. Service migration for deadline-varying user-generated data in mobile edge-clouds. *IEEE World Congress on Services, SERVICES 2018*, pages 53–54, 2018. 44

25 Ed Garsten. Sharp growth in autonomous car market value predicted but may be stalled by rise in consumer fear. *Forbes*, Aug 2018. 20

26 Fanyu Gong, Ziwei Sun, Xiaodong Xu, Zhao Sun, and Xiaosheng Tang. Cross-tier handover decision optimization with stochastic based analytical results for 5g heterogeneous ultra-dense networks. In *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6. IEEE, 2018. 27, 40, 42

27 Xinjie Guan, Xili Wan, Jun-bo Wang, Xinxin Ma, and Guangwei Bai. Mobility aware partition of mec regions in wireless metropolitan area networks. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–2. IEEE, 2018. 21

28 Hannes Hartenstein and LP Laberteaux. A tutorial survey on vehicular ad hoc networks. *IEEE Communications magazine*, 46(6):164–171, 2008. 26

29 Jinho Hwang, Yun-Wu Huang, Maja Vukovic, and Nikos Anerousis. Enterprise-scale cloud migration orchestrator. In *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 1002–1007. IEEE, 2015. 59

30 Razi Iqbal, Talal Butt, Omair Shafiq, Manar Talib, and Tariq Umar. Context-aware data-driven intelligent framework for fog infrastructures in internet of vehicles. *IEEE Access*, PP:1–1, 10 2018. 59

31 Hyoungju Ji, Sunho Park, Jeongho Yeo, Younsun Kim, Juho Lee, and Byonghyo Shim. Ultra-reliable and low-latency communications in 5g downlink: Physical layer aspects. *IEEE Wireless Communications*, 25(3):124–130, 2018. 19

32 Yejee Lee, Bongjhin Shin, Jaechan Lim, and Daehyoung Hong. Effects of time-to-trigger parameter on handover performance in son-based lte systems. In *2010 16th Asia-Pacific Conference on Communications (APCC)*, pages 492–496. IEEE, 2010. 40

33 Jun Li, Xiaoman Shen, Lei Chen, Dung Pham Van, Jiannan Ou, Lena Wosinska, and Jiajia Chen. Service Migration in Fog Computing Enabled Cellular Networks to Support Real-Time Vehicular Communications. *IEEE Access*, 7:13704–13714, 2019. 23, 63

34 Xin Li, Yifan Dang, and Tefang Chen. Vehicular Edge Cloud Computing: Depressurize the Intelligent Vehicles Onboard Computational Power. *IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2018-Novem:3421–3426, 2018. 43, 44

35 Siyi Liao, Jianhua Li, Jun Wu, Wu Yang, and Zhitao Guan. Fog-enabled vehicle as a service for computing geographical migration in smart cities. *IEEE Access*, 7:8726–8736, 2019. 43, 44

36 Qianyu Liu, Chiew Foong Kwong, Sibo Zhang, Lincan Li, and Jing Wang. A fuzzy-clustering based approach for madm handover in 5g ultra-dense networks. *Wireless Networks*, pages 1–14, 2019. 41, 42

37 Shaoshan Liu, Liangkai Liu, Jie Tang, Bo Yu, Yifan Wang, and Weisong Shi. Edge computing for autonomous driving: Opportunities and challenges. *Proceedings of the IEEE*, 107(8):1697–1716, 2019. 22, 23, 58

38 Huimin Lu, Qiang Liu, Daxin Tian, Yujie Li, Hyoungseop Kim, and Seiichi Serikawa. The Cognitive Internet of Vehicles for Autonomous Driving. *IEEE Network*, 33(3):65–73, 2019. 22

39 Guiyang Luo, Quan Yuan, Haibo Zhou, Nan Cheng, Zhihan Liu, Fangchun Yang, and Xuemin Sherman Shen. Cooperative vehicular content distribution in edge computing assisted 5G-VANET. *China Communications*, 15(7):1–17, 2018. 21

40 Redowan Mahmud, Ramamohanarao Kotagiri, and Rajkumar Buyya. Fog computing: A taxonomy, survey and future directions. In *Internet of everything*, pages 103–130. Springer, 2018. 23

41 Stefano Mariani, Giacomo Cabri, and Franco Zambonelli. Coordination of Autonomous Vehicles: Taxonomy and Survey. 2020. 37

42 Nagib Matni, Jean Moraes, Lucas Pacheco, Denis Rosario, Helder Oliveira, Eduardo Cerqueira, and Augusto Neto. Experimenting long range wide area network in an e-health environment: Discussion and future directions. *IWCMC*, 2020. 69

43 Iago Medeiros, Lucas Pacheco, Denis Rosário, Cristiano Both, Jéferson Nobre, Eduardo Cerqueira, and Lisandro Granville. Quality of experience and quality of service-aware handover for video transmission in heterogeneous networks. *International Journal of Network Management*, page e2064, 2018. 41, 42

44 Iago Medeiros, Lucas Pacheco, Denis Rosário, Cristiano Both, Jéferson Nobre, Eduardo Cerqueira, and Lisandro Granville. Quality of experience and quality of service-aware handover for video transmission in heterogeneous networks. *International Journal of Network Management*, 0(0):e2064. e2064 nem.2064. 49, 53

45 Abdelrahim Mohamed, Oluwakayode Onireti, Seyed Amir Hoseinitabatabaei, Muhammad Imran, Ali Imran, and Rahim Tafazolli. Mobility prediction for handover management in cellular networks with control/data separation. In *2015 IEEE International Conference on Communications (ICC)*, pages 3939–3944. IEEE, 2015. 21

46 F. Mohammadnia, C. Vitale, M. Fiore, V. Mancuso, and M. A. Marsan. Mobile small cells for adaptive ran densification: Preliminary throughput results. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–7, April 2019. 19

47 Emmanuel Ndashimye, Sayan K Ray, Nurul I Sarkar, and Jairo A Gutiérrez. Vehicle-to-infrastructure communication over multi-tier heterogeneous networks: A survey. *Computer networks*, 112:144–166, 2017. 20, 21

48 James Nightingale, Pablo Salva-Garcia, Jose M Alcaraz Calero, and Qi Wang. 5g-qoe: Qoe modelling for ultra-hd video streaming in 5g networks. *IEEE Transactions on Broadcasting*, 64(2):621–634, 2018. 18

49 Tao Ouyang, Zhi Zhou, and Xu Chen. Follow Me at the Edge: Mobility-Aware Dynamic Service Placement for Mobile Edge Computing. *IEEE Journal on Selected Areas in Communications*, 36(10):2333–2345, 2018. 43, 44

50 Lucas Pacheco, Iago Medeiros, Hugo Santos, Helder Oliveira, Denis Rosario, and Eduardo Cerqueira. A handover algorithm for video sharing over vehicular networks. *LADC*, 2019. 68

51 Lucas Pacheco, Denis Rosario, Eduardo Cerqueira, and Leandro Villas. Service migration in edge computing environments for connected autonomous vehicles. *SBRC*, 2020. 68

52 Lucas Pacheco, Denis Rosario, Leandro Villas, Eduardo Cerqueira, and Torsten Braun. Service migration in heterogeneous smart city networks. *MSWIM*, 2020. 68

53 Lucas Pacheco, Helder Rosario, Denis Oliveira, Eduardo Cerqueira, and Torsten Braun. Service migration for connected autonomous vehicles. *ISCC*, 2020. 68

54 Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. CRAWDAD dataset epfl/mobility (v. 2009-02-24). Downloaded from ⟨https://crawdad.org/epfl/mobility/20090224⟩, February 2009. 49

55 Lorenzo Rubio, Juan Reig, and Narcís Cardona. Evaluation of nakagami fading behaviour based on measurements in urban scenarios. *AEU-International Journal of Electronics and Communications*, 61(2):135–138, 2007. 63

56 Lambros Sarakis, Theofanis Orphanoudakis, Helen C Leligou, Stamatis Voliotis, and Artemis Voulkidis. Providing entertainment app lications in vanet environments. *IEEE Wireless Communications*, 23(1):30–37, 2016. 26, 27

57 Susana Sargento, André Cardote, Carlos Ameixieira, Filipe Neves, and Jorge Dias. Method and apparatus for multi-network communication in vehicular networks. *US Patent US9439121 B2*, 2016. 42

58 Elmar Schoch, Frank Kargl, Michael Weber, and Tim Leinmuller. Communication patterns in vanets. *IEEE Communications Magazine*, 46(11):119–125, 2008. 27

59 Syed Adeel Ali Shah, Ejaz Ahmed, Muhammad Imran, and Sherali Zeadally. 5G for Vehicular Communications. *IEEE Communications Magazine*, 56(1):111–117, 2018. 18

60 Ketyllen Da Costa Silva, Zdenek Becvar, and Carlos Renato Lisboa Frances. Adaptive hysteresis margin based on fuzzy logic for handover in mobile networks with dense small cells. *IEEE Access*, 6:17178–17189, 2018. 40, 42

61 Meryem Simsek, Adnan Aijaz, Mischa Dohler, Joachim Sachs, and Gerhard Fettweis. The 5G-Enabled Tactile Internet: Applications, requirements, and architecture. *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pages 61–66, 2016. 38, 59

62 Zhenjie Tan, Hua Qu, Jihong Zhao, Shiyu Zhou, and Wenjie Wang. Uav-aided edge/fog computing in smart iot community for social augmented reality. *IEEE Internet of Things Journal*, 2020. 36

63 Luca Tartarini, Marcelo Antonio Marotta, Eduardo Cerqueira, Juergen Rochol, Cristiano Bonato Both, Mario Gerla, and Paolo Bellavista. Software-defined handover decision engine for heterogeneous cloud radio access networks. *Computer Communications;*, 115:21 – 34, 2018. 51

64 Jeffrey C Ulin. *The business of media distribution: Monetizing film, TV, and video content in an online world*. Routledge, 2019. 20

65 Ricard Vilalta, Selva Vía, Fermín Mira, Ramon Casellas, Raul Muñoz, Jesus Alonso-Zarate, Apostolos Kousaridas, and Markus Dillinger. Control and management of a connected car using sdn/nfv, fog computing and yang data models. In *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, pages 378–383. IEEE, 2018. 23

66 Heng Wang, Bin Wang, Bingbing Liu, Xiaoli Meng, and Guanghong Yang. Pedestrian recognition and tracking using 3d lidar for autonomous vehicle. *Robotics and Autonomous Systems*, 88:71–78, 2017. 37

67 Shangguang Wang, Jinliang Xu, Ning Zhang, and Yujiong Liu. A survey on service migration in mobile edge computing. *IEEE Access*, 6:23511–23528, 2018. 37

68 Xiaodong Xu, Xiaoxuan Tang, Zhao Sun, Xiaofeng Tao, and Ping Zhang. Delay-Oriented Cross-Tier Handover Optimization in Ultra-Dense Heterogeneous Networks. *IEEE Access*, 7:21769–21776, 2019. 41, 42

69 Ilsun You, Yuh Shyan Chen, Sherali Zeadally, and Fei Song. A brief overview of intelligent mobility management for future wireless mobile networks. *Eurasip Journal on Wireless Communications and Networking*, 2017(1), 2017. 19

70 Ashkan Yousefpour, Caleb Fung, Tam Nguyen, Krishna Kadiyala, Fatemeh Jalali, Amirreza Niakanlahiji, Jian Kong, and Jason P. Jue. All one needs to know about fog computing and related edge computing paradigms: A complete survey. *Journal of Systems Architecture*, 98(December 2018):289–330, 2019. 35

71 Xiang Yu, Maolin Guan, Mingxia Liao, and Xia Fan. Pre-Migration of Vehicle to Network Services Based on Priority in Mobile Edge Computing. *IEEE Access*, 7:3722–3730, 2019. 43, 44, 58

72 Di Zhang, Yuezhi Zhou, Xiang Lan, Yaoxue Zhang, and Xiaoming Fu. Aht: Application-based handover triggering for saving energy in cellular networks. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9. IEEE, 2018. 42

73 Ke Zhang, Yuming Mao, Supeng Leng, Yejun He, and Yan Zhang. Predictive Offloading in Cloud-Driven Vehicles: Using Mobile-Edge Computing for a Promising Network Paradigm. *IEEE Vehicular Technology Magazine*, 12(2):36–44, 2017. 23

74 Zhongliang Zhao, Lucas Guardalben, Mostafa Karimzadeh, José Silva, Torsten Braun, and Susana Sargento. Mobility prediction-assisted over-the-top edge prefetching for hierarchical vanets. *IEEE Journal on Selected Areas in Communications*, 36(8):1786–1801, 2018. 21

75 Zhongliang Zhao, Mostafa Karimzadeh, Lucas Pacheco, Hugo Santos, Denis Rosario, Torsten Braun, and Eduardo Cerqueira. Edge-based mobility management with transferable reinforcement learning trajectory prediction. *IEEE TNSM*, 2020. 69

76 Kan Zheng, Xiaoli Zhang, Qiang Zheng, Wei Xiang, and Lajos Hanzo. Quality-of-experience assessment and its application to video services in lte networks. *IEEE Wireless Communications*, 22(1):70–78, 2015. 52