Optimising End User QoE for DASH Enabled Video Streams Over Bandwidth Constrained Links in Urban HetNets

Timothy Casey, MSc.

A Dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

Dublin City University

School of Electronic Engineering

Supervisor: Dr. Gabriel-Miro Muntean

September 2020

DECLARATION

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____

ID No.: **13211188**

Date: **10-09-2020**

To my wife Avilene and sons Luke, Kieran and Jake

ACKNOWLEDGEMENTS

The support by the Institute of Technology Carlow through their PhD funding program for staff is gratefully acknowledged.

Firstly, I would like to thank my PhD supervisor Dr. Gabriel-Miro Muntean for his guidance, understanding and advice throughout this entire research process.

To my colleague and friend Caroline Byrne for her continuous encouragement throughout it all.

Last but by no means least a very special message of thanks to my wife Avilene for her unwavering support over the years while working fulltime to progress her own career and projects.

Kilkenny, August 2020

Tim Casey

LIST OF PUBLICATIONS

T. Casey, G-M. Muntean, “Delivery of High Definition Video Content over Bandwidth Con-strained Links in Heterogeneous Wireless Networks”, IEEE International Symposium on Broad-band Multimedia Systems and Broadcasting (BMSB), June, 2019

T. Casey, G-M. Muntean, “Reducing Stalling Events during DASH Video Playback in Heterogeneous Multi-Network Wireless Environments”, IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), June, 2017

T. Casey, G-M. Muntean, “MPEG-DASH-based Framework for Improving End-user Video Experience in Heterogeneous Multi-Network Wireless Environments”, IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), June, 2016

T. Casey, G-M. Muntean, “Scan-Or-Not-To-Scan – Balancing Network Selection Accuracy and Energy Consumption”, IEEE International Wireless Communications and Mobile Computing Conference (IWCMC), August, 2015

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

LIST OF ABBREVIATIONS

A

ABR: Adaptive Bit Rate

AIS: Adaptive Interface Selection

AMPS: Advanced Mobile Phone Service

AP: Access Point (Wi-Fi)

AUC: Authentication Center

B

BSA: Basic Service Area

BSC: Base Station Controller

BSS: Base Station Subsystem

BTS: Base Transceiver Station

C

CAPEX: Capital Expenditure

CBR: Constant Bit Rate

CLS: clear-to-send

CS: Circuit Switched

CSMA/CA: Carrier Sense Multiple Access/Collision Avoidance

D

DASH: Dynamic Adaptive Streaming over HTTP

DHCP: Dynamic Host Configuration Protocol

DL: Down link

E

EIR: Equipment Identity Register

EE: Energy Efficiency

EHT: Extremely High Throughput

EPC: Evolved Packet Core

EPS: Evolved Packet System

F

FCC: US Federal Communications Commission

FDM: Frequency Division Multiplexing

G

3GPP: Third Generation Partnership Project

GB: Gigabyte

GERAN: GSM EDGE Radio Access Network

GGSN: Gateway GPRS Support Node

GMSC: Gateway Mobile Switching Center

GPRS: General Packet Radio Services

GSN: GSM Support Nodes

GSM: Global System for Mobile Communication

H

HAS: Http Adaptive Streaming

HetNet: Heterogeneous Network

HD: High Definition

HLR: Home Location Register

I

IBSS: Independent Basic Service Set

IEEE: Institute of Electrical and Electronics Engineers

IF: Influence Factor

IoT: Internet of Things

IP: Internet Protocol

IR: Infrared

ISC: International Switching Center

ISM: Industrial Scientific Medical

ITU: International Telecommunications Union

IWF: Internetworking Function

L

LA: Location Area

LAN: Local Area Network

LTE: Long Term Evolution

M

M-2-M: Machine-to-Machine

MAN: Metropolitan Area Network

MCT: Machine Type Communications

MDF: MPEG-DASH based Framework

MGW: Media Gateway

MIH: Media Independent Handover

MIMO: Multiple-input Multiple-output

MPD: MPEG-DASH Media Presentation Description

MPEG: Moving Picture Expert Group

mps: metres per second

MS: Mobile Station

MSC: Mobile Switching Center

MSE: Media Source Extensions

MU-MIMO: Multiple User Multiple-input Multiple-output

N

NMT: Nordic Mobile Telephony

NS3: network simulation software

NSS: Network Switching Subsystem

O

OFDM: Orthogonal Frequency Division Multiplexing

OFDMA: Orthogonal Frequency Division Multiple Access

OMC: Operation and Maintenance Center

OMSS: Operation and Maintenance Subsystem

OTT: over-the-top

OPEX: Operational Expenditure

P

PAR: Project Authorization Request

PDN: Packet Data Network

PDP: Packet Data Protocol

PoA: Point of Attachment

PS: Packet Switched

PSTN: Packet Switched Telephone Network

Q

QAM: Quadrature Amplitude Modulation

QoE: Quality of Experience

QoS: Quality of Service

R

RAN: Radio Access Network

RAT: Radio Access Technology

RDC: Remaining Data Cap

RDM: Resolution Discovery Module

RSSI: Received Signal Strength Indication

RTCP: RTP Control Protocol

RTP: Real-time Transport Protocol

RTS: request-to-send

RTSP: Real-time Streaming Protocol

RNC: Radio Network Controller

S

SAE: System Architecture Evolution

SE: Spectrum Efficiency

SGSN: Serving GPRS Support Node

SINR: Signal to Interference plus Noise Ratio

SMS: Short Message Service

SoC: System on a Chip

SON: Self-Organising Network

SONS: Scan-Or-Not-to-Scan

T

TCP: Transmission Control Protocol

TDM: Time Division Multiplexing

TSN: Time-Sensitive Networking

U

UDP: User Datagram Protocol

UE: User Equipment

UL: Uplink

UMTS: Universal Mobile Communications System

URL: Uniform Resource Locator

UTRAN: UMTS terrestrial radio access network

V

VLC: Video Lan Client

VLR: Visited Location Register

VoIP: Voice over IP

W

W-CDMA: Wideband Code Division Multiple Access

WLAN: Wireless Local Area Network

WIPS: Wireless Intrusion Protection Switching

ABSTRACT

Timothy Casey

Optimising End-user QoE for DASH Enabled Video Streams over Bandwidth Constrained Links in Urban HetNets

Mobile users in urban HetNet environments contend with reduced Quality of Experience (QoE) while streaming video due to fluctuations in available bandwidth. Adaptive Bit Rate technologies have been developed to deal with changes in wireless link conditions but they do not address the problem of streaming HD video over bandwidth constrained links. Service providers impose data-caps on pay-as-you-go accounts which is problematic since much of the downloaded data consists of video files and breaching the data-cap can result in degraded performance and high excess usage charges.

User devices are typically dual-homed having both a cellular interface and a Wi-Fi interface although some devices may be equipped with additional interfaces such as Bluetooth Low Energy. Urban HetNets provide users with the opportunity to switch to a network that best suits their needs. Users seek to protect their data allowances by downloading as little data as possible over cellular networks by connecting to Wi-Fi whenever possible. However, under certain conditions it can be detrimental to the user's QoE to attempt to connect to Wi-Fi APs as the connection periods might be so short that little or no data can be downloaded or the connection attempt might fail. In addition, mobile devices are constrained by their battery capacity and it is important that energy consumption be reduced where possible.

This thesis presents a solution for managing video streaming in urban HetNets for mobile users, protecting their data-caps, optimising QoE and reducing energy consumption. The proposed solution consists of the following three novel components:

a) the **Scan-Or-Not-to-Scan (SONS) framework** containing the SONS utility function
b) **MPEG-DASH-based Framework (MDF)** which manages SD video streams
c) **Adaptive Interface Selection (AIS)** which manages HD video streams over bandwidth constrained wireless links

The proposed components were evaluated using modelling, simulation and hybrid test solutions consisting of Linux Virtual Machines and Network Simulator 3 networks. The results showed that the proposed solution can reduce the amount of data transferred over cellular networks, maintain QoE of SD video streams by reducing the number of stalling events and enable streaming of HD video over bandwidth constrained links.

CHAPTER 1 INTRODUCTION

This chapter documents the growth of video traffic on the Internet and introduces the challenges faced by urban commuters who wish to both protect their data allowances and maintain their QoE while streaming video. Following this, the problem statement for this thesis is outlined and the solution proposed here is presented. The novel contributions of the solution in this thesis are introduced, the thesis objectives are outlined and the thesis structure is detailed.

1.1 Research Motivation

Modern day commuters in urban areas have a presumption of being able to connect to the Internet whenever they want, wherever they are located. They carry sophisticated mobile devices such as smartphones or tablet computers and frequently have access to high capacity, low latency communication networks. During their daily commutes they connect to the World Wide Web for news, weather, social media and entertainment purposes. Much of the online content that they consume during their journeys consists of streamed video or other multi-media materials and they expect their Quality of Experience (QoE) while viewing this type of content to be high.

However, reality frequently does not match their expectations. Mobile users wishing to enjoy a good QoE when consuming video content on the go face multiple challenges such as:

- Uneven or patchy mobile broadband coverage
- Fluctuations in available bandwidth
- Data caps implemented by the service providers
- Limited battery capacity

Recent years have seen spectacular growth in the amount of data consumed by mobile users [1]. This growth has been driven by increased mobile device capabilities combined with improved cellular network coverage which enables users to view high definition (HD) video, play online games, and stay connected to the Internet almost anywhere at any time. Subscriber Quality of Experience (QoE) requirements and the user's expectation that they

can always be connected has driven mobile service providers to upgrade their networks to increase network speed, network capacity and network coverage area.

In many developed countries effective mobile broadband is a reality, but availability can be uneven with coverage concentrated in areas having the highest population densities. For example, during the first 3 months of 2019 smartphone users in Germany could on average access the Internet over 4G networks approximately 82% of the time while in urban areas [2]. Mobile phone users in rural areas of Germany on the other hand could only connect to 4G networks 73.5% of the time, these rural dwellers spent over a quarter of their time connected to 2G and 3G networks.

The rapid adoption of powerful smartphones capable of displaying high definition video content has had a profound impact on both communication networks and the types of traffic that flow across them. Video content that requires high capacity networks for successful delivery has become the dominant traffic type and will continue to grow into the future. A current report from Cisco [1] estimates that almost 79% of the world's mobile data traffic will be video by 2022 a nine-fold increase from 2017. It is also estimated that 69% of all data traffic originating on mobile networks will be offloaded to Wi-Fi networks by 2022.

However, merely increasing raw network speeds does not guarantee that the end user will have a satisfactory QoE when consuming video content. Opensignal's 2019 report [3] analysed the consumer's mobile video experience in 69 countries worldwide. Measurements were taken from approximately 8 million devices over a period of 4 months using a dedicated mobile phone application. The report highlighted the fact that the relationship between network speed and end-user QoE is complicated. Analysis showed that when network speeds are relatively low QoE and connection speed are closely related but once network speeds increased above an average download speed of 15 Mbps the speed of the connection had little impact on the quality of the streamed video playback. For example, of the 69 countries examined South Korea had the fastest average connection speed but ranked 21 when it came to video experience.

Clearly pure connection speed does not determine user QoE in countries having the fastest network speeds. Latency and consistency of connection speed are important; very high connection speeds are not necessary to stream video content over mobile networks, but a video stream requires consistency of connection in order to avoid stalling events and low

latency to avoid excessive delays in beginning playback. A network delivering 40 Mbps connection speeds one moment and a 3 Mbps connection the next will provide a less satisfactory overall experience when streaming video than a network capable of delivering a constant connection speed of 15 Mbps. It is extremely difficult to maintain consistent download speeds in mobile networks due to their shared capacity. Fluctuations in connection speed arise from frequent changes in the number of connected users and the demands that they place on the shared resource. For instance, smartphone users in cities in Australia [4] can experience a 15 Mbps reduction in download speeds over 4G networks depending on the time of day.

In the context of end user video experience only 11 of the 69 countries surveyed in [3] achieved a rating of Very Good, none achieved a rating of Excellent, 55 achieved a rating of Fair or Good and 3 countries ranked as poor. These results mean that for most of the world the mobile video experience is in need of improvement. In many countries video load times are slow, stalling events are common to varying degrees and connections struggle with high definition (HD) video formats.

The improvements in mobile device capabilities brought about by ever more powerful CPUs, graphics cards and high definition screens have taken place hand in hand with consumer demands for lighter and thinner physical devices. This trend towards lighter and thinner devices has resulted in lighter and thinner battery packs. The reduction in battery size coupled with the development of ever more power-hungry system components and applications has led to a situation in which users rarely see a full day's device operation on a single battery charge.

Many organisations and industry leaders forecast huge increases in the amount of network traffic due to the demand for and delivery of video and multi-media content. Cisco [1] predicts that the average smartphone user will generate 11 GB of data per month by 2022, however as of August 2018 the top 10% of users consumed 45 GB of data per month and 46% of users generated more than 10 GB of data per month. It is also predicted that rapid and ubiquitous Internet of Things (IoT) development and deployment will help drive this increase in network traffic.

The demand for video content has placed mobile networks under severe strain and in an attempt to control the amount of data consumed by user's service providers impose 'data

caps'. Data caps are limits placed on the amount of data that a subscriber can download over a cellular network during a specified time period. Breaching this cap can lead to very high excess usage charges and even throttling of the users down link capacity. Regardless of the reason for their existence data caps are commonplace and they exert a powerful influence on subscribers. The imposition of data caps has a greater impact on users with lower incomes and young adults than many other users. This cohort of users is the most likely to use fixed price, monthly pay-as-you-go data plans that have the smallest data allowances. Many users within this group are also likely to use their mobile device as the primary means of connecting to the Internet.

A report from the Pew Research Centre [5] states that almost 37% of smartphone owning American adults "mostly" use a smartphone to access the Internet, a doubling of the number from 2013. The report also states that 58% of young adults, aged 18-29 years of age, are more likely to use their smartphone to go online, up from 41% in 2013. This growing trend towards smartphone-based Internet access is evident across all age groups and income brackets. A minority of Americas rely exclusively on their smartphones for Internet access with 17% of US adults reported as "smartphone-only Internet users". This means that they report owning a smartphone but do not have a high-speed Internet connection at their place of residence. Lower income adults are most likely to be "smartphone-only Internet users" with almost 25% of this cohort having no other means of accessing the Internet. Obviously, these mobile users with appropriately equipped devices seek alternative ways to connect to the Internet to reduce the amount of data downloaded over their cellular connection and to protect their data cap. This is especially important in times of crisis and uncertainty such as the Covid-19 pandemic. In this scenario it is vital for users to conserve their data allowance so that they can access public health and safety information, use health related apps such as the Health Services Executive's Covid-19 tracker app, access government guidelines, social services and reliable information online over cellular networks when Wi-Fi may not be available.

A previous report by OpenSignal [6] revealed that in many countries around the world smartphone users spent a considerable amount of time connected to Wi-Fi Access Points (APs). Rather surprisingly the country whose citizens spent the largest proportion of their time connected to Wi-Fi APs was the Netherlands. Dutch mobile users spent approximately 63.8% of their time connected to Wi-Fi. This was despite the Netherlands having some of

the world's highest average speeds for 4G cellular networks. Clearly Wi-Fi remains an important wireless technology for mobile users regardless of the availability of high speed, high capacity cellular networks.

In addition to voluntary connections to Wi-Fi networks initiated by end-users the mobile network service providers themselves also make frequent use of Wi-Fi. A common strategy adopted by service providers for relieving the strain on overburdened cellular networks is to offload data onto Wi-Fi networks. Cisco [1] predicts that by 2022 the amount of data offloaded from smartphones to Wi-Fi will be 59% of data traffic, up from 57% at the end of 2017. Data offloading to Wi-Fi has increased despite the deployment of 4G networks having faster speeds and greater capacity than previous cellular networks. The growth in the amount of offloaded data is due to the 4G networks attracting users with high data usage devices which has also led to so-called "unlimited" 4G data plans having data caps. It is expected that the 5G networks currently being deployed will also resort to data caps, data offloading and connection throttling to control the amount of data on the new networks. This strategy has the advantage of reducing service provider's capital expenditure (CAPEX) and reducing energy costs (OPEX) as well as expanding coverage for subscribers. With the introduction of an increasing number of 5G networks over the coming years it is predicted that both data caps and connections speeds will be greater than ever. These increases in speed and capacity will be met by high usage applications such as Virtual Reality (VR) and Augmented Reality (AR) which will have increased data requirements. It is estimated that by 2022 71% of 5G data traffic will be offloaded to alternative networks [1].

Technology used in cellular networks and in all other wireless communications networks continually evolves in order to meet user requirements and demands. The demand for 'always on' Internet connectivity and the rapid rise in the consumption of video content drove the development and deployment of 3G and 4G networks. The development of 4G (LTE and LTE-Advanced) heralded the arrival of what could be truly described as mobile broadband. These networks have the speed and capacity to meet both the current requirements and expectations of users.

Early adapters of 4G technology saw tremendous gains in terms of download speeds and reductions in latency. However, as the adoption rates for LTE/LTE-Advanced have increased rapidly the benefits of the technology for new users has not been as marked.

Current users experience speeds and capacity at approximately 50% of the levels experience by the early adapters.

It has always been the case with new communication technologies that as the benefits of the technology (ease of use of the technology, increased speeds and capacity) becomes apparent to end users the adoption rates rise rapidly. All communication networks have finite capacity and sooner or later the system reaches its limits after which the original benefits reduce for adopters of the technology. Although 4G systems have yet to reach their limits it is inevitable that eventually they will do so after which more advanced and efficient systems will be developed and deployed.

Social media sites such as Facebook see video delivery as a key component of their offerings and push video and graphic content towards their huge user base. Demand for video services over Facebook during the 2020 Pandemic [7] have sky-rocketed over a very short period of time. The business model of many of the providers of so called 'free' services on the Internet is to harvest and sell user data to enable targeted advertisements and revenues have surged. It should be remembered that advertisements also consume a significant amount of bandwidth on both fixed and mobile networks. Delivery of advertising content has a negative impact on end user experience as it increases load times for web pages and consumes some portion of a user's data plan that the user must pay for.

With the ever increasing demand for network capacity and a rise in mobile data consumption comes an increase in energy consumption and capital expenditure for service providers. 5G is being touted as solution to the problem of how to meet the need for faster, higher capacity networks. In a simplistic sense the proposed 5G systems will build upon the foundation provided by current 4G systems. With regards to 5G not everything is certain, nevertheless, some already proven technologies such as cognitive radio (software defined radio), small cell deployment, co-operative systems, Self-Organising Networks (SONs) and green Multi-mode RF seem to be solid candidate technologies for 5G components.

Growth in mobile data traffic far exceeds growth in voice traffic and has done so since 2009 and currently it is estimated that Voice over IP (VoIP) accounts for less than 0.5% of all traffic on mobile networks. In addition to rapidly increasing levels of all types of traffic but particularly video traffic, the communications networks of mobile service providers face yet another challenge. The predicted growth in the numbers of interconnected devices, in the

form of the Internet of Things (IoT), will result in end users being tracked, monitored and served by tens, if not hundreds of machines. There are of course wildly conflicting numbers of IoT devices forecast but there seems to be general agreement that the number will exceed 6.4 billion excluding smartphones, tablets and computers [8]. In order to support human activities, regulate human habitats, enable smart cities, assist in transportation systems and inter-vehicle networks devices will need to communicate and cooperate with each other. Machine-to-machine (M2M) communications of this nature will require very stringent latency of less than 1 ms [8].

Any 5G system will be a heterogeneous wireless network or HetNet and users will face all the challenges currently faced by them in heterogeneous networks including efficient network selection, connection delay and UE energy conservation. The vision for 5G is that of a converged system consisting of multiple communicating technologies that support a wide and varied range of applications. These applications are expected to include multi-GB per second mobile Internet, vehicle-to-vehicle (V2V) networking, vehicle to infrastructure communications, Machine Type Communications (MTC), public safety applications, etc.

Initial deployments of millimetre-wave, very high speed, short range 5G technologies are not without their problems [9]. Practical, real world tests have shown that the effective range of the technology varies with implementation but currently ranges from approximately 122m to 244m. The short range makes it difficult and expensive to achieve widespread coverage in urban areas, reports suggest that in order to achieve comprehensive coverage of an area it is necessary to install equipment every 160m. Because millimetre-wave 5G supports very high speeds it makes sense to deploy the technology in those areas of a city that attract the greatest number of users such as shopping malls, college campuses, sports arenas, etc. This suggests islands of 5G connectivity scattered through larger areas of 4G and 3G coverage, enabling UEs to fall back to 4G when they move out of 5G coverage.

Additionally, over-heating is an issue with 1st generation 5G devices [10]. For smartphones thermal throttling is a fact of life, it is usually a result of intensive 3 D gaming or from a device being left in direct sunlight for a long period of time e.g. mounted on a vehicles dashboard or left on a windowsill. When a System on a Chip (SoC) generates a lot of heat that cannot be dissipated quickly enough the CPUs react by slowing down their operations and thereby generating less heat, for 1st generation 5G systems this is a problem for the modem. 4G LTE smartphones employ a single SoC that combines all the usual computer

components with the LTE modem in a single unit. The first generation 5G device design requires the same chip working in conjunction with a separate chip for the 5G modem and a module containing several chips for the 5G antennas. As a result, 5G components require more space than 4G and they also generate more heat. When the heat becomes excessive in a 5G device and cannot be dissipated quickly enough the 5G components are shut down and the device switches to 4G operation. It is clear that despite the growing number of 5G system deployments that both Wi-Fi and 4G technologies will have an important role to play in communications for many years to come.

1.2 Research Issues

Despite the advances achieved in cellular network technologies such as LTE/LTE-Advanced and the wide-spread deployment of 4G networks Wi-Fi remains an important component of modern mobile communication eco-systems. Even in countries that have extensive 4G infrastructures users have been reported to spend more than 60% of their time connected to Wi-Fi access points (APs) [6]. Future 5G systems will be heterogeneous network systems having Wi-Fi as an essential component of these systems.

Multi-homed, intelligent user equipment will leverage the characteristics of these HetNets to provide users with enhanced Quality of Experience (QoE) while making efficient use of network resources and conserving limited energy resources. However, HetNets are not without their problems having no centralized control, large differences in both speed/capacity and effective range of co-existing networks, as well as gaps in coverage. In urban environments the movement of large numbers of end-users from one coverage area to another can result in large, frequent fluctuations in available bandwidth.

Fluctuating bandwidth can have a serious negative impact on user QoE when consuming video content. In an attempt to overcome the effects of fluctuating bandwidth on viewer satisfaction dynamic adaptive streaming over http (DASH) technologies have been developed. The media content for delivery to DASH enabled clients is broken into segments of varying bitrates with the bitrate of a requested segment being dependent on the current network conditions being experienced by the client. If the available bandwidth falls below the level required to stream the currently selected segment the client will request the next segment having a lower bitrate than the current segment. Conversely, if the amount of available bandwidth happens to increase then the client will request a segment having a

higher bitrate than the current segment. Sudden switches between segments of differing bitrates can impact on user QoE. In addition, stalling events caused by insufficient available bandwidth have significant impact on user QoE.

The existence of multiple networks implies the need for network selection algorithms that will enable the UE to select the network most appropriate to the user's needs. As well as selecting the most appropriate network to connect to it is also important for the algorithms to be able to decide not to connect to any available network if this is the best course of action to take.

Mobile users seeking to use Wi-Fi APs face two fundamental challenges:

- The relatively short range of Wi-Fi itself
- The amount of time required to detect and connect to an AP (the connection delay)

The connection delays associated with Wi-Fi continue to be a major concern. A mobile user traveling through the coverage area of an AP has but a short period of time in which to establish a connection and transmit or receive data. Frequently, the dwell time or amount of time a mobile user remains within range of an AP is less than or equal to the connection delay. Any attempt to connect to the AP under these conditions results in reduced data transmissions and an increase in energy consumption.

1.3 Problem Statement

Owners of mobile devices such as smartphones who wish to enjoy high definition (HD) video content while on the move face many challenges. The simple fact that they are in motion is a significant contributor to many of the problems that they experience. Wireless communications have always, by their very nature, been unreliable subject as they are to frequent, unpredictable disruptions. A mobile user will experience fluctuations in the quality of their wireless connection as they pass from areas having good wireless conditions to areas having poor wireless conditions. The change from good to bad (and vice-versa) can be rapid, often only requiring a shift in position by the user of a few metres. The number of current connections to the Wi-Fi AP or mobile base-station and distance from the Point of Attachment (PoA) also have a large impact on the Quality of Experience of a mobile user.

Mobile devices and smartphones are often multi-homed having at least one cellular interface and one Wi-Fi interface. This basic configuration gives rise to the possibility, in a heterogeneous multi-network wireless environment, of using the strengths of one wireless technology to compensate for the weaknesses in another wireless technology. For example, the cellular interface can establish a connection having an effective range measured in kilometres which can be used to maintain a communications session when the Wi-Fi interface, having an effective range of only tens of metres, moves out of range of an access point (AP).

In order to be able to use the strength of one wireless technology to compensate for the weakness in another wireless technology requires the mobile device to be able to first detect available networks and to then make a decision as to which one it wants to connect to. A great many network selection strategies have been proposed that employ various selection criteria such as signal strength but few if any take into consideration the rate of speed at which the user is travelling and none, to the best of our knowledge, employ the user's data-cap as an input to their decision making process.

Portable devices rely entirely on their users for mobility, the user controls when movement occurs, where the device is moved to, the direction of travel and the speed at which the device is travelling. This makes user behaviour and in particular user speed, something to be taken into consideration when deciding at which point to begin the network detection process. The network detection process is costly for users of mobile devices in terms of energy consumption. Many devices such as smartphones automatically scan for available networks regardless of current conditions so that even when networks are detected it is not always possible to establish a useful connection. The usefulness of any mobile device is determined by the amount of energy it has stored in its battery and all current mobile devices are constrained in this way. Many users of mobile devices struggle to get a full day of usage from a single charge and obviously balancing energy conservation against providing useful services to the user is difficult but must always be considered.

Energy consumption can be reduced by only initialising the network detection process when a realistic chance of establishing a useful connection exists. A challenge for mobile users in HetNets is deciding when, if ever, to begin the network detection and selection process to conserve energy whenever possible.

Fluctuations in available bandwidth are a common characteristic of wireless communications. These fluctuations can have a serious, detrimental impact on a user's QoE when viewing streamed video content over a wireless communications link. Dynamic adaptive streaming techniques have been developed to address the issue of fluctuations in available bandwidth. However, the end users QoE can be negatively impacted on by events such as breaks in connectivity and stalling events in the video playback which can occur even when using adaptive streaming techniques.

As previously stated, many users are constrained in the amount of data that they can download by the data caps imposed on them by their service provider. In order to protect their data allowance users will change their Point of Attachment (PoA) from a cellular network to Wi-Fi whenever possible. This can lead to a decrease in the user's QoE due to increases in the number of stalling events and increased interference due to devices operation on the same frequency bands. The challenge for mobile users is how to maintain or increase their QoE while protecting their data cap. Methods used to maintain or increase QoE should seek to minimise the amount of data downloaded over a cellular network and limit the impact this may have on energy consumption where possible.

1.4 Solution Overview

This thesis introduces a novel de-centralised, user-centric system that enables mobile users to maintain or increase their QoE in heterogeneous, multi-network wireless environments, minimise the amount of data downloaded over cellular connections and conserve energy. The proposed solution models the process of making a decision as to when to begin the network selection process as a utility function.

The system makes use of a scan or no scan solution that removes automatic Wi-Fi scanning for mobile devices and replaces it with a dynamic, adaptive Wi-Fi scanning model. Scans are only undertaken when there is a realistic expectation of a successful connection to Wi-Fi being made. QoE is maintained and frequently improved by load balancing HD video content over multiple, heterogeneous wireless links using real-time playout buffer levels as an input into the decision making process. End user data-caps are protected by offloading as much of the video content as possible to the Wi-Fi links. The solutions novel mechanisms are as follows:

- Scan-Or-Not-to-Scan (SONS)
- MPEG-DASH based Framework (MDF)
- Adaptive Interface Selection (AIS)

1.5 Thesis Contributions

1.5.1 Scan-Or-Not-to-Scan (SONS) - SONS is a novel framework that employs a utility function that assists a mobile system in deciding whether to run a network selection algorithm when in the presence of multiple wireless networks. The decision to run or not to run a network selection algorithm provides an opportunity to save energy and maintain data transfer rates. Activating wireless interfaces and scanning for available networks in a deliberate manner as opposed to automatic scanning and interface activation can save energy. The number of unnecessary and potentially unsuccessful scans and connection attempts is reduced, the total amount of time that a user is disconnected is reduced by avoiding needless disruption of established connections and QoE for end-users of streamed video content is improved by reducing the number of handovers and changes in video bitrate. The SONS framework also contains the **SONS utility function,** a novel utility function that takes the users speed over the ground, remaining data-cap and AP coverage area as inputs and generates a utility score. This utility score is used by both the MPEG-DASH-based Framework (MDF) and Adaptive Interface Selection (AIS) in their operations.

1.5.2 MPEG-DASH based Framework (MDF) a generic technology agnostic framework that seeks to optimise the performance of MPEG-DASH enabled clients streaming SD video content in multi-network wireless environments. MDF employs the SONS mechanism to determine when conditions are suitable for a user to attempt to switch their point of attachment from a cellular network to an alternative Wi-Fi network. Attempting to connect to alternative networks only when a reasonable possibility of establishing a useful connection exists reduces disruption to connectivity. MDF also matches the requested video segments with device capabilities (e.g. screen size) to avoid downloading segments of a higher definition than can be utilised and to protect the user's data-cap.

1.5.3 Adaptive Interface Selection (AIS) is a mechanism for streamed HD video delivery over bandwidth constrained links in wireless HetNets which improves end-user QoE by employing both Wi-Fi and cellular bandwidth resources simultaneously when necessary. In-

itial video segments and associated files are downloaded via the mobile devices cellular interface, then if Wi-Fi conditions permit downloading of segments is carried out over the Wi-Fi interface and the cellular interface is disconnected. AIS monitors the level of the playout buffer of the DASH enabled client. When the Wi-Fi network conditions are unable to support HD video streaming, indicated by the playout buffer being depleted faster that it can be replenished, the cellular interface is temporarily reactivated, and additional segments are downloaded over it. Both interfaces are used to download video segments simultaneously until the playout buffer has been replenished. If conditions are not conducive to Wi-Fi operations the Wi-Fi interface is shutdown to prevent unnecessary scanning operations being carried out and all video segment downloads occur via the cellular interface.

This strategy greatly reduces the number of stalling events by boosting the overall bandwidth being used to stream the video content. Multiple stalling events in streamed video have been shown to negatively impact an end-user's QoE, significantly reducing the number of stalling events improves QoE of the viewer. AIS also reduces energy consumption by disabling interfaces when not required. It also reduces the amount of data downloaded over the cellular networks by offloading the streaming function to Wi-Fi where possible.

In order to deal with the fluctuations in available bandwidth that mobile users experience, modern video streaming employs adaptive bitrate (ABR) algorithms. MPEG-DASH enabled clients adapt the bitrate of requested video segments to match as closely as possible the available bandwidth of the wireless link in use. If link conditions are poor or the link is congested, the ABR algorithm will select the lowest available bitrate supported by the video stream. This approach works well for standard video content since there is no restriction on using low bitrates.
However, for users wishing to view High Definition (HD) content, there is a lower bound to the bitrate below which they cannot go if the content is to be considered HD. ABR algorithms, forced to maintain HD bitrates over wireless links with constrained bandwidth, experience significant numbers of stalling events. High numbers of stalling events, particularly towards the end of a video, have a serious negative impact on end-user QoE.

This work proposes AIS to address this issue. AIS leverages the multi-homed nature of modern mobile devices and the segmented nature of video content prepared for DASH-enabled clients to overcome the bandwidth constraints that hobble ABR algorithms in the context of

delivering HD video. AIS uses multiple wireless interfaces in parallel to download content to maintain the playout buffer level to prevent stalling events. By minimizing the number of stalling events and streaming at HD compatible bitrates end-user QoE is maintained. Downloading as little content as possible over mobile networks helps protect the end-user's mobile data-cap.

1.6 Thesis Objectives

This thesis seeks to provide a system of integrated components designed to maximise a mobile users' QoE while watching DASH enabled video streams, minimise the amount of data downloaded over cellular network connections while streaming video and reduce the number of unnecessary Wi-Fi scanning and connection operations to minimise streaming disruption and conserve energy. These elements can operate as standalone components or be tightly integrated into a single, unified system. Such a system abstracts from the mobile user the decision making process around deciding when to attempt to connect to Wi-Fi APs and when not to, it also automatically determines if a video stream is SD or HD and invokes the appropriate resolution specific sub-module to deal with it.

This thesis sets out a number of specific objectives:

1. To develop a process to decide when to initiate Wi-Fi scanning operations that takes into consideration a user's remaining data-cap, AP coverage areas and the user's speed over the ground
2. Enable a mobile user to maximise Quality of Experience while streaming SD video
3. Enable a mobile user to maximise Quality of Experience while streaming HD video
4. Reduce the overall amount of data transferred over cellular connections
5. Reduce energy consumption by deactivating wireless interfaces when possible

1.7 Structure of the Thesis

The thesis has been structured in 7 chapters as follows:

Chapter 1 Introduction – establishes the motivation for the research activity carried out, the problem statement is discussed and an overview of the proposed solution is presented. This chapter also details the contributions to the advancement of the state of the art.

Chapter 2 Technical Background – discusses the technical background to the work related to wireless data communications including Wi-Fi and cellular communications with reference to 5G systems

Chapter 3 presents the related works

Chapter 4 introduces the novel Scan-Or-Not-to-Scan (SONS) framework. The SONS utility function is introduced and examples of use included. Simulation models and scenarios are presented, methodology is described and results are presented.

Chapter 5 describes MDF, a generic technology agnostic framework that seeks to optimise the performance of MPEG-DASH enabled clients streaming Standard Definition (SD) video in urban HetNet environments. Simulation models and scenarios are presented, methodology is described and results are presented.

Chapter 6 introduces Adaptive Interface Selection (AIS). AIS leverages the multi-homed nature of modern mobile devices and the segmented nature of video content prepared for DASH-enabled clients to overcome the bandwidth constraints that hobble ABR algorithms when streaming HD content. Simulation models and scenarios are presented, methodology is described and results are presented.

Chapter 7 Concludes the thesis and presents possible directions for future work

CHAPTER 2 TECHNICAL BACKGROUND

This chapter presents the technical background to the work of this thesis. It describes the major technological fields involved including IEEE 802.11 wireless communications, mobile phone networks, 4G and 5G communications systems, Quality of Service, Quality of Experience, Multimedia Streaming and adaptive streaming strategies.

2.1 Wireless Communications Overview

While there are a great many wireless technologies are in use today they have basic underlying system components in common that are always the same, a device capable of transmitting an encoded signal, the transmission medium itself and a receiving station capable of capturing the wireless transmission and decoding it. Broadly speaking, every wireless technology offers the following benefits

- Mobility - users of the technology are freed from the restrictions of being tethered to a single location. Users may move but the data that they depend on is typically stored at a fixed location. Enabling access to stored data from anywhere, at any time can lead to productivity gains.
- Network deployment is simplified and takes less time - in many situations, in city centres or over bodies of water, building of physical infrastructure may be prohibitively expensive, bureaucratically difficult or both. Wireless technologies reduce the need for wired infrastructure making them faster to deploy.
- Flexibility - Having no physical points of connectivity means that users can come and go as they please with no need of technical assistance, user administration is automated and reduces the load on the network administrators. More importantly, the absence of physical connectivity means that users can connect to the network and maintain their connectivity even while in motion.
- Cost - The cost of deploying wireless networks is quite often, but not always, considerably less that their wired equivalent. In any event, if the service provider wants to target mobile users there is simply no alternative.

In addition to its many benefits wireless communications also has various drawbacks including:

- Planning issues – it is very difficult to ensure comprehensive coverage in all areas
- Scarcity of suitable sites for infrastructure – many of the most suitable sites for wireless infrastructure in urban areas has already been occupied making expansion difficult
- Reliability issues – blocking of signals, signal fading, multi-path propagation, and interference make wireless communications unstable. This is often as a result of a mobile user changing position
- Security – Denial of Service (DOS) attacks, Man in the Middle attacks and many other types of attacks are difficult to prevent in wireless environments where the signals propagate in free space
- Shared spectrum – the available bandwidth must be shared amongst all users and this can result in poor performance, as the number of users increased the share of bandwidth per user decreases
- Physical environment – the physical environment in which the wireless network is deployed may change at short notice impacting on network performance. For example, the growth of vegetation such as leaves on trees can absorb wireless signals. The network operator may have little or no control over the physical environment itself

2.2 Evolution of Mobile Telephony

2.2.1 GSM

In 1946 the first commercial, car-based telephony service, operated by AT&T, was approved for use by the US Federal Communications Commission (FCC). The following year, 1947, saw AT&T introduce the concept of radio frequency reuse or the cellular network system [11]. This is a concept that was to underpin all mobile communication systems that followed.

For many years following after their 1946 introduction mobile telephony systems remained reliant on motor vehicles for their mobility due to the size, weight and energy requirements of the equipment. Despite the obvious limitations, similar systems were built in various

countries during the 1950's and 1960's but the number of users was low reaching a few thousand at best.

Initially, the early mobile communication systems were deployed by national monopolies and fixed-line service providers. It was only after mobile communications became internationalised that a surge in subscriber numbers and usage occurred. The analogue Nordic Mobile Telephony (NMT) system, introduced in the Nordic countries in 1981, was the first international mobile communications system. Because NMT was an international, cross-border system it was necessary to develop the ability for customers to 'roam' from a home network to another network. Enabling subscribers to obtain telephony services outside their home networks also provided an impetus for growth in the mobile phone market as well as attracting new entrants into the mobile communications business.

Also in 1981 the analogue Advanced Mobile Phone Service (AMPS) was introduced in North America and various other analogue technologies were deployed in other countries. All of these early analogue systems had a number of short-comings in common, the equipment itself remained unwieldy and power hungry and was typically carried in motor vehicles, 'cross-talk' between customers frequently occurred and the voice quality of calls was inconsistent.

2.2.2 GSM System Architecture

The Global System for Mobile Communication (GSM) is a digital system developed to replace the original, first generation analogue communication systems. [3] Digital systems have certain advantages over analogue systems, voice traffic can be encrypted, digital systems make more efficient use of scarce radio spectrum resources and digital systems can provide data services such as SMS text messaging and video calls. The use of digital technologies also enables the use of efficient compression techniques and multiplexing of traffic.

The first commercial GSM phone call was made by Radiolinja in Finland in 1991, initially GSM was used almost exclusively for voice communications but this was to quickly change following the sending of the first SMS text message in 1992. SMS messaging quickly gained popularity and continues to be a staple of inter-personal communications today. The basic components of a GSM system are presented in Figure 1.

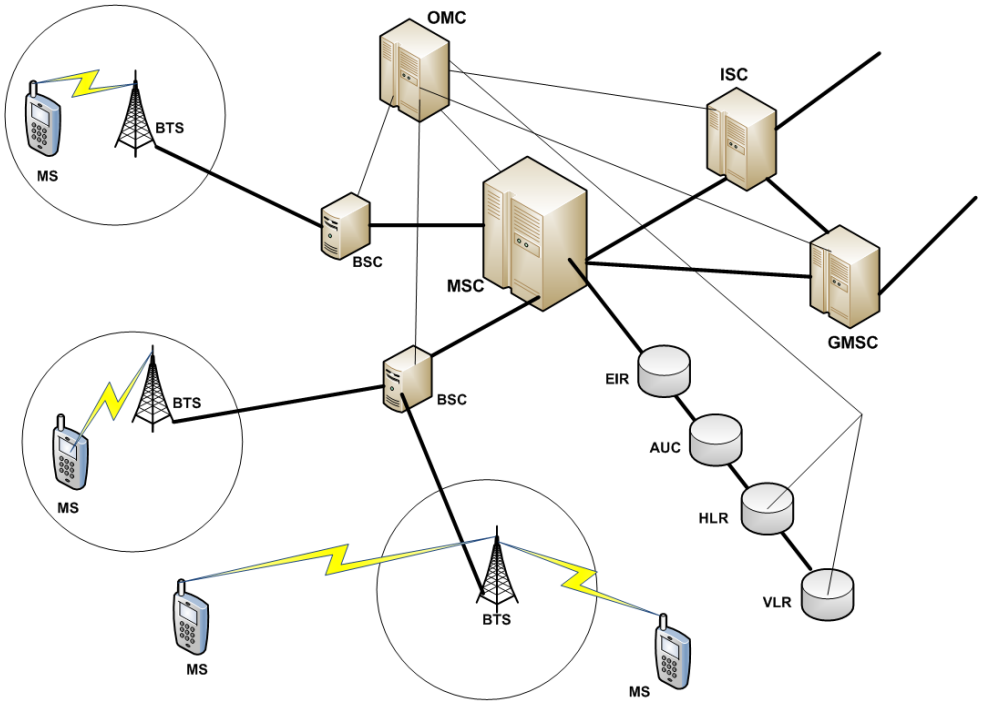

Figure 1 Basic Components of a GSM System

A subscriber has a Mobile Station (MS) that can communicate with a base station over the air. In a GSM system the base station is known as a Base Transceiver Station (BTS), the BTS has equipment for transmitting and receiving radio signals in addition to some equipment for signal and protocol processing.

In order to minimise BTS size, necessary control and protocol intelligence is maintained in the Base Station Controller (BSC). Each BSC will contain protocol functions to, among other things, allocate channels, manage channel setup and tear down and manage handovers. It is usual for a single BSC to control several BTS and the BSC may be connected to a BTS via a fixed, physical link or over a dedicated, point-to-point radio link. The radio access network consists of both the BTS and the BSC.

2.2.3 Switching in a GSM network

In GSM networks user traffic is routed through a switch, the Mobile Switching Centre (MSC) that carries out all of the functions that an equivalent switch in a fixed telephone network would. These functions include data forwarding, path search and service feature processing; in addition, the MSC must also take into consideration user mobility and the administration of radio resources.

Additional functionality is also required from the MSC to handle the registration of mobile user locations and to take care of the handover of a connection in the event that a mobile user changes from one cell to another. A GSM network may contain multiple MSCs with each MSC having responsibility for a particular section of the network, for example that portion of the network that covers a city. A dedicated Gateway MSC (GMSC) is used to handle calls terminating in or originating from a fixed telephone network.

The Internetworking Function (IWF) takes care of inter-networking between a cellular network and a fixed network. It maps the protocols used in the cellular network with the equivalent fixed network protocols. Traffic to international telephone networks and to other mobile networks is routed via the International Switching Centre (ISC) of the respective country.

2.2.4 GSM Network Databases

GSM networks contain several different databases, for example, the current location of a mobile user is held in the Home Location Register (HLR) and the Visited Location Register (VLR). The location data is stored to enable the network management and switching functions to establish a call to the correct BTS. User profiles are also stored in these databases; the user profile is required for various administration tasks as well as for billing and charging activities. Network security functions make use of two other databases, security related data such as encryption keys and authorisation keys are stored in the Authentication Centre (AUC) while equipment related data is held in the Equipment Identity Register (EIR).

The Operation and Maintenance Centre (OMC) centralises the organisation of network management. OMC functions include subscriber administration, terminal administration, nework configuration, network operations, charging data, performance monitoring and network maintenance.

A GSM system may be divided into three subsystems, the Base Station Subsystem (BSS) which is concerned with the radio access network, the Network Switching Subsystem (NSS) which is concerned with the core network and the Operation and Maintenance Subsystem (OMSS) which is concerned with the management network.

A hierarchical relationship exists between the Base Transceiver Station (BTS), the Base Station Controller (BSC) and the Mobile Switching Centre (MSC). A GSM network is divided into one or more MSC groups; each MSC group is composed of at least one Location Area

(LA). A Location Area is made up of several cell groups and each cell group is assigned to a BSC. There is at least one BSC for each Location Area.

2.2.5 General Packet Radio Services (GPRS)

In the development of cellular networks towards the provision of mobile broadband GPRS was an important step. Efficient access to IP networks was enabled through the use of packet-orientated transmission technology.

GPRS was built upon the pre-existing GSM architecture which was extended through the use of two new GPRS nodes, the SGSN and the Gateway GPRS Support Node (GGSN) (Figure 2). Integration of GPRS into the GSM infrastructure required the development of a new category of nodes known as GSM Support Nodes (GSNs). The routing and delivery of data packets between mobile stations (MS) within the cellular network and external packet data networks (PDNs) was the responsibility of the new GSNs.

Delivery of data packets within a particular service is carried out by the service area's Serving GPRS Support Node (SGSN). The work of an SGSN also includes MS authentication and the attach/detach function, as well as logical link management and the routing and transfer of data packets. A SGSN's location register holds the location information and user profiles for all GPRS users registered with the SGSN.

The interface between the internal cellular network and external packet data networks (PDNs) such as the Internet is provided by a Gateway GPRS Support Node (GGSN). A GGSN converts GPRS packets received from a SGSN into the appropriate Packet Data Protocol (PDP) format, for example IP, and forwards the packets out to the destination external network. Data packets travelling into the cellular network from an external PDN pass through the GGSN. The GGSN converts the PDN address i.e. the destination IP address into the GSM address of the target device. Re-addressed data packets are then forwarded to the appropriate SGSN.

In order to support this activity, the GGSN stores the current SGSN addresses and profiles of registered users in a location register. A GGSN can act as an interface to an external network for multiple SGSNs and conversely a SGSN can route its data packets to different GGSNs if available.

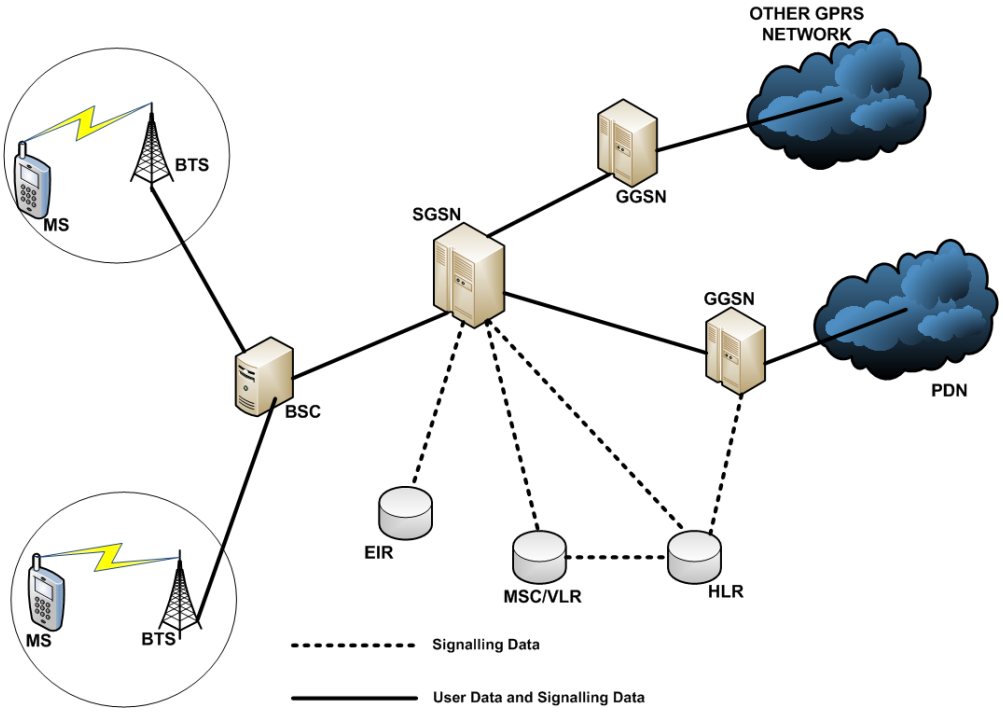

Figure 2 Basic GPRS Architecture

GPRS has the ability to support Quality of Service (QoS) and for each PDP context in use an individual QoS profile can be generated that details service requirements such as latency, throughput and reliability. In order to use GPRS services a MS must first obtain an address appropriate to the external PDN to be contacted e.g. an IP address. A PDP context is then created, the context describes the session parameters such as QoS, GGSN to be used, PDP address, PDP type, etc. Dynamic address allocation systems such as DHCP are also used to support large numbers of mobile users.

Once an active PDP context has been created for a mobile station data packets from the external network addresses to the MS can be routed to the relevant GGSN. Data packets are sent from the GGSN to the MS's current SGSN which in turn forwards the data packets to the MS.

The GPRS air interface is packet orientated and any MS with multi-slot capability can transmit on multiple timeslots within a TDMA frame. There is also separate allocation of uplink and downlink with physical channels only being assigned for the duration of a transmission session. Radio resources are more efficiently utilised through the use of this flexible channel allocation strategy.

2.2.6 Predecessors to 3GPP Long-Term Evolution (LTE)

LTE's design was the result of collaboration between various regional and national standardisation bodies. This group of bodies is known as Third Generation Partnership Project (3GPP) and they were also responsible for the predecessor to LTE which was called the Universal Mobile Communications System (UMTS). The Universal Mobile Telecommunications System (UMTS) was an evolution of Global System for Mobile Communications (GSM) and the same network architecture is used by both UMTS and GSM.

This architecture consists of three basic components, the core network, radio access network (RAN) and the mobile phone or user equipment (UE). Figure 3 depicts this common architecture and as can be seen the core network itself is divided into two domains, the circuit switched (CS) domain and the packet switched (PS) domain. The circuit switched domain behaves in a similar fashion to a traditional fixed-line telephone system. It is responsible for routing voice traffic through the geographical area in which the service provider responsible for the call operates. Telephone calls can be made to landlines because the circuit switched domain interfaces or communicates with the public switched telephone networks (PSTN) of other operators.

The circuit switched domain also communicates with the circuit switched domains of rival network operators. This enables a customer of one mobile network provider to call a customer of another mobile operator. Data streams (email, web pages, etc.) flow between the core networks packet switched domain to which a customer is connected and external packet data networks (PDNs). These packet data networks include networks such as the Internet.

Traffic flows across the core network's two domains in very different ways. In the circuit switched domain a procedure known as circuit switching is used. This procedure establishes a dedicated bi-directional connection for each telephone call to ensure a constant data rate and minimal delay. It is the same approach used by traditional telephony systems and while effective it is very inefficient. The established connection is allocated sufficient capacity to be able to deal with worst case scenario in which both parties speak simultaneously. Due to the fact that this situation rarely arises the connection is typically over-provisioned. In addition, this approach is not suited to data transfers which tend to be 'bursty' in nature i.e. rapidly varying in data rates.

The packet switched domain was created to address this issue and as a result it employs a different approach to transporting data. In the packet switched domain, data to be transported is first segmented into units known as packets. Each of these packets carries additional addressing information that identifies both the source of the data and the destination device to which it is being sent. Routers within the network read destination address of every packet that it receives and forwards each packet through the appropriate interface towards its final destination. Instead of reserving network resources for a single connection for some period of time the packet switching approach shares network resources amongst all users as necessary. This makes more efficient use of network resources since there is no over-provisioning of a connection but in the event that too many devices try to send data simultaneously congestion can occur on the network resulting in delays. Communications between a user's mobile phone and the core network are dealt with by the radio access network (RAN).

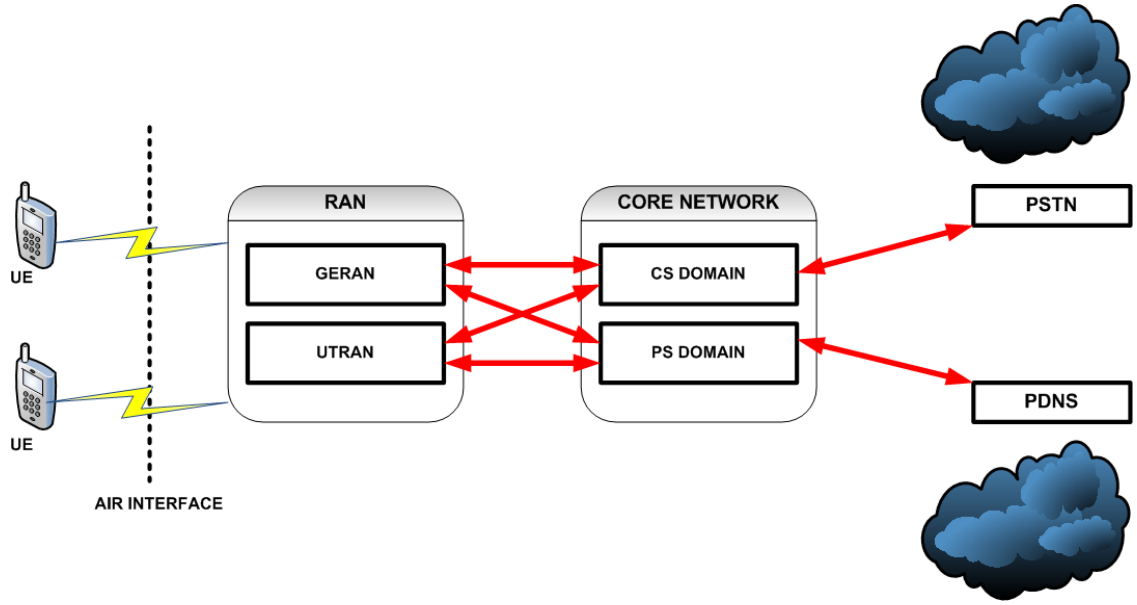

Figure 3 Common GSM/UMTS Architecture

In Figure 3 we can see two separate RANs, the GSM EDGE radio access network (GERAN) and the UMTS terrestrial radio access network (UTRAN). Although each of these RANs employs the different radio techniques of GSM and UMTS they can share a common core network between them. Officially a user's mobile device is known as the user equipment (UE) but it is commonly referred to as a mobile. The user equipment connects to the RAN over the air interface (aka the radio interface). Downlink (DL) refers to the direction from

the network to the user equipment while the term uplink (UL) refers to the direction from the user equipment to the network.

One of the most powerful features of mobile phones is that they can be used outside the coverage area of their service provider a situation referred to as roaming. The process of roaming is enabled through roaming agreements between network operators. A rival operator's network is referred to as the visited network and the subscribers own network is referred to as their home network

2.2.7 Radio Access Network (RAN) Architecture

In the UMTS radio access network shown in Figure 4 below, the base station or Node B is the most significant component. A typical Node B will be equipped with one or more sets of antennas which are used to communicate with UEs in one or more sectors. In many cases, a Node B will employ a set of 3 antennas, with each antenna used to control a sector. Each sector spans an arc of 120 degrees and a mobile phone network can contain several thousand such base stations.

The name 'cellular network' is derived from the units called 'cells' into which a mobile phone network might be divided. The term cell can be used in two different ways, in the USA the term refers to a group of sectors controlled by a single Node B and in a European context the term refers to the same thing as a sector. Cells are limited in both size and capacity; they enable more efficient use to be made of available spectrum through frequency reuse. The size of a cell is the maximum range at which the UE can successfully transmit to and receive from the Node B. The capacity of a cell is the maximum combined data rate of all the UEs in the cell.

Cell sizes themselves can vary greatly depending on their type. Macro cells have the greatest size with a cell size measured in kilometres and they provide wide-area coverage in both rural and urban areas. In the more densely populated urban areas micro cells with a size of several hundred meters might be deployed. Various techniques enable micro cells to have greater collective capacity than a single cell of an equivalent size to the collective. Pico cells having coverage areas measured in tens of meters are found in large indoor environments such as shopping malls and offices. Finally, femto cells are designed for use in domestic environments and have an effective range measured in meters.

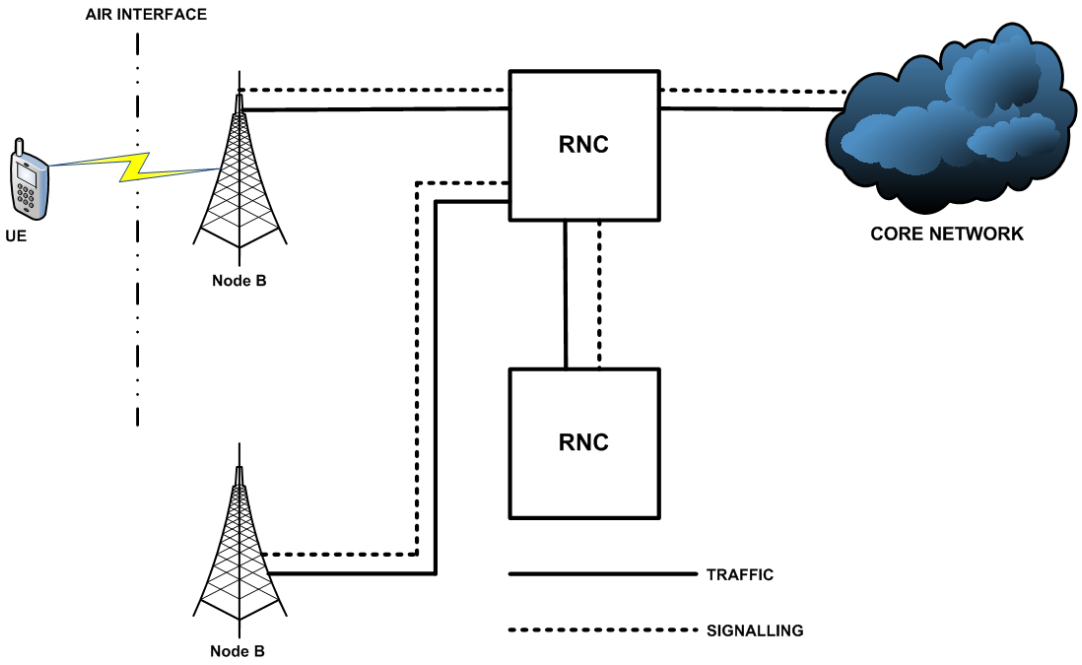

Figure 4 UMTS Radio Access Network

A vital function of the air interface is the separation of Node B transmissions from those of UEs in order to prevent them interfering with each other. In UMTS this can be achieved in one of two ways using either Frequency Division Multiplexing (FDM) or Time Division Multiplexing (TDM). In the case of Frequency Division Multiplexing the air interface achieves segregation between the UE and Node B transmissions by having the UE use one carrier frequency and the Node B use a different carrier frequency. With Time Division Multiplexing both the UE and the Node B transmit on the same carrier frequency but at different times. Another important function of the air interface is the separation of different Node Bs and UEs from each other.

When a mobile user moves through the coverage area of a Node B it will eventually reach the edge of the cell and begin moving into the adjacent cell. At this point the UE must switch its point of connection to the base station of the cell it is entering. A UE that is engaged in an active communication session can switch its point of connection using a process known as a handover. A UE that is in standby mode can use a process known as cell reselection to ensure that it is ready to communicate should the need arise. UEs in UMTS networks can establish communications with the Node B in the new cell while maintaining its connection to the Node B in the previous cell. This is known as a soft handover or make before break.

Devices known as Radio Network Controllers (RNC) control groups of Node Bs and each RNC has two main tasks. They are responsible for passing user voice information and data packets between the core network and the Node Bs. The RNC is also responsible for controlling the UEs radio communications by means of signalling messages that are transparent to the end user. For example, the RNC tells the UE when to hand over from one base station to another base station. Typical networks might contain some tens of RNCs with each RNC controlling hundreds of base stations.

Overall a GSM RAN is of similar design to a UMTS RAN. In the GSM RAN a base station is referred to as a base transceiver station (BTS) and the controller is referred to as a base station controller (BSC). In the case where a UE supports both GSM and UMTS the network can hand it over between the two different RANs. The process of handing over between GSM and UMTS is known as an inter-system handover.

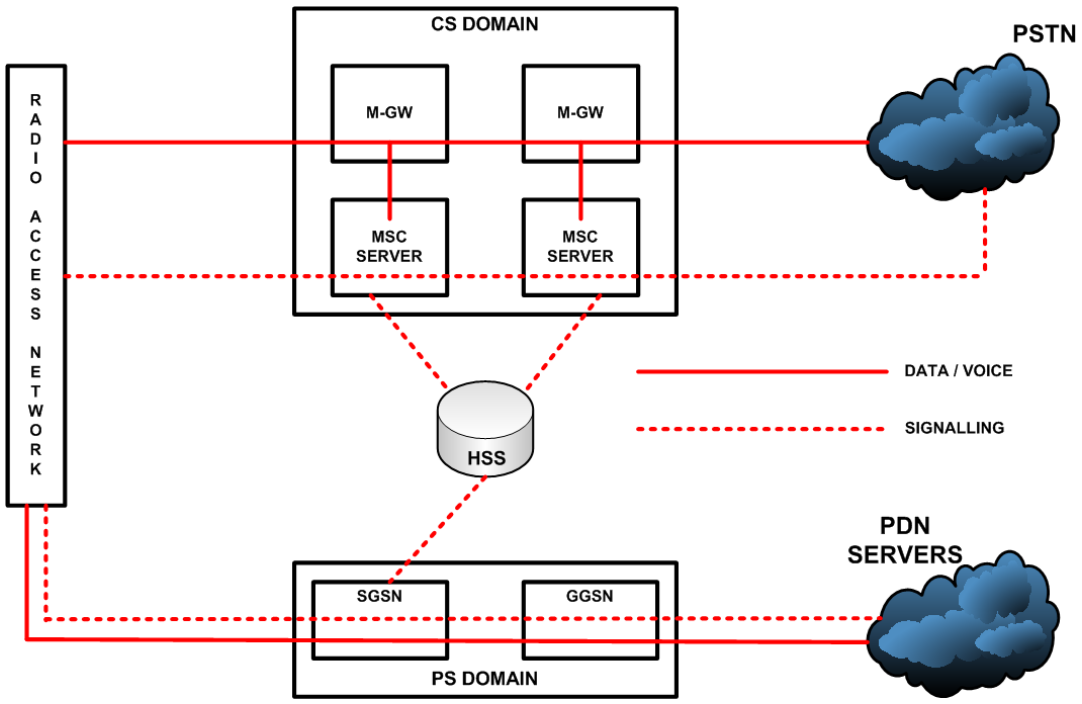

Figure 5 Internal architecture of the GSM/UMTS core network

Media Gateways (MGWs) within the circuit switched domain (CS) (Figure 5) route voice calls within the network. The Mobile Switching Centre (MSC) servers are responsible for setting up, managing and tearing down calls through the use of signalling messages. A typical cellular network might contain a few such devices.

Gateway GPRS support nodes (GGSNs) in the packet switched domain (PS) act as interfaces to external data networks and servers. Data packets are routed between base stations and the GGSNs by the Serving GPRS support nodes (SGSNs). SGSNs handle signalling messages that establish, manage and terminate data streams. Only a few such devices are deployed in a typical network.

A central shared database named the home subscriber server hold information on all the networks subscribers. It combines the function of and replaces two earlier system components, the home location register (HLR) and the authentication centre (AUuC).

2.2.8 Long-Term Evolution (LTE)

In the early years of mobile telephony voice was the dominant type of traffic on cellular networks. Data traffic on mobile networks remained low for a variety of reasons such as latency issues, network capacity, UE capabilities and the cost to end users. From approximately 2010 onwards the amount of data traffic carried on cellular networks began to increase dramatically with part of the reason for this increase in data traffic being improvements in network performance itself. However, many would argue that the most significant driver in data traffic increases was the introduction of smartphones such as the Apple iPhone (in 2007) which was quickly followed by Android based devices from 2008.

There were several ways in which smartphone adoption drove data consumption. These smartphones were more capable, had bigger and better screens and were designed to support third part application development. This resulted in the proliferation and use of mobile applications. Improvements in screen size and resolution coupled with more powerful CPUs enable the enjoyment of games and video. In addition, in a bid to drive data traffic volumes on their networks mobile operators had previously offered flat rate, unlimited data plans to subscribers. The outcome was a situation in which users were not initially motivated to control their data consumption. The early unlimited data plans for subscribers have been phased out in the intervening years. Modern flat rate data plans now come with a usage limit or data

cap. Exceeding this data cap can result in very high excess usage charges and as a consequence many users now actively monitor their data usage.

In approximately 2010 2G and 3G networks began to become congested and both latency and user dis-satisfaction increased. This situation highlighted the need for increased network capacity and performance.

2.2.8.1 Mobile Telecommunication System Capacity

The theoretical limit on the data rate achievable by any communication system can be calculated using the following formula:

$$\mathbf{C = B\log_2(1 + SINR)}$$

Where SINR is the signal-to-interference plus noise ratio, B is the bandwidth of the communication system in Hz and C is the channel capacity in bits s^{-1}. In theory it is possible for a communications system to transmit data from a sender to a receiver without any errors provided that the data rate is less than the channel capacity. In the context of cellular networks C is the maximum data handling capacity; it equals the combined data rate of all UEs in the cell.

2.2.8.2 Strategies for Increasing System Capacity

Essentially there are three main methods by which the capacity of a mobile communication system might be increased. Possibly the most important means by which overall system capacity can be increased is through the use of smaller cell sizes. For cellular networks channel capacity is the maximum data rate that a cell is capable of handling. Through a process of reducing cell sizes and deploying additional base stations the overall capacity of the system can be increased.

Another method is to increase the amount of available bandwidth. The International Telecommunications Union (ITU) and both regional and national regulators manage radio spectrum. The growth in the adoption and use of mobile telecommunications resulted in increased allocation of spectrum for 2G and 3G systems. However, suitable radio spectrum is a finite resource which must be shared with diverse legacy systems such as military communications and radio astronomy. As with any finite resource there are limits to how much spectrum could be allocated to mobile communication systems.

The third and final method to increase a communication system's capacity is to improve the technology used in the system. This strategy has several benefits, it enables more efficient use to be made of available spectrum, and it can make communications more robust. Long-Term Evolution (LTE) was the next stage in the drive for improved cellular communication systems.

2.2.8.3 Additional Motivations for LTE

An additional motivation for the development of LTE was the need to reduce both Operational Expenditure (OPEX) and Capital Expenditure (CAPEX). In order to meet increasing demands for capacity operators of 2G and 3G networks need to deploy additional resources. They need to deploy and maintain two core networks, the circuit switched (CS) domain and the packet switched (PS) domain. Depending on congestion levels within the packet switched domain it is possible to transport voice traffic over the PS domain using Voice over IP (VoIP). By converging both voice and data traffic onto a single network, operators can reduce both CAPEX and OPEX.

2.2.8.4 Reduce Latency for Data Applications

Delays in the order of 100ms are possible for data applications in 3G networks during data transfers between network components and over the air interface. While voice applications can tolerate delays of 100ms other real time applications such as interactive gaming cannot. The desire to reduce latency (end-to-end delay) on the network is an important driver for LTE deployments.

2.2.8.5 Reducing Network Complexity

Over the years the complexity of GSM/UMTS communication system grew as a result of having to maintain backwards compatibility with legacy devices and the need to add new features. Building a new system from scratch enables the system designers to improve performance and reduce complexity by not having to support legacy devices.

2.2.9 LTE High Level Architecture

The aim in developing LTE was to create a communications system capable of remaining competitive over a timescale of ten years or more. This was to be achieved by providing the low latency and high data rates that were anticipated to be required by existing and future applications and users.

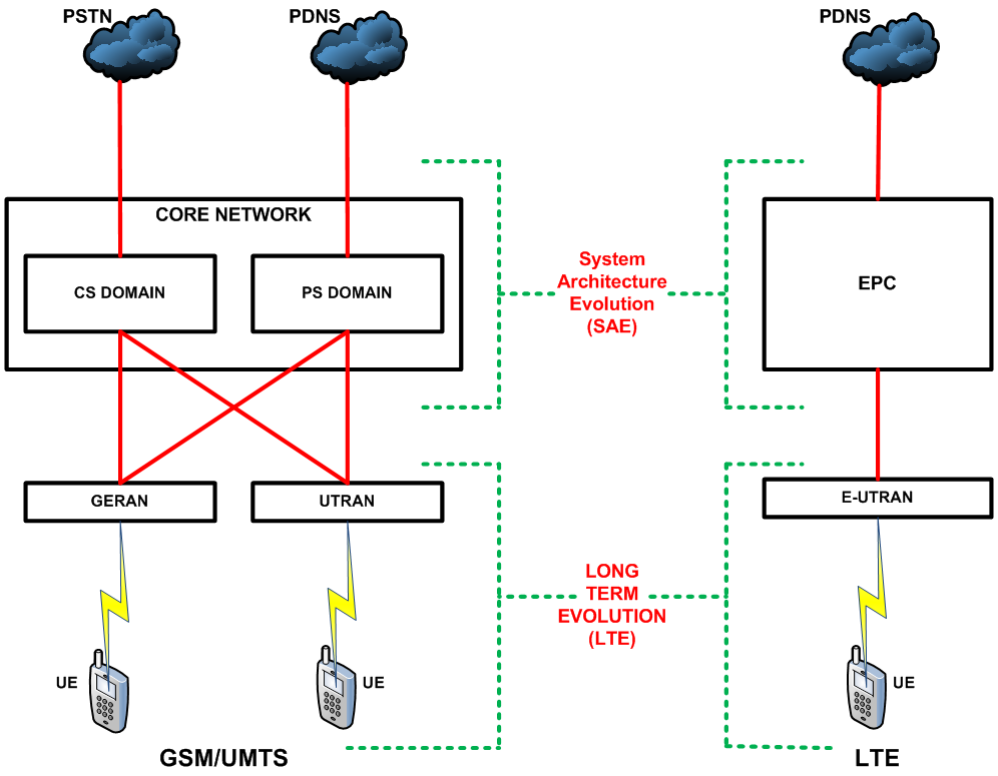

Figure 6 GMS/UMTS and LTE architectures

As can be seen from Figure 6 the evolved packet core (EPC) is a direct replacement for the GSM/UMTS packet switched domain. No equivalent replacement was added for the GSM/UMTS circuit switched domain (CS). From the beginning LTE was optimised for data traffic with voice traffic being handled by other techniques. Radio communications between the EPC and the UE are handled by the evolved UMTS terrestrial radio access network (E-UTRAN) which was a direct replacement for the UTRAN. The internal operation of an LTE mobile device is very different from that of a UMTS mobile but it is still referred to as user equipment or UE.

Two 3GPP work items gave rise to the LTE architecture. These were system architecture evolution (SAE) that was concerned with the core network and Long Term Evolution (LTE) that was concerned with the RAN air interface and UE. Evolved Packet System (EPS) is the official name by which the whole system is known while LTE only refers to the evolution of the air interface. However, regardless of official usage LTE has become the name by which the system is now known.

2.2.10 LTE Summary

LTE, as originally used, is concerned with the UE, the air interface between the UE and the RAN itself. The specifications developed required LTE to be capable of delivering a peak data rate of 100Mbps on the downlink and 50Mbps on the uplink. The system that was eventually deployed exceeded these requirements having a peak data rate of 300Mbps on the downlink and 75Mbps on the uplink. It must be remembered that these data rates can only be achieved under ideal conditions and can never be achieved in real world environments. In comparison to LTE, UMTS W-CDMA has a peak data rate of 14.4Mbps on the downlink and 5.76Mbps on the uplink.

For time-sensitive applications such as voice and real-time gaming latency is a very important issue and in the context of LTE there are two aspects of the issue to be considered. One is the time taken for data to travel between the UE and the fixed network, the other is the time required by the UE to transition from a low-power standby state to an active state. In the case of the time taken for data to travel from the UE to the core network the requirements were that the delay should be less than 5ms when there was no congestion on the air interface. In the case of transitioning from standby mode to active mode following a user initiated action the requirements state that the time taken should be less than 100ms.

Bandwidth, coverage areas and mobility also had their own set of requirements. LTE is optimised for users travelling at speeds up to 15Kmh, will work with high performance at speeds up to 120Kmh and also supports speeds up to 350Kmh. A variety of bandwidths can be used with LTE ranging from 1.4 MHz up to a maximum of 20 MHz. LTE cell sizes were optimised at 5Km, would work at sizes up to 30Km but with degraded performance and would support cells having a maximum size of 100Km.

Data packets within the evolved packet core (EPC) are routed using the Internet Protocol (IP) and devices using IPv4, IPv6 and dual stack IPv4/IPv6 implementations are supported. In

GSM/UMTS systems an IP based connection is only established on request and the connection is torn down when the session terminates. In contrast to this, in LTE the EPC establishes a basic IP connection for a device when it is powered on and connects to the network. This connection enables always-on connectivity with external networks and is only torn down when the device is powered off.

2.3 The IEEE 802.11 Family of Standards

In 1997 the Institute of Electrical and Electronics Engineers (IEEE) released the 802.11 standard [13], its genesis lay in the 1985 decision by the FCC to allow unlicensed access to the ISM band of radio frequencies. 802 is the IEEE general designation for network standards and the "11" suffix refers to standards dealing with wireless local area networks. IEEE 802.11 wireless LAN standards are a family of standards developed by group 11 of the IEEE LAN/MAN Standards Committee (IEEE802).

The 1997 original version of the standard specified two raw data rates of 1 and 2 Megabits per second (Mb/s) to be transmitted in the unprotected Industrial Scientific Medical (ISM) band of frequencies at 2.4GHz. The standard also specified the same raw data rates for transfer using infrared (IR) signals.

The media access method specified by the standard was Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) which was identical to the Ethernet media access method. After error handling and correction were applied the practical channel capacity was limited to approximately 65% of the theoretical maximum limit. The original 802.11 standard was soon extended with the 802.11b amendment that helped to popularise the 802.11 standard.

2.3.1 802.11b

The 802.11b amendment [14] to the original IEEE 802.11 standard was ratified in 1999. The 802.11b amendment increased the raw data transfer rate to a maximum theoretical rate of 11 Mb/s. However, due to the overhead imposed by the CSMA/CA scheme employed the maximum practical throughput achievable is approximately 5.9 Mb/s for TCP and 7.1 Mb/s for UDP. As is the case with the original standard, the 802.11b amendment also operates in the unprotected ISM 2.4GHz band of frequencies and is adversely affected by interference.

Products that supported 802.11b were quickly brought to market, significant reductions in price combined with a marked improvement in throughput (in comparison to the original 1 and 2 Mb/s of 802.11) saw the adoption of 802.11b as the definitive wireless standard of the period.

Wireless network interface cards designed for 802.11b can operate at rates of up to 11Mb/s maximum, however, in the presence of interference or poor channel quality they can reduce the transfer rate to 5.5 Mb/s, 2 Mb/s and finally 1 Mb/s if circumstances require. At the lower rates they are less susceptible to interference and signal attenuation due to the use of less complex and more redundant methods of data encoding.

2.3.2 802.11a

Another amendment to the 802.11 standard, 802.11a [15], was ratified in 1999. Unlike the original 802.11 and the 802.11b amendment, both of which operate in the ISM 2.4GHz band of frequencies, 802.11a is designed to operate in the 5GHz frequency band. 802.11a employs the same core protocol as the original standard but uses 52 sub-carrier Orthogonal Frequency Division Multiplexing (OFDM) to achieve a raw maximum data rate of 54 Mb/s. In practice, a realistic rate of approximately 24 Mb/s is achievable. As is the case with 802.11b the data transfer rate can be reduced under adverse wireless conditions. The fall-back rates for 802.11a are 48, 36, 34, 18, 12, 9 and finally 6 Mb/s. 802.11a equipment is not inter-operable with 802.11b equipment and vice-versa.

The 2.4GHz ISM band of frequencies does not require a license for use and is therefore very heavily used by a multitude of devices ranging from microwave ovens to remote control toys, each of which is a potential source of interference. Since 802.11a utilises the more heavily regulated 5GHz band of frequencies 802.11a compatible devices have much less interference to deal with. However, the 5GHz carrier frequency restricts operation to near line of sight requiring the deployment of more APs than would be the case with 802.11b. In comparison to signals transmitted by 802.11b equipment, signals emanating from 802.11a equipment are less able to penetrate walls, foliage and other obstructions leading to reduced coverage and an increased number of blind-spots.

2.3.3 802.11g

The 802.11g amendment [16] was ratified in June 2003, it utilises the same band of frequencies as 802.11b, that is the ISM 2.4GHz band. Although 802.11g is capable of a maximum raw data rate of 54 Mb/s, in practice a throughput of 24.7 Mb/s is likely to be achieved. 802.11g equipment and networks are fully backward compatible with 802.11b. A severe downside to this backwards compatibility is that in some older 802.11g networks the presence of a station using 802.11b can reduce the speed of the network to that of 802.11b.

Because 802.11g devices operate in the ISM 2.4GHz band of frequencies they are subject to the same sources of interference as 802.11b devices. On the surface, taking into account a raw data rate of 54Mb/s, 802.11g devices and networks seem capable of delivering a much higher throughput than 802.11b but the actual results achieved were not as impressive as might be expected. This is due to multiple factors including interference from other devices and conflicts with 802.11b equipment (mentioned previously), limited channelization having only three non-overlapping channels and a higher data rate that is often more susceptible to interference than 802.11b. Often the data rates achieved by 802.11g are very close to those achieved by 802.11b.

2.3.4 802.11n

802.11n-2009 [17] improves on 802.11g by using various techniques such as MIMO and channel bonding. Multiple-input Multiple-output (MIMO) is used to increase the capacity of the radio link. MIMO uses multiple transmitters and receivers at both the AP and the station to increase the overall amount of data transmitted. The technology leverages multipath, a natural radio phenomenon, which occurs when a transmitted radio signals are bounced off multiple surfaces resulting in the transmitted signals reaching the destination antenna at slightly different times and on slightly different paths. Previously, multipath was a source of interference that degraded radio performance, however, MIMO increases the data rate by enabling the antennas at the receiver to combine signals arriving at different times and on different paths. 802.11n operates in both the 2.4GHz and 5GHz frequency bands and provides a significant increase in the maximum net data rate from 54 Mbits/second to 600 Mbits/second. Channel widths have doubled from the 20MHz of 802.11g to 40MHz; the increase in channel width has been achieved through the use of channel bonding. Channel bonding is a technique that allows separate channels to be combined into a larger single

logical channel. In 802.11n the two 20 MHz channels to be bonded into a single 40 MHz must be adjacent. Doubling the channel bandwidth enables twice the data throughput.

2.3.5 802.11ac

The 802.11ac 2013 amendment [18] defines modifications to the 802.11 physical and MAC layers which support a multi-station throughput of at least 1 Gbit/second and a maximum single link throughput of at least 500 Mbits/second. Some of the technologies implemented in 802.11ac are extensions of advances introduced in the preceding 802.11n amendment such as MIMO and channel bonding. 802.11ac uses a greater number of antennas, an increased number of spatial streams and wider RF channels. The new channel widths are 80MHz and 160MHz in contrast to the 40MHz channel bandwidth see in 802.11n. The 80MHz channel width is mandatory, the 160MHz channel size is optional and again, channel bonding is employed in order to realise the required channel widths. A channel width of 80MHz is achieved through bonding two adjacent 40MHz channels. The 160MHz channels are defined as two 80MHz channels that may or may not be contiguous. 802.11ac also introduced new mechanisms such as MU-MIMO or Multi-user MIMO which enable the AP to transmit simultaneously to multiple stations. All previous 802.11 communication sessions had been either point-to-point (AP to station or station to AP) or broadcast (AP transmitting beacon frames). In 802.11ac the AP can transmit different streams to different stations simultaneously through the use of beam-forming. MU-MIMO does not of itself increase the amount of bandwidth available to individual stations rather it increases the networks overall utilisation by enabling data transmissions to multiple stations at the same time.

2.3.6 802.11ax

The IEEE 802.11ax standard [19], also known as Wi-Fi 6, builds on the strengths of 802.11ac, and it couples the freedom and high speed of gigabit wireless with the predictability found in licensed radio such as LTE. It allows enterprises and service providers to support new and emerging applications on the same Wireless LAN (WLAN) infrastructure, while delivering a higher grade of service to older applications. IEEE 802.11ax access points support more clients in dense environments and it provides more predictable performance for advanced applications such as 4K video, Ultra HD, wireless office, and Internet of Things (IoT). Flexible wake-up time scheduling lets client devices sleep much longer than is the case with

802.11ac, and wake up to less contention, extending the battery life of smart phones and other battery constrained devices. IEEE 802.11ax offers the following improvements:

- Denser modulation using 1024 Quadrature Amplitude Modulation (QAM), enabling a more-than-35-percent speed burst
- Orthogonal Frequency Division Multiple Access (OFDMA)-based scheduling to reduce overhead and latency
- Robust high-efficiency signalling for better operation at a significantly lower Received Signal Strength Indication (RSSI)

In addition, 802.11ax is a dual-band 2.4-GHz and 5-GHz technology, thus enabling 2.4-GHz-only clients to gain some of its benefits without modification. 802.11ax 2.4-GHz also supports significant increases in the range of Wi-Fi, adding standards-based sounding and beam forming. Importantly, 802.11a/g/11n/11ac monitoring and wireless intrusion protection systems (Wireless Intrusion Protection Switching (WIPS)) can continue to decode most management frames such as beacon and probe request/response frames, even when sent in the new 802.11ax packet format. IEEE 802.11ax has designed for maximum compatibility, coexisting efficiently with 802.11a/g/n/ac devices. Its new preamble (HE-SIG-A/B) follows the traditional 802.11a/g/n/ac preamble and extensions to request-to-send/clear- to-send (RTS/CTS) procedures for multiuser to help avoid collisions with older single-user mode users.

2.3.7 802.11be

A candidate for the next amendment of the IEEE 802.11 standard is IEEE 802.11be Extremely High Throughput (EHT) [20] which may be designated as Wi-Fi 7. 802.111be is expected to be built on 802.11ax and focuses on indoor and outdoor operations at both stationary and pedestrian speeds utilising the 2.4, 5 and 6 GHz frequency bands. The 802.11be amendment is being developed on an ongoing basis, with an initial draft expected by March 2021, and a final version predicted for early 2024. The primary candidate features mentioned in the 802.11be Project Authorization Request (PAR) are:

- 320 MHz bandwidth and more efficient utilization of non-contiguous spectrum
- Multi-band/multi-channel aggregation and operation

- 16 spatial streams and Multiple Input Multiple Output (MIMO) protocols enhancements
- Multi-Access Point (AP) Coordination (e.g. coordinated and joint transmission),
- Enhanced link adaptation and retransmission protocol (e.g. Hybrid Automatic Repeat Request (HARQ))
- If required, adaptation to regulatory rules specific to 6 GHz spectrum
- Integrating Time-Sensitive Networking (TSN) extensions for low-latency real-time traffic (IEEE 802.11aa)

In addition to the features mentioned in the PAR, there are many newly introduced features including the following examples:

- Newly introduced 4096-QAM (4K-QAM)
- Contiguous and non-contiguous 320/160+160 MHz and 240/160+80 MHz bandwidth
- Frame formats with improved forward-compatibility
- Enhanced resource allocation in OFDMA
- Support of direct links, managed by an access point

It has been predicted that Wi-Fi 7 speeds will be four times that of Wi-Fi 6 at approximately 30 Gbps with the first Wi-Fi 7 chipset likely be made available for testing in 2021.

2.3.8 802.21

IEEE 802.21 [21] is a standard that is concerned with developing algorithms to enable seamless handovers between networks of the same type (horizontal handovers) or between networks of different types (vertical handovers). It is the 802 standard for handover services and although it can be applied to either type of handover it is primarily aimed at vertical handovers.

The standard aims to assist with handover initiation, network selection, and interface activation. Key benefits are described as enabling optimum network selection, providing seamless roaming to maintain connections and lower power operation for multi-radio devices. 802.21 uses multiple services to optimise vertical handovers, the key services are Link Layer Triggers (state change, predictive triggers and network initiated triggers), Network Information

(available networks, neighbour maps and network services) and Handover Commands (client initiated, network initiated and vertical handovers).

Link Layer Triggers or events used include link up or link down, link going down (predictive) and changes to link parameters. Network initiated events may refer to load balancing or other network operator operations. The use of these triggers is intended to minimize connectivity disruptions during link switching.

The Media Independent Information Service will rely on 802.21 Information Servers being deployed; the servers will hold a Global Network Map, a list of all available networks, neighbour maps and information on higher layer services. This will enable information about all available networks to be obtained via a single radio e.g. a cellular radio used to indicate the presence of suitable Wi-Fi networks. A common format for information representation can be used across different networks. The Information Service can help with network discovery and selection leading to more effective handover decisions.

IEEE 802.21 is concerned with facilitating handovers between networks and various categories of handovers are supported. These include

- Terminal Controller Handovers in which the terminal makes use of some Media Independent Handover (MIH) services
- Terminal Initiated, Network Assisted Handovers in which the terminal makes use of the MIH Information Service

Network Initiated and Network Controlled Handover in which the network makes use of MIH Event and Command Services in addition to Information Service knowledge to decide if a handover is required or desired, to decide the target network and to command the terminal to undertake the handover.

A mobile node will be able to detect whether or not an 802.11 network supports MIH functions through information contained within the 802.11 Beacon frames. The new standard is designed for use with both existing and evolving network technologies.

2.3.9 802.11 Channel and Frequency Usage

The ISM frequency bands (2.4GHz and 5GHz) are utilised by the 802.11 standards [22] as follows. The ISM 2.4GHz band is used by 802.11b, 802.11g and 802.11n-2.4, while the ISM 5GHz band is used by 802.11a, 802.11ac and optionally 802.11n.

Spectrum in each ISM band is divided into channels with each channel being assigned a centre frequency and spectral mask. The channel's spectral mask determines the effective bandwidth of the channel and the channels are overlapping. In the ISM 2.4GHz band there are 14 channels with the centre frequency of each channel being separated from its neighbours by 5MHz. Channel 1 is centred on a frequency of 2.412 GHz, Channel 2 is centred on a frequency of 2.417GHz and so on. Not all channels are available for use in every country, depending on the country in which the technology is being deployed some channels might be restricted or unavailable.

The spectral mask assigned to each channel requires the signal for the channel to be attenuated by at least 30dB from its peak power at +/- 11MHz from the centre frequency and by -50dB at +/- 22MHz from the centre frequency. This gives each channel an effective width of 22MHz. It is sometimes assumed that because the spectral mask only defines power output restrictions up to +/- 22MHz from the centre frequency that the channels energy does not extend beyond the +/- 22MHz limit. In reality, if a transmitter is sufficiently powerful the signal can have a negative, interfering impact beyond the boundary of the spectral mask.

Although it is often stated that Channels 1, 6 and 11 do not overlap this is not strictly true. It would be more precise to say that given the separation of the channels the attenuation on one channel should be sufficient to present minimal interference with a signal on another channel. This depends, of course, on the transmitter power of device using the other channels.

2.4 IEEE 802.11 Network Overview

The IEEE 802.11 [23] standard allows for interacting components that enables the building of a wireless local area network that can support mobile stations or nodes. The most basic element of this is the Basic Services Set (BSS).

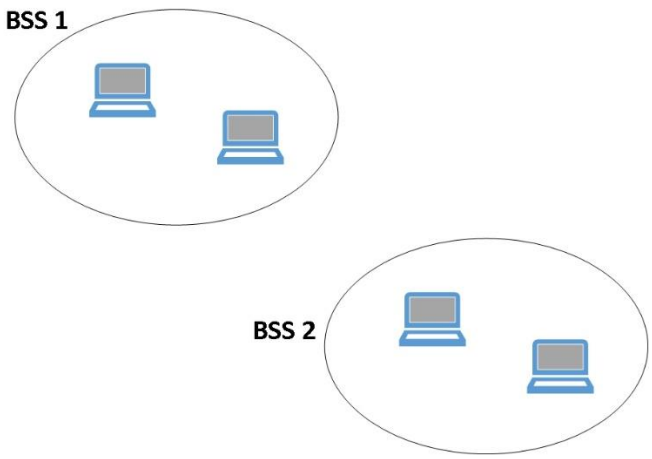

Figure 7 Basic Service Sets

Figure 7 depicts two Basic Service Sets (BSSs), each of which contains two stations or nodes. In wireless networking it is important to realise that well defined coverage areas do not exist. Due to its nature, the propagation characteristics of the wireless environment are both dynamic and unpredictable. Large differences in signal strength can result from small changes in station position or direction of travel. These effects can also occur regardless of whether a station is mobile or stationary. This is due to the fact that other moving objects, large vehicles for example, within the environment can have an impact on the propagation of signals between communicating stations. The oval shapes used to symbolise each BSS can be viewed as the coverage area of the Basic Service Set. Stations that are members of a BSS can communication with each other as long as they remain within the coverage area. If a mobile node moves out of its BSS it will not be able to communicate directly with the other members of the BSS.

2.5 802.11 Network Components

Although multiple amendments to the original 802.11 standard [23] now exist they share common design features. Each infrastructure type network has four basic physical components (Figure 8).

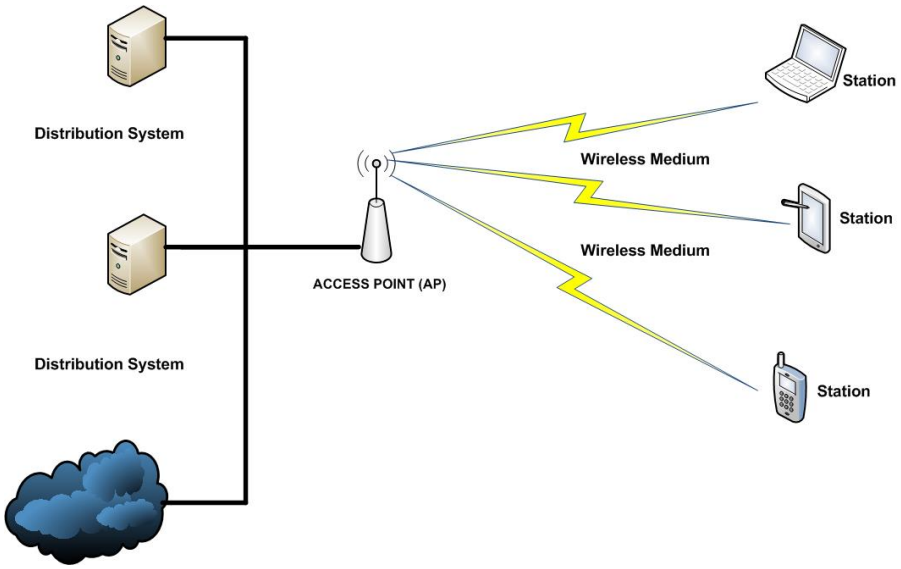

Figure 8 802.11 Network Components

Distribution System - the physical backbone network that carries data between Access Points (APs) or between an AP and the core network

Access Points (AP) - APs provide a bridge between the wireless and wired portions of a network. Frames designed for transport over the wireless media cannot be transported over the wired portion of the network without modification. The distribution system is typically, but need not be, Ethernet based and the AP must convert traffic from one frame type to the other in order to enable the exchange of data. APs also perform many other functions but this bridging/conversion function is perhaps the most important

Wireless Medium - in all 802.11 networks data is transferred over wireless medium, the frequencies employed depend on the version of 802.11 in use

Stations - stations are devices equipped with wireless network interface cards capable of sending and receiving data over the wireless medium. Typically, stations connecting to wireless networks are mobile/portable devices such as laptops, tablets, smart-phones or some other hand-held computer. However, there is no requirement for a station to be a portable device. In situations where the deployment of a wired network infrastructure is not feasible desktop computers, printers, scanners, servers, etc. may be connected through a wireless LAN with little or no mobility.

2.6 Basic 802.11 Network Types

The Basic Service Set (BSS) is the basis of an 802.11 wireless network [23] and the Basic Service Area (BSA), defined by the propagation characteristics of the transmission medium, is the space within which communications can take place [23]. Stations within the basic service area can communicate with each other. There are two types of basic service set (BSS)

- Independent Basic Service Set (IBSS) (Figure 7)
- Infrastructure Basic Service Set (Figure 10)

2.6.1 Independent BSS (Ad-hoc)

Stations in an IBSS or ad-hoc wireless network communicate directly with each other without the services of an AP (Figure 9). In order to be able to communicate stations must be in range of each other. The smallest ad-hoc network possible consists of two stations. Typically, IBSS or ad-hoc networks are created for a specific purpose and have a relatively small number of members e.g. to support a meeting by enabling the exchange of documents. When the meeting ends the ad-hoc network is disbanded.

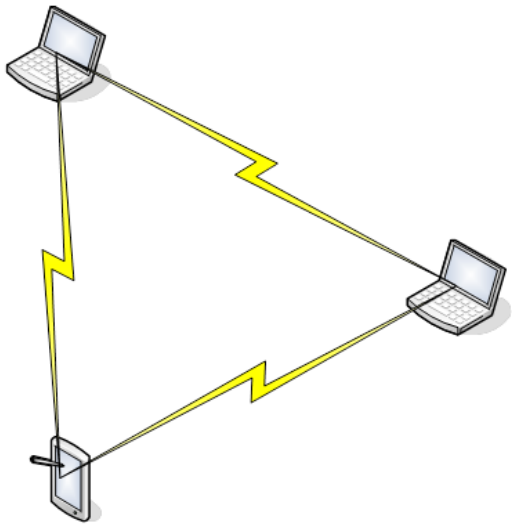

Figure 9 Ad-hoc Wi-Fi network

2.6.2 Infrastructure Basic Service Set

An infrastructure BSS differs from an IBSS in that the services of an AP are used. All communications within the infrastructure BSS take place via the AP (Figure 10). If a station wishes to communicate with another station the frame is sent to the AP and the AP forwards it on to the target station. Replies from the target station are sent to the AP and then onto the originating station. No direct communication between stations takes place. In an infrastructure BSS the service area is defined by the distance from the AP at which a station can receive a transmission.

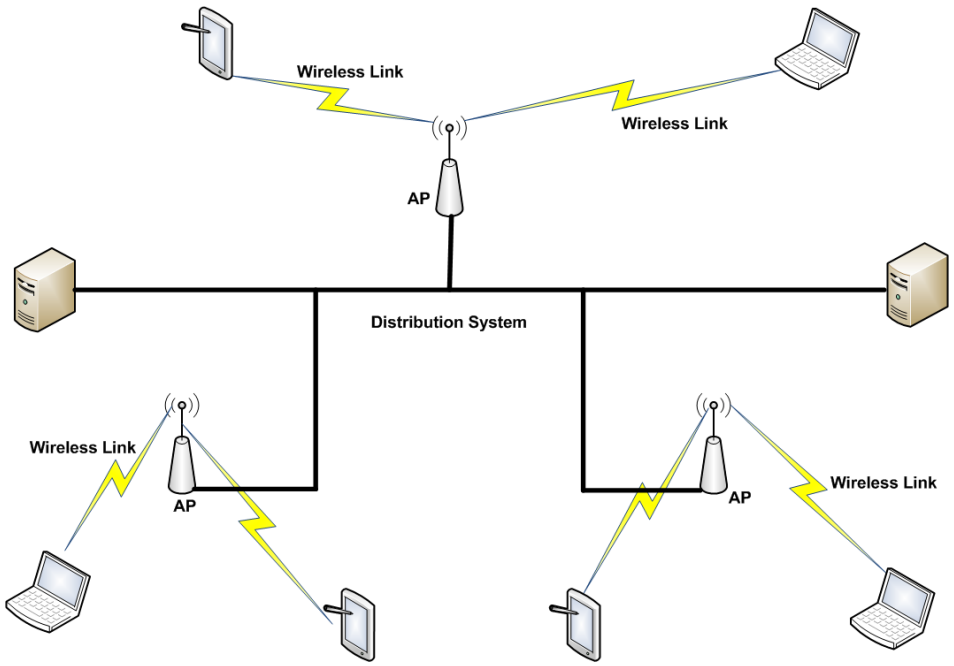

Figure 10 Infrastructure mode Wi-Fi network

Multi-hop communications within an infrastructure BSS may appear to be sub-optimal but it does present certain advantages. The use of Access Points enables battery powered devices to employ energy conservation strategies while continuing to send and receive frames. Stations can shut-down their wireless interfaces to conserve energy and power up at predetermined intervals to send data and receive frames buffered at the AP. This behaviour requires the AP to be aware of the energy saving scheme and to be prepared to buffer frames for the hibernating station.

Although stations can only communicate with each other via the AP there is no restriction on how far apart the stations are as long as they remain in range of the AP. This increases the geographical spread of the stations considerably as the AP acts as a relay between stations.

When operating within an infrastructure BSS stations must associate with an AP in order to gain access to network services. When the station associates with the AP it, in effect, joins that particular wireless network. The process of association is always initiated by the 'client' or station; an AP may allow or deny the association request based on the contents of the request. A wireless interface card belonging to a station may only associate with one AP at a time, but there is no restriction to the number of wireless interface cards per station. In reality, a station may be associated with multiple APs via multiple wireless interface cards.

There is no limit set by the 802.11 standard [23] with regard to the number of stations that might be associated with a single AP, however the practicalities of operating within shared spectrum do force limits on the number of stations that will be permitted to associate with an AP.

2.7 Quality of Service (QoS)

Broadly speaking Quality of Service (QoS) can be defined as a description or measurement of the overall performance of a service such as a telephone system or computer network and in particular the performance as seen by the system or network users [24]. In the context of the performance of a network service it can be assessed quantitatively using error rate, bit rate, availability, jitter, throughput, transmission delay, etc.

The basic concepts of QoS in the context of a network can be illustrated through the following example. Nodes on the network host various applications some of which exchange data with other applications running on other network nodes. Data is passed from the application layer to the lower layers before being put onto the network. In this context QoS refers to the ability of the network to handle the frames placed onto it so that the service requirements of the communicating applications are met. In order to be able to do so the network requires the following:

- A fundamental traffic handling mechanism
- The ability to identify different types of traffic to ensure that it should be considered by the traffic handling mechanism
- The network must be able to control the traffic handling mechanisms

The rate at which applications generate traffic varies greatly depending on the functions carried out by the application. Applications require the network to be able to handle their traffic at the rate at which it is being generated. All applications can tolerate some level of delay in the network and some level of variation in the delay itself. However, an application that uses VoIP will tolerate far less delay and jitter than an email application can. Some applications such as video streaming can tolerate some level of data loss while others such as file transfer programs cannot. These requirements can be represented by QoS parameters such as:

- Bandwidth – the rate at which an application's traffic must be carried by the network
- Latency - the amount of delay in traffic delivery tolerated by the application
- Jitter – variation in latency
- Loss – percentage of lost data an application can tolerate

Since network resources are finite there will be occasions on which there will be insufficient network resources available to meet demand. When network resources are insufficient to meet current demand congestion occurs. Network devices may react to congestion in one of two ways:

- Store excess packets in temporary buffers until the congestion subsides
- Discard packets to reduce congestion

As a result of these behaviours applications will experience either delays or data loss during periods of congestion. The traffic forwarding capacity of network interfaces and the availability of buffers for temporary storage are the resources required to provide QoS on a network. Internal decision making mechanisms on network devices prioritise traffic to determine which packets get access to these resources. Network devices that support QoS do so by making intelligent decisions on which types of traffic to allocate resources to. When congestion occurs a device might decide to queue traffic from a delay tolerant application in a local buffer and to immediately forward on traffic belonging to an application that cannot tolerate delays. In this case memory resources are allocated to the delay tolerant traffic and interface and capacity resources are allocated to the traffic that is less delay tolerant, of course the device also has the option to discard all traffic if necessary.

In order to be able to prioritise traffic in this manner it is necessary to classify the traffic in some way and to associate each type of traffic with certain resources. Resource allocation can be achieved by separating traffic arriving at a device into separate queues based on traffic classification. An algorithm that handles queue-servicing determines the rate at which traffic from each queue is allowed onto the network. This determines the amount of network resources allocated to each traffic class. The greater the amount of available resources allocated to a particular class of traffic the better its QoS. End-to-end QoS depends on the contributions made by each of the networks components on the path from source to destination.

2.8 Quality of Experience (QoE)

Quality of Experience has been defined as "the degree of delight or annoyance of the user of an application or service. It results from his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state" [25]. It has also been defined as "the overall acceptability of an application or service, as perceived subjectively by the end-user" [26].

Quality of Experience can be extremely difficult to measure as a great many seemingly unrelated influence factors impact on it. For example, producers of multimedia content such as movies and video games create "experiences" and may attempt to influence the viewers/consumers emotional state through the manipulation of images, the use of emotional story lines and the use of sounds and music to build atmosphere or tension. At the senders (the creators) end "meaning" is related with the creator's intentions while at the "receivers" end the "meaning" is drawn from experiencing and interpreting the content. It is very likely that the meaning developed by the receiver will differ significantly from the meaning intended by the creator. According to the authors of [25] an Influence Factor (IF) is "any characteristic of a user, system, service, application or context whose actual state or setting may have influence on the Quality of Experience for the user". They also insist that Influence Factors must not be regarded in isolation as they are often inter-related.

Influence Factors (IF) are grouped by the authors into three categories:

1) Human Influence Factors – which are "any variant or invariant property or characteristic of a human user. The characteristic can describe the demographic and socio-economic background of the user, their physical and mental constitution or the user's emotional state". Human IFs are complex and strongly inter-related
2) System Influence Factors – the "properties and characteristics that determine the technically produced quality of an application or service". They are related to media capture, coding, transmission, storage, rendering, reproduction/display and the communication of information.

System IFs can be divided into four sub-groups:

a) Content Related System IFs – content type and content reliability e.g. colour depth, texture, 2D/3D, etc.
b) Media Related System IFs – media configuration factors e.g. resolution, sampling rate, frame rate, media synchronisation, etc.
c) Network Related System IFs – data transmission over the network, bandwidth, delay, jitter, data loss, error rate, throughput
d) Device Related System IFs – end systems or UE as well as network devices along the end-to-end communication path, device capabilities such as screen size, resolution, CPU, battery life time, audio, etc.

3) Context Influence Factors – factors that "embrace any situational property to describe the user's environment in terms of physical, temporal, social, economic task and technical characteristics" [33]

2.8.1 Differences between QoS and QoE

The differences between QoS and QoE can be reduced to a number of factors:

1. Scope – telecommunications are typically the focus of QoS while QoE is applied to a wider domain. QoE can be applied to areas that do not involve communication systems, for example HD video played in a domestic setting
2. Focus – QoS is focused on the performance aspects of physical communications systems. QoE on the other hand is focused on the end-user's assessment of system performance influenced by factors such as the user's own expectations, the content and the context in which the media is consumed among many other factors
3. Methods –QoS can be assessed in a reasonably straight-forward way using an analytical approach and empirical measurements. QoE assessment is far from straightforward and requires a multi-disciplinary and multi-methodology approach

It must be remembered that QoE is, in a great many instances, very dependent on QoS and various components of a system can have a large impact on some aspects of QoE. [27]

2.9 Multimedia Streaming

Video streaming methods can be broadly divided into two categories, adaptive and non-adaptive streaming methods. Early streaming methods were non-adaptive in that they did not alter the bitrate of the content based on the current amount of bandwidth available to the client device. Video content was delivered using various transport protocols such as the Real-time Transport Protocol (RTP) [28] and the Real-time Streaming Protocol (RTSP) [29].

RTP provides end-to-end network transport functions suitable for applications transmitting real-time data. These applications include audio and video delivered over unicast or multicast services if supported by the underlying network. The protocol is not concerned with either Quality of Service (QoS) or resource reservation in the network. Typically, RTP runs on top of UDP in order to make use of UDPs checksum and multiplexing services. RTP is not confined to using only UDP and can be used with any other suitable network or transport protocol. RTP has no mechanisms for ensuring timely delivery of packets and does not provide any QoS guarantees. There is no guarantee of packet delivery, no prevention of out-of-order delivery and no assumption that the underlying network will deliver packets in sequence. RTP includes separate sequence numbers that enable the receiver to reconstruct the sender's original packet sequence. The sequence numbers may also be used to identify a packets correct location within a stream without having to decode the packets in sequence.

RTP consists of two closely linked components, namely:

1. RTP itself which is used to carry data having real-time properties
2. The RTP Control Protocol (RTCP) which monitors QoS and also conveys information regarding participants in on-going sessions. RTCP also helps synchronise multiple streams.

RTP can compensate for jitter through the use of buffers and by using sequence numbers can detect out of sequence packets both of which are common occurrences on IP networks. The protocol is used in conjunction with an associated profile and payload format.

Real-time Streaming Protocol (RTSP) [29] is an application layer protocol developed to control real-time data deliveries. It establishes and controls either single or multiple time-synchronised streams of continuous media such as audio and video. A presentation description

defines the set of streams to be controlled. There is no concept of an RTSP connection, a server maintains a session identified as being an RTSP session. The RTSP session is not tied to a transport layer connection such as TCP. In the course of an RTSP session an RTSP client can open and close multiple reliable connections to the RTSP server and request data. An RTSP client may also choose to use an unreliable, connectionless protocol such as UDP for data transport. RTSP controlled streams may use RTP but RTSP does not depend on the protocol. Although RTSP is similar in syntax and operation to HTTP/1.1 it differs in a number of important aspects. These differences include:

- An RTSP server must by default retain state in almost all cases unlike HTTP
- Both an RTSP server and client can issue requests
- Data is typically carried out of band by a different protocol
- RTSP is defined to use ISO 10646 rather than ISO 8859-1

In the context of RTSP a presentation is a set of one or more media streams presented to the client as a complete media feed using a presentation description. The presentation description contains information on one or more media streams within the presentation such as information on the content, network addresses and the set of encodings in use. An RTSP URL may be used to identify each media stream and presentation.

A presentation description file defines the overall presentation itself as well as the properties of the media which makes up the presentation. The presentation description file does not need to be stored on the media server and can be obtained by the client using a variety of means such as email or HTTP download. A description of the media streams that make up the presentation including language, encoding and other parameters that enable the client to select the most appropriate combination of media are contained within the presentation description file. The presentation description also identifies each media stream that is controllable by RTSP.

These media streams are identified by means of RTSP URLs which point to the media server controlling that particular media stream. The RTSP URL is also used to name the media stream on the server. For the purpose of load sharing several media streams may be located on different servers, for example audio can be stored on one server and video content on another. The various transport modes supported by the server are also outlined in the description file.

In addition to the media parameters the destination address and port number to be used must also be determined and several modes of operation are supported by RTSP:

- Unicast – a port number selected by the client and the source address of the RTSP request are used by the server to stream media to.
- Multicast (server chooses the address) – the server decides on the multicast port number and address. This is typical for a live streaming event.
- Multicast (client chooses the address) – if the server is to take part in an existing multicast event the multicast address, port number and encryption key are given by the conference description.

2.10 Adaptive Bitrate Streaming

Adaptive bitrate (ABR) [30] [31] [32] streaming is a method of streaming multimedia content over communications networks that takes into consideration fluctuations in the amount of available bandwidth. While protocols such as RTP and RTSP ran on top of TCP and UDP adaptive streaming technologies run almost exclusively over HTTP. The basic idea with adaptive streaming is that an adaptive bitrate enabled client monitors the available bandwidth and adjusts the requested media stream to match. This approach requires the source material to be provided in multiple different bitrates. The client application switches between the various available source material bit rates depending on the amount of bandwidth available on the link. Adaptive bitrate streaming can reduce the amount of buffering required, speedup the initial playback and provide good QoS for both high bandwidth and low bandwidth links.

The adaptive bitrate streaming implementations currently in use encode the source content at multiple bitrates. Each version of the content is divided in segments of short duration and depending on the particular implementation the duration of the segments can vary but are typically between 2 seconds and 10 seconds in length.

A client streaming application begins operations by first obtaining, usually by download, a manifest file. This manifest file contains details of the different available bitrates and the segment durations. The first segment requested by the client is typically a low bitrate segment, if the client determines during the download that the link capacity is greater than the segment bitrate it will request the next higher bitrate segments for subsequent downloads. A later reduction in available bandwidth would result in the client requesting segments at lower

bitrates, the client attempts to download the highest bitrate segment that the current bandwidth will support. Requesting a low bitrate segment as the initial segment also reduces the amount of time the end user must wait until media playback begins.

Adaptive bitrate streaming provides good end user experience as the client player will always attempt to select the highest bitrate media segment supported by the current available bandwidth. Because adaptive bitrate streaming typically runs over HTTP there is little difficulty in traversing firewalls or NAT enabled devices. Since all operational logic is on the client side the requirement for persistent connections between server and client is reduced. No session state information for client's needs to be held on the server which makes it less complicated to scale. Adaptive bitrate systems based on HTTP have the potential to be more operationally complex than systems implemented using RTP and RTSP. This is due to the need to encode the source material at multiple bitrates and at multiple segment lengths, essentially creating multiple copies of the same content on each server. Increased storage requirements, the work of the encoding operations and the difficulty in maintaining consistent quality across all delivery systems are all factors that need to be considered. However, in practice the potential for additional complexity has been nullified by adaptive bitrate's lower costs and better scalability. Adaptive bitrate streaming technology runs over HTTP and can therefore make use of the same HTTP web server used to deliver most of the content on the Internet thereby greatly reducing deployment costs.

Adaptive bitrate streaming has been implemented in MPEG-DASH, Adobe HTTP Dynamic Streaming, Apple HTTP Live Streaming, Microsoft Smooth Streaming and many others. However, at the time of writing MPEG-DASH [35] is the only HTTP adaptive bitrate technology that has been ratified as an international standard (ISO/IEC 23009-1:2012) in April 2012.

Despite all of their benefits ABR technologies are not without their limitations particularly in relationship to latency. For example, it is often the case that if you view a live action event such as a soccer match on both a traditional TV broadcast and an over-the-top (OTT) streaming service you will observe the video stream being several seconds behind the TV broadcast. The delay or lag in the video stream can be as high as 18 seconds or more but is on average less than 10 seconds [34] and is frequently long enough for comments on the action to appear on social media before the action takes place on the video stream. This is not an ideal situation especially for sports fans and reduces their enjoyment of the event.

In the context of live ABR streaming we can define latency as the degree of delay between the time the camera captures the action and the time the action is viewed on the user's screen. As is the case with non-live video streaming ABR technologies segment the content in segments of varying duration as part of the streaming process. Video stream latency is caused by many factors including:

- Video encoding pipeline duration
- Ingest and packaging operations
- Transport protocol used and network propagation time
- Content Delivery Network (CDN)
- Segment length selected
- Media player settings such as buffer level

Although video latency is the result of many different factors the media player itself can be responsible for much of it. Latency in live video streams will never be eradicated but it can be significantly reduced with some effort. Scalable lower latency can be achieved over HTTP using standard DASH technology.

2.10.1 Dynamic Adaptive Streaming over HTTP (DASH)

Dynamic Adaptive Streaming over HTTP (DASH) [35], also known as MPEG-DASH is an ABR technique that enables the streaming of appropriately formatted content over the Internet from standard HTTP servers. As is the case with other ABR strategies the video content to be served is broken into segments of various durations and bitrates, and in essence multiple versions of the content are stored on the server. When a video prepared for MPEG-DASH distribution is requested a manifest file called a Media Presentation Description (MDF) file is first downloaded. The MPD file is an XML document that describes information regarding the available video segments, this information includes segment duration, segment bitrate, timing, etc.

Based on the current link conditions the DASH enabled media player will select the video segment with the highest bitrate supported by the network connection in use. If the amount of available bandwidth reduces the bitrate of the requested segments is also reduced, conversely if the amount of available bandwidth increases the bitrate of the requested segments

also increases. This strategy enables the media player to adapt to fluctuations in available bandwidth and reduce the number of stalling or re-buffering events.

MPEG-DASH is an international standard and should not be confused with a transport protocol — the transport protocol that MPEG-DASH uses is TCP. It uses the existing HTTP web server infrastructure that is used for delivery of essentially all World Wide Web content and is both audio/video codec agnostic.

2.11 5th Generation Systems (5G)

Many organisations and industry leaders forecast huge increases in the amount of network traffic due to the demand for and delivery of video and multi-media content. It is also predicted that rapid and ubiquitous Internet of Things (IoT) development and deployment will help drive the increase in network traffic. With an increase in demand for network capacity and with a rise in mobile data consumption comes an increase in energy consumption and capital expenditure for service providers. 5G is being touted as solution to the problem of how to meet the need for faster, higher capacity networks [36].

In a simplistic sense proposed 5G systems will build upon the foundation provided by current 4G systems. 5G is an evolution that has been described as a "convergence of Internet services with legacy mobile networking standards leading to what is commonly referred to as the 'mobile Internet' over Heterogeneous Networks (HetNets), with very high speed broadband. With regard to 5G nothing can be considered to be certain, indeed the very term 5G is considered by some to be nothing more than hype or marketing. Nevertheless, some already proven technologies such as cognitive radio (software defined radio), small cell deployment, co-operative systems, Self-Organising Networks (SONs) and green Multi-mode RF seem to be solid candidate technologies for 5G components.

Mobile devices such as the smartphones and light weigh tablets that drive the continuously rising levels of mobile traffic have become ubiquitous. These devices and the media streaming services that serve them enable the consumption of High Definition (HD) video almost anywhere and at any time. In addition, social media sights such as Facebook see video delivery as a key component of their offerings and push video and graphic content towards their huge user base.

This growing trend towards video as a primary component of Internet traffic will inevitably increase the demand for additional capacity on networks. It should also be remembered that advertisements consume a significant amount of bandwidth on both fixed and mobile networks. Delivery of advertising content also has a negative impact on end user experience as it increases load times for web pages and consumes some portion of a user's data plan that the user must pay for.

Growth in mobile data traffic far exceeds growth in voice traffic and has done so since 2009. Currently it is estimated that Voice over IP (VoIP) accounts for less than 0.5% of all traffic on mobile networks. In addition to rapidly increasing levels of traffic the communications networks of mobile service providers face yet another challenge. The predicted growth in the numbers of interconnected devices, in the form of the Internet of Things (IoT), will result in end users being tracked, monitored and served by tens, if not hundreds of machines.

In order to support human activities, regulate human habitats, enable smart cities, assist in transportation systems and inter-vehicle networks devices will need to communicate and cooperate with each other. Machine-to-machine (M2M) communications of this nature will require very stringent latency of less than 1ms.

The authors of [36] identify what they consider to be the key building blocks for 5G some of which are outlined in the following section.

2.11.1 Evolution of existing Radio Access Technology (RATs)

It is highly unlikely that 5G systems will employ a specific radio access technology, it is reasonable to assume that future systems will consist of further evolutions of existing RATs and novel systems. If this is to be the case then it would make economic sense to start addressing the forthcoming shortfall in network capacity by improving existing RATs in terms of spectrum efficiency (SE), energy efficiency (EE) and latency. It may also be necessary for service providers to advocate or support the notion of radio access network (RAN) sharing controlled through some centralised mechanism. Current widely deployed RATs include Wi-Fi, 3G/4G and LTE/LTE-Advanced.

2.11.2 Hyper-Dense Small Cell Deployments

Hyper-Dense Small Cell Deployment represents another approach to solving the coming capacity requirements while bringing additional energy efficiency to the system. This approach also known as a HetNet can assist in improving an area's spectral efficiency (bits/second/Hz/m^2). Broadly speaking HetNets can be implemented in one of two ways that is Multi-tier HetNet in which a cellular system is overlaid with small cells of the same technology i.e. micro, pico or femto cells. The second option is known as multi-RAT HetNet and consists of overlaying a cellular system with small cells of differing technologies i.e. Wi-Fi or Wi-MAX.

Qualcomm has demonstrated [37] that adding small cells can scale the capacity of a network in a linear fashion with the network capacity doubling each time the number of cells doubles. This approach is not without its drawbacks, reducing the size of the cells increases the inter-cell interference and the associated required control signalling. In order to combat the increase in inter-cell interference advanced inter-cell interference management techniques are necessary at the system level in conjunction with complementary interference cancelling methods at the user equipment (UE).

Small cell deployment was the focus of LTE R-12 [38] in which the New Carrier Type (NCT) aka the Lean Carrier was implemented in order to enable the host macro-cell to assist small cells. This approach allows increase efficiency in the control plane for mobility management, resource allocation, synchronisation, etc. through the macro-layer while enabling a spectrally efficient and high capacity data plane through the small cells. In addition, reduced cell sizes can help improve energy efficiency by reducing the physical distance between the UE and the network edge which shrinks the power budget of the wireless links.

2.11.3 Self-Organising-Network (SON)

Networks having the ability to self-organise are considered to be a key component of 5G systems. The increase in the density of small cell networks means the need for self-organising networks gains momentum. It is predicted that large amounts of wireless traffic will be generated indoors and in order to be able to handle this volume of traffic very dense deployments of small cells will be necessary. In many cases these small cells will be installed and maintained by the end-users and will not be controlled by service providers. To enable this

scenario indoors, small cell devices will need to be simple to install ('plug n play' model) and be self-configuring. These small cell devices will also need to be self-organising in order to cooperate in an intelligent, co-operative manner with neighbouring small cell devices in order to minimise inter-cell interference.

2.11.4 Machine Type Communications (MTC)

Another aspect of 5G will be the need to facilitate inter-device communication. Machine Type Communications (MTC) describes a communication session in which one or both end-points is a machine. This type of communication introduces two major challenges in that the number of devices will be extremely large and the latency on the communications links must be extremely low. Ericsson [39] forecasts that the number of connected devices on its future networks will exceed 16 billion stating that "anything that can benefit from being connected will be connected". An increasing demand for real-time services and the ability to remotely control mobile devices and machines such as vehicles over the network will require extremely low latency rates for MTC. A latency rate of less than one millisecond, enabling the 'tactile Internet' [40] will require a 20x improvement in latency from 4G to 5G.

In order to achieve this low latency rate fibre to the home/premises is necessary while keeping the physical distance of the wireless radio access link as short as possible.

2.11.5 Development of Millimetre-wave Radio Access Technology

Current RATs are approaching Shannon's capacity limit and spectrum below 3GHz is increasingly congested. In response to this situation research into exploiting cmWave and mmWave bands for mobile communications is being actively pursued. The use of mmWave for mobile communications faces a multitude of challenges including the fact that at these wave lengths path loss is relatively higher in comparison to sub 3GHz bands. Blocking of signals by moving people, objects and vehicles also presents a serious challenge. Higher penetration loses for mmWave signals means that indoor users may not be able to avail of RATs situated out of doors.

In spite of its shortcomings the use of mmWave communications does present certain advantages. For example, mmWave bands provide a large amount of spectrum, at 60GHz there

is 9GHz of available unlicensed spectrum, in stark contrast to the global allocation of spectrum for all mobile communications of approximately 780MHz [36]. A second advantage is that mmWave communication enables the use of small antenna sizes with small separation distances between them. This permits multiple antennas to be placed in a small area, ideal for mobile devices.

However, the use of mmWave bands for outdoor wireless communications faces serious obstacles. MmWave signals may be subject to significant levels of attenuation during periods of heavy rain due to the fact that raindrops are approximately the same size as the wavelength (mm) leading to scattering. Foliage loss is high mmWaves limiting propagation during certain times of the year in areas with extensive vegetation such as wooded parks and suburban neighbourhoods. Use of mmWaves for outdoor communications may require the support of a backup cellular system operating in legacy sub 3GHz bands [36].

2.11.6 Network Slicing

Network slicing [41] is a technique that enables a network operator to create dedicated virtual networks over a common network infrastructure that are tailored to provide the specific functionality required by a service or customer. This technique employs the same principles behind software defined networking (SDN) and network functions virtualisation (NFV) in fixed networks. It supports the creation of multiple virtual networks on a common shared physical infrastructure which can then be customised to meet the specific needs of applications, services, devices, customers or operators. In the context of 5G, a single physical network will be sliced into multiple virtual networks that can support different radio access networks (RANs), or different service types running across a single RAN. Network slicing may be implemented in the RAN but it is expected to be implemented primarily to partition the core network.

Each virtual network (network slice) consists of an independent set of logical network functions that support the requirements of the particular use case. They are optimised to provide the resources and network topology for the specific service and traffic that will use the slice. Complete isolation between the slices insures that no slice can interfere with the traffic in another slice reducing the risks associated with introducing and running new services. This approach also helps improve security, in the event that a cyber-attack breaches one slice the attack is contained and cannot spread beyond that slice. Network slices are transparent to the

user and the user experience of the network slice will be the same as if it was a physically separate network.

2.11.7 5G Service Based Architecture (SBA)

5G Service-Based Architectures (SBA) [42] provide a modular framework from which common applications can be deployed using components of varying sources and suppliers. The 3GPP defines a Service-Based Architecture (SBA), whereby the control plane functionality and common data repositories of a 5G network are delivered by way of a set of interconnected Network Functions (NFs), each with authorization to access each other's services. Network Functions (NFs) are self-contained, independent and reusable and can take on the role of either a producer of services or a consumer of services. A Service Based Interface (SBI) is used to expose the functionality of each Network Function service. The SBA uses a centralized discovery mechanism that takes advantage of a NF Repository Function (NRF). Records of available NF instances and their supported services are held in the NRF. Other NF instances can subscribe and be notified of registrations from NF instances of a specified type. Service discovery is supported by the NRF through receiving Discovery Requests from NF instances and the NRF maintains a record of which NF instances support specific services. The Unified Data Management (UDM) provides services to other SBA functions, such as the NEF. The UDM holds information in local memory, however, it may also be stateless, storing information externally within a Unified Data Repository (UDR). The UDM provides authentication credentials while also being employed by the Access and Mobility Management Function (AMF) and the Session Management Function (SMF) to retrieve subscriber data and context.

2.11.8 Redesigning Backhaul Links

The improvements to Radio Access Network (RAN) technology necessary to reach the capacity levels and transmission speeds required for 5G cannot take place in isolation. For the system to operate at the required level backhaul links need to be re-engineered in order to be able to handle the massive increases in the amount of data traversing the networks. Failure to do so would result in the backhauls themselves becoming bottlenecks. The challenge becomes greater as the number of hyper-dense small cell deployments increase as each cell will require a backhaul link of some type. To meet capacity demands and latency targets

fibre will be required for ‘last mile’ links to homes and businesses. Alternative communication mediums such as TV whitespace will also have an important part to play as reliable backhauls that don’t interfere with cells or RANs.

2.11.9 Energy Efficiency

Energy efficient design is necessary in all aspects of future communication systems from RAN and backhauls to end user equipment. Energy efficient design provides multiple benefits, reducing the amount of energy required to transmit data can help reduce costs to end users even as data rates improve and the amount of data transmitted increases. Improved battery capacity and longevity reduces e-waste and enables mobile users to remain connected for longer. Mobile operators can increase revenue by reducing OPEX through energy savings. Intelligent user equipment (UE) can provide the best possible quality of experience (QoE) while reducing energy consumption. In developing nations where the greatest growth in user numbers can be achieved electricity supplies outside the major urban areas can be very unreliable where they exist at all and charging mobile device can be relatively expensive. In these cases, energy efficiency is of particular importance.

2.11.10 Allocation of new Spectrum for 5G

Wireless communications in the coming years will require the use of new spectrum. It is highly unlikely that the expected increase in data traffic can be met solely through improvements in spectral efficiency and the deployment of hyper-dense small cell networks. Some telecom companies estimate that up to ten times more spectrum will be necessary to meet the demand for network capacity.

2.11.11 Spectrum Sharing

The allocation of new spectrum for wireless communications is a slow and painful process. In order to meet user needs efficient use of available spectrum is vital. Cognitive or software defined radio might be employed to exploit under-utilised spectrum such as TV whitespace. Radio Access Network (RAN) Virtualisation - Virtualisation of the RAN would allow sharing of the wireless infrastructure among multiple operators. Network virtualisation needs to be pushed from the wired core i.e. switches and routers out to the edge of the network. Sharing infrastructure among operators would help reduce both operational expenditure (OPEX)

and capital expenditure (CAPEX) for the operators concerned. RAN virtualisation would require some form of centralised control for operational intelligence and a convergence between wired and wireless components of the system. Access points containing multiple technologies and software defined networking would create an immensely flexible network.

2.12 5G Architecture

The vision for 5G (Figure 11) is that of a converged system consisting of multiple communicating technologies that support a wide and varied range of applications. These applications are expected to include multi-GB per second mobile Internet, vehicle-to-vehicle (V2V) networking, vehicle to infrastructure communications, Machine Type Communications (MTC), public safety applications, etc.

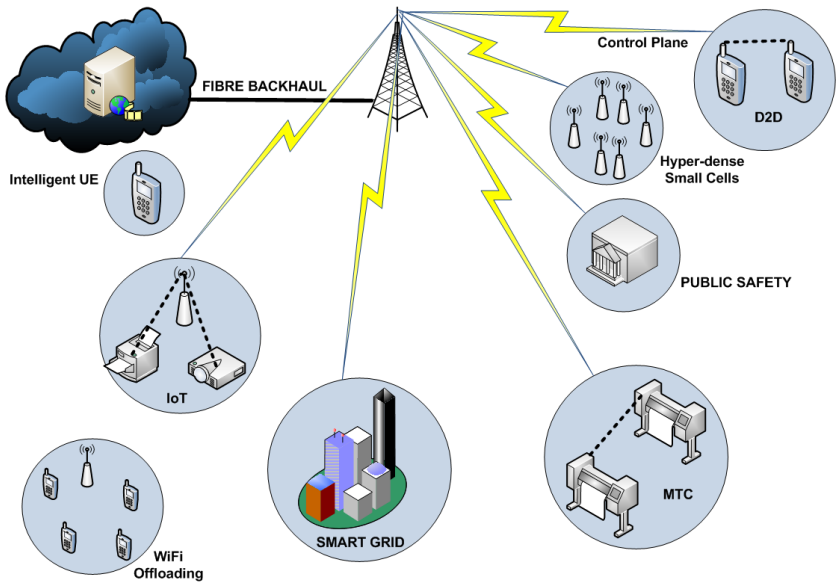

Figure 11 5G Architecture

CHAPTER 3 RELATED WORK

In the previous chapter, existing standards for wireless communications, mobile telephony, multimedia streaming, QoS, QoE and adaptive bitrate streaming were introduced. This chapter presents the state-of-the-art research projects in fields related to the work of this thesis. In order to be in a position to address the problem outlined in Chapter 1, it is necessary to show the state of current research solutions before presenting the novel solutions of this thesis.

3.1 Introduction

This chapter presents a review of the research in the following areas: wireless network detection, wireless network selection, energy conservation in mobile devices, data offloading, mobility patterns and rates of speed in order to take advantage of the opportunities for mobile connectivity presented by heterogeneous wireless environments we must first detect available networks within range of our mobile device. Network detection has been an active area of research for many years and network detection strategies have become ever more sophisticated as the number and type of deployed networks has increased over time.

3.2 Network Detection

In order for a Wi-Fi enabled device to connect to a wireless network it must first find a suitable network or networks. This is achieved through a process known as scanning, scanning activities can be classified as being either active or passive. In active scanning the UE transmits Probe Requests on the appropriate channels and then waits for Probe Responses from Wi-Fi access points within range. In the event that the UE does not receive a Probe Response within a predetermined period of time it moves to the next channel in the sequence, transmits a Probe Request and again waits for a Probe Response. If no Probe Responses are received on any of the available channels the UE cycles through the channels again and will continue to do so until either the battery dies or the user intervenes.

In passive mode the UE simply listens on each channel for beacon frames that are transmitted by available APs at regular intervals, the beacon frames are typically sent every 100ms. Passive scans can take longer to complete than active scans since the UE must wait and listen for a beacon frame as opposed to actively probing for available APs. If the UE listens on a

channel shortly after a beacon frame has been sent and misses it the UE can potentially be waiting another 100ms before the next beacon frame is sent. Another possible issue for UEs engaged in passive scanning is that if they don't wait for a long enough period of time on a channel they run the risk of switching channels before the next beacon frame is transmitted. Passive or active scanning for available wireless networks is the first step in the network discovery process and is therefore inevitable. The scanning process has been shown to be the most significant contributor to Wi-Fi connection delays and has been the focus of a great deal of research over the years.

Mishra, Shin and Arbaugh in [43] describe their work in analysing the IEEE 802.11 MAC Layer handoff process. They divide the entire handoff delay into three separate sub-delays which they describe as:

1) Probe delay – the period of time taken to send a Probe Request and either receive a Probe Response or have the relevant timers expire
2) Authentication delay – the delay incurred during the exchange of authentication frames between the UE and the access point (AP)
3) Association delay – the latency incurred during the exchange of association frames

Experiments were conducted using an indoor wireless network with the aim of accurately measuring the handoff delay. Measurements were taken on two co-located wireless networks using standard APs from two different vendors. The wireless cards used by the UEs came from three different suppliers. The results obtained from the experiments showed that the most significant contributor to the handoff delay was the probe delay component. The probe delay was found to account for more than 90% of the overall handoff delay. This was regardless of the combination of APs and wireless interfaces used in the experiments. The authors also observed that even in the number of frames exchanged between the UE and the AP handoff the probe phase accounted for 80% of the frames in all cases. It was also found to be the case that the hardware used in the experiments i.e. the various APs and wireless cards also affected the handoff delay. Large variations in delays were seen with each one of the multiple combinations of hardware used. Regardless of the variation in the overall length of the handoff delay the probe phase was always responsible for the greatest part of the delay.

Pi, et al [44] examine the issue of why establishing a connection to a Wi-Fi AP takes such a long time. The authors conducted measurement studies on 5 million mobile phone users in four cities associating with 7 million APs in 0.4 billion Wi-Fi sessions to better understand the Wi-Fi setup process in the real world. Data was captured using the popular "Wi-Fi Manager" application on Android based smart phones. It showed that in urban areas as many as 45% of mobile devices fail in establishing a Wi-Fi connection to an AP within range and that 15% of successful Wi-Fi connections take longer than 5 seconds to become established. The "Wi-Fi Connection Setup Cost" is defined by the authors as "the time span between the moment a user clicks on the SSID (service set identifier) name of the AP s/he wishes to connect to and the time his/her device obtains the IP address". It is recognised that obtaining an IP address does not guarantee access to the Internet. Four sub-phases of the overall connection process are identified:

1) Scan – the purpose of the scan phase is the detection of APs within range of the UE
2) Association – necessary in order to establish a connection the association process follows a set pattern of authentication request, authentication response, association request and association response. The four packets used are sent and received in sequence. When the UE has received the association response the MAC-Layer connection has been established. The authentication request/response component is a legacy feature of the older WEP standard.
3) Authentication – a common feature in modern Wi-Fi environments, it consists of a four packet MAC-Layer handshake
4) DHCP – the client UE interacts with a DHCP server to obtain an appropriate IP address

The time cost associated with each of the sub-phases identified was recorded using a modified version of the Wi-Fi Manager app. The modified application which was equipped with Wi-Fi association breakdown was deployed to 12, 472 selected devices which generated 706,000 connection attempts. The data collected showed that 45% of Wi-Fi connection attempts failed for a variety of reasons including but not confined to Timeouts (14%) and DHCP failure (9.4%). Approximately 15% of the connection attempts examined took longer than 5 seconds to complete with 5% taking longer than 10 seconds. For UEs experiencing large connection time events the scan sub-phase was found to be the main culprit.

The authors reported little difference in connection time patterns between Wi-Fi networks using the 2.5 GHz and 5 GHz frequency bands. There was also little difference in connection time costs for users roaming in enterprise wireless networks and connection time costs for users in their home networks.

Pi, et al, propose a machine learning based AP selection algorithm that classifies candidate APs into one of two classes, FAST or SLOW. Technical features of the various AP models are used as inputs into the machine learning model. The selection algorithm actively avoids APs that have been classified as SLOW. Measurement tests that compared a baseline algorithm which simply used signal strength as a selection criterion with the author's algorithm were conducted. Evaluation of the results demonstrated that using the authors approach could reduce connection failures from 33% to 3.6% and 80% of connection costs could be reduced by 10X.

In [45] Castignani ,et al, present the results of experiments on the discovery process in 802.11 networks which focused on the length of time a UE must wait before receiving a response from an AP. The authors characterise the Wi-Fi discovery process using two metrics full scanning failure and full scanning latency. They define full scanning failure as the failure to discover any point of attachment on any available channel. Full scanning latency is the total amount of time taken to scan all available channels. Full scanning latency is represented as a function of MinCT, MaxCT and the probability of finding activity on a particular channel.

MinChannelTime (MinCT) and MaxChannelTime (MaxCT) are two timers defined in the IEEE 802.11 standard. MinCT and MaxCT determine the amount of time that a UE must sit on a channel awaiting a Probe Response having sent a Probe Request on that channel. The MinCT timer defines the maximum amount of time a UE must wait for the first Probe Response. If a response is not received at the UE within MinCT the channel is considered to be empty and the UE restarts the process on the next channel in the scanning sequence. On the other hand, if a Probe Response is received within the MinCT the UE remains on the channel until the MaxCT timer has expired in order to check for additional Probe Responses that may have been sent by other APs on that particular channel.

The authors conducted a set of experiments in both a simulated environment and a physical test-bed. In the experiments they examine the impact of both MinCT and MaxCT on full scanning latency and full scanning failure. They developed two strategies for setting the

timer values during the experiments. Their first strategy was based on the use of fixed time periods for the timers, a similar approach to the one taken by some existing open source 802.11 drivers such as MadWiFi and ath5k. The second strategy involved dynamically adapting MinCT and MaxCT values during the scanning process.

Every Wi-Fi discovery process is unique due to the multiple possible combinations of AP deployments and user speed and movement patterns, as a result it is not feasible to determine a set of 'best' values for MinCT and MaxCT. Castignani, et al, aim to find a trade-off between a minimal full scanning latency and a minimal full scanning failure. The basic idea is to dynamically lower MinCT and MaxCT values when a point of attachment is discovered on a channel and to increase the MinCT and MaxCT values if not point of attachment is found.

This approach enables the UE to reduce the amount of time spent on a channel once points of attachment have been discovered. In contrast a set of fixed MinCT and MaxCT times would cause the UE to spend the same amount of time on each channel regardless of the existence of any APs. By reducing the timer durations when candidate APs are discovered the overall, full scanning latency is reduced and a full scanning failure avoided. Increasing the timer durations when no candidate APs have been found introduces the danger of increasing the full scanning latency but gives the UE more time on each channel to listen for possible Probe Responses. This reduces the chances of missing a Probe Response and therefor reduces the risk of a full scanning failure.

The authors tested their proposed methods using both simulations and real world test beds. They observed that the adaptive strategy gave a better balance between full scanning latency and full scanning failure than the fixed timer strategy. Results showed that in almost all proposed scenarios the adaptive strategy offered a better percentage of candidate APs, minimised the number of full scanning failures (max 2%) and kept a low and controlled full scanning latency of between 190ms and 434ms.

The 802.11 family of standards subdivide their allocated frequency bands into smaller channels with the specific number of channels depending on geographic location. Subdividing the spectrum in this way presents two challenges:

- There are multiple channels that must be scanned in search of points of attachment during each discovery phase leading to large scanning delays

- There are few non-overlapping channels (3 in 2.4 GHz and 8 in 5 GHz)

A very simplistic way in which to reduce the overall scanning latency is to only scan a subset of all available channels instead of scanning all of the available channels. It is obvious that by scanning only half of the available channels would immediately result in a 50% reduction in overall scanning latency. One problem with this approach is that the UE runs the risk of never finding potentially available APs on the un-scanned channels. When taking the approach of only scanning a subset of the available channels the process for selecting the channels to include is of the utmost importance.

Shins, et al, in [46] develop a handoff procedure to reduce the MAC-Layer handoff delay. Through experimentation they determined that the scanning delay portion of the discovery process was responsible for 90% of the total handoff latency. Their work focuses on reducing the overall scanning delay by using a selective scanning algorithm and by reducing the number of time the selective scanning algorithm itself is required.

The selective scanning algorithm works as follows:

- When first invoked the algorithm sends Probe Requests all available channels and listens for Probe Responses
- A channel mask is set by turning on the bits for all channels on which a Probe Response was received. In addition, bits for channels 1,6 and 11 (non-overlapping channels) are also set as they are likely to be used by APs
- Select the 'best' AP (the one having the strongest signal strength) and connect to it
- The channel the UE is currently connected on is removed from the list of candidate channels by resetting the mask bit

In the event that no APs are discovered using the current mask invert the mask bits and rescan the channels indicated as eligible by the mask. If it is the case that no APs are detected clear the channel mask and scan all the channels

The authors state that using this algorithm reduced the handoff latency in the experiments to a value between 30% and 60% of the original handoff delay times. An AP cache was used to further reduce scanning latency; the cache held a list of neighbour or adjacent APs based on previously detected signal strength. If a handoff is required, the cache is checked for suitable APs to connect to. If the cache does not contain any previously detected APs the UE conducts a scan using the selective scanning algorithm, if a candidate AP is found in the

cache the UE attempts to connect to it. If the UE fails to establish a connection to the AP identified in the cache the selective scanning algorithm is invoked. The use of caching enabled the authors, under controlled lab conditions, to reduce delays by over 90%.

In [47] the authors employ a strategy of passive scanning as a means of reducing the overall scanning delay. They advocate that UEs should not, as is usual, wait for a disconnection event to occur or to suffer degraded performance before seeking a new point of attachment. In other words, UEs should be proactive and not merely reactive and should continuously monitor the performance of all APs operating on the UEs current channel and on all overlapping channels. By continuously monitoring beacon frames from APs and capturing long term trends in link quality the UE can make handover decisions before being forced to. Because the next point of attachment has already been selected scanning delay is greatly reduced. In the event that no AP is available on the current channel a full scan can be implemented to find the next point of attachment. The authors propose that that UEs handoff to APs operating on the UEs current channel. This behaviour is based on the result of experiments which demonstrated that in-band handoff is significantly shorter than scan based handoffs and result in low packet loss and delay variability. The authors acknowledge that choosing a point of attachment based on information regarding APs operating in the UEs current channel is not optimal. For example, an AP with far better signal strength characteristics could well be operating on a nearby channel but might not be detected. However, the authors feel that for delay sensitive applications the longer delay incurred by out of band scanning is unacceptable.

The 802.11's standard active scanning algorithm is triggered by deterioration in the strength of the signal received from the AP to which the UE is currently attached. Proactive monitoring of the metrics used to trigger scanning in UEs may also be used to prevent full scanning behaviour in 802.11. In [48] the authors suggest the use of higher than normal threshold values to be used in order to trigger the scanning process earlier than usual. The aim is to begin scanning while the UE is still attached to a usable AP and capable of transmitting and receiving data. However, Wi-Fi enable interfaces cannot send and receive data while engaged in scanning operations. In order to circumvent this restriction, the authors propose subdividing the scanning or discovery phase into sub-phases. The objective is to enable the UE to alternate between sending or receiving data and scanning for new points of attachment. Data is sent and received in the intervals between the scanning sub-phases.

The authors discuss two versions of the scheme, one in which all available channels are scanned and a second version in which only a subset of the available channels are scanned. In the first approach, known as smooth handoff, there is actually no reduction in the overall amount of time required for the scanning phase. It is simply interleaved with other operations and because it does not occur in a single contiguous block of time the scanning delay is less apparent to the end user. The second approach, known as greedy smooth handoff, does reduce the overall scanning delay. It achieves this by only scanning a subset of the channels and in the event that a suitable AP is found it connects to the AP without scanning any of the other channels. If no suitable AP is found in the subset of channels the UE continues and scans all available channels.

Both of these schemes require the UE to begin scanning operations earlier than is strictly necessary. In order to trigger this behaviour higher threshold values are used to force the UE into performing scans. This strategy can lead to an unnecessarily high frequency of scans and to combat this behaviour an adaptive algorithm is used to dynamically change the threshold value used as a scanning trigger. For example, when a UE first begins to scan it employs a high threshold value. If the signal strength of the AP to which it is attached falls below this artificially high threshold it triggers scans for nearby APs. The UE scans all available channels but finds that all detected APs report signal strengths below the current high threshold value in use. In order to address this situation, the threshold value is dropped to a lower value and the channels are rescanned. When the UE finds an AP to which it can connect using the new, lower threshold value it does so. Following a successful connection to the new AP the threshold value is increased with the expectation that the next AP discovered will have a better signal quality.

Early work by Velayos and Karlsson [49] focused on techniques to reduce handoff time in 802.11b networks. They analysed the handoff process by dividing it into three sequential phases which they named detection phase, search phase and execution phase. The detection phase was the period of time during which it became apparent that a handoff to another AP would be required. The search phase was concerned with discovering alternative points of attachment within range. In the execution phase the actual handoff itself was performed. The duration of each of the three phases was measured and it became clear that the detection and search phases accounted for most of the handoff time. In fact, the execution phase was so

short that the authors state that “its reduction will not significantly decrease the total handoff time”.

Velauos and Karlsson [49] point out that the need to handoff to another AP is detected at the link-layer after several non-acknowledged frames. The primary factor in determining the duration of the detection phase is the number of failed frames permitted before action is taken. It can vary between Wi-Fi card manufacturers because if a frame is not acknowledged the UE cannot determine the loss of the frame was due to a collision, congestion in the cell or simply the fact that the UE has moved out of range of the AP. A handoff can be initiated by either the AP or the UE and the subsequent actions taken will vary based on who initiated the handoff. In the case of the handoff being initiated by the network the detection phase is reduced to a single disassociation message sent by the AP. Handoffs initiated by the UE are by far the most common and in this case the detection phase includes the UE having to detect deteriorating radio link quality based on received signal strength reported by the physical layer or some number of failed frame transmissions.

The authors focus their study on optimising the detection phase based on the number of failed frames. As previously stated, the main issue is how to determine the cause of the lost frames. From their measurements the authors observed that the UE would initially assume collisions as being the cause and would react by retransmitting the frames at lower bitrates. In the event that retransmissions remained unsuccessful radio fading is assumed and probe requests are sent to check the link. It was only after several unanswered Probe Requests that the AP was considered to be out of range and the search phase began. The authors decided on a different approach, one that required the UE to initiate the search phase as soon as collisions could be ruled out as the cause of lost frames. If temporary signal fading was the cause the AP selected following the search phase would likely be the current AP and a handoff would not be executed. The authors determined that three consecutive collisions, even in saturated cells, was a very rare event.

Based on this observation they formulated their link layer detection algorithm as follows, “if a frame and its two consecutive retransmissions fail the UE can discard collision as the cause of failure and start the search phase, there is no need to explicitly probe the link”. The authors tested their detection algorithm under the same conditions as those used when taking their initial measurements and found it to be approximately 300 times shorter than the previous fastest measured detection phase. They also state that in order to achieve these results active

scanning must be performed. While they also suggest shortening the MinCT and MaxCT timers to reduce the overall scanning delays they do not suggest only scanning a subset of the available channels.

Ramani and Savage in [50] describe SyncScan, a low cost technique for continuously tracking adjacent APs by synchronising short listening periods at the UE with periodic beacon frame transmissions from the APs. Using SyncScan the authors propose to replace the large transient overhead incurred by actively scanning for APs with a scheme involving a continuous process of passively monitoring available channels for APs within range. The disruption to data transmissions caused by channel switching is minimised by synchronising the short UE listening periods with AP beacon frame transmissions.

Wi-Fi APs transmit beacon frames at regular intervals, typically every 100ms, in order to alert potential clients to their existence and to synchronise state information with already associated client UEs. At the core of SyncScan [50] is a staggered schedule of periodic beacon frame transmissions across all available channels. For example, APs on channel 1 would transmit their beacon frames at time t or as close to it as possible. APs on channel 2 would transmit their beacon frames at $t + d$, APs on channel 3 would transmit their beacon frames at $t + 2d$ and so on across all available channels.

UEs associated with an AP on channel 1 could detect APs on channel 2 by switching to channel 2 d ms after receiving a beacon frame from the AP on channel 1. The UE can use this pattern of behaviour to locate all APs within range while minimising the amount of time that it is out of contact with its associated AP. When the need for a handover arises the overall delay is reduced to that of authentication and association. SyncScan also offers the opportunity to make better handoff decisions by continuously monitoring the signal strength of multiple APs rather than that of the AP to which it is currently connected.

The authors acknowledge that SyncScan does introduce additional complexity; the synchronisation of beacon frame transmission times and client listening periods requires accurate clocks. Over short time frames the authors claim that clock drift is negligible for commercial grade APs, over longer time frames clock synchronisation is required and the authors suggest the use of the Network Time Protocol (NTP). Synchronisation presents its own potential risks, for example multiple APs operating on the same channel run the risk of all transmitting their beacon frames at the same time leading to collisions and the loss of frames. To combat

this potential problem, the beacon generation time can be randomly varied over some small window e.g. 3ms. A client UE sitting on the channel for the entire window period could expect to receive most of the beacon frames on the channel.

SyncScan has a hidden cost, although it removes the transient overhead associated with the irregular full scanning phase it introduces a smaller but regular overhead. When a client is listening on other channels it cannot send or receive data on its own channel. It may also miss transmitted frames while checking other channels leading to the possible need for missed frames to be retransmitted.

Ramani and Savage state that the most obvious benefit of their scheme is the substantial reduction in handoff times from around 400ms to just a few milliseconds. Continuous scanning can aid in the discovery of APs with better signal strength before the currently connected APs signal strength degrades to a point below the threshold value. This approach enables handoffs to be made earlier avoiding the potential disruption caused by being forced to begin scanning operations due to a loss of connectivity.

The authors of [51] introduce a strategy to reduce the handover delay by reducing scanning delay. Their approach relies on the idea that APs that are nearer to the UE should have higher signal strength. The authors propose the use of a Media Independent Information Server (MIIS) which stores the channel numbers that APs are using along with the co-ordinates at which each AP is sited. The UE requests the information from the MIIS and when a handoff is triggered the UE calculates its current location and then determines which of the APs is physically closest. The UE then sends a Probe Request on the channel used by its nearest AP. By only probing a single channel scanning delay is greatly reduced. In the event that this approach fails the UE can revert to standard scanning behaviour. The author's strategy assumes that MIIS servers are available for use.

Chang, et al., in [52] propose two enhancements to the active scanning mechanism in order to reduce overall scanning delay based on the IEEE 802.11ai standard. IEEE 802.11ai is a fast initial link setup (FILS) amendment intended to enable a UE to achieve secure link setup in less than 100ms. Within a FILS environment the authors seek to improve the effectiveness and reliability of the scanning process using their Enhanced Active Scanning Scheme 1 and Enhanced Active Scanning Scheme 2.

Enhanced Active Scanning Scheme 1 operates in the same way as active scanning with some enhancements aimed at reducing the number of transmitted management frames. The enhancements must be implemented at both the UE and the AP. Scheme 1 operates as follows, when an AP receives a Probe Request it will not send a Probe Response if:

- An appropriate Probe Response is already queued in the transmission buffer waiting to be sent or transmission of a beacon frame is scheduled.
- The AP itself is already in the 'filter list' included as part of the Probe Request. The filter list is a list of all APs from which the requesting UE has already received a response.

Enhanced Active Scanning Scheme 2 operates in the same way as Scheme 1 with one additional enhancement. When an AP receives a Probe Request it will reply with a Probe Response that contains information on all neighbour APs. When an AP receives a Probe Response from another AP it will flush Probe Response messages from its own buffer. The authors tested their enhanced scanning schemes using simulations, results showed that the proposed enhanced scanning schemes are generally more efficient than existing schemes. There was also a significant reduction in the number of management frames transmitted.

The 802.11 standard requires a Mobile Node (MN) to scan all the possible channels to discover available Access Points (AP's). In an attempt to reduce interference several channels will be empty and due to having to make a full scan the MN ends up wasting time scanning empty channels, which results in a high connection delay. This connection delay can be reduced by simply confining the scanning procedure to a limited set of channels, i.e. those currently being used by nearby access points (AP). In [53] the authors propose Intelligent Scan which uses the Media Independent Information Server (MIIS) to inform the MN of the channel configuration information of the network or the surrounding AP's. In the proposed scheme the MN acquires channel configuration information from MIIS server and then uses that information to scan a sub-set of channels being used by the surrounding access points. Only scanning a sub-set rather than scanning all possible channels results in reduced handover delay. The authors define a parameter NET_CHANNEL_CONFIG which is passed in the query by the MN to the MIIS, so that the MIIS knows which information is being requested by the MN.

3.3 Network Selection

A mobile user operating in a heterogeneous multi-network environment equipped with a device capable of connecting to the various technologies deployed in the area must decide on which available network they wish to use. Network selection strategies consider many metrics when deciding which of the available wireless network is 'best' suited to the needs of a particular user. Many network selection strategies have been proposed over the years but they do not typically address the issues of mobile user speed. It has been demonstrated that the rate at which a mobile user is travelling has a significant impact on the amount of data they can transfer within a heterogeneous multi- network wireless environment.

Network selection schemes for multi-homed mobile devices in heterogeneous multi-network wireless environments must also be as efficient as possible. Schemes that require the collection of and processing of multiple criteria will take longer to reach a decision that less complicated schemes. Mobile users in environments in which technologies with limited ranges, such as Wi-Fi, are deployed will have less time to detect, analyse and connect to a network. Even at modest pedestrian speeds the window of opportunity for network detection and selection is small.

Over the years there has been a wide variety of approaches to developing network selection schemes based on various input metrics and decision making mechanisms. Regardless of the approach taken all network selection mechanisms can be classified as either network centric or user centric.

In many ways the problem of network selection for network centric approaches is one of resource allocation. That is directing the end-user towards networks that have the greatest number of resources available at that particular moment. In the case of network centric solutions some type of centralised control is required to amalgamate information on current network conditions, conduct analysis, reach a decision and recommend candidate networks to end users. This approach assumes that cooperation can be achieved between wireless networks and end user's; it also incurs additional overhead costs from the transmission of information and coordination messages. Users are required to cooperate with the centralised controller and possibly with each other. However, computation, analysis and decision mak-

ing take place within the network, this reduces the workload on user equipment at the expense of introducing latency. It also provides a network wide view of wireless conditions rather than a narrow localised one.

3.3.1 Network Selection Strategies

The authors of [54] describe a network centric decision making process used to determine 'candidate networks' that can provide the best Quality of Service (QoS) to the end-user. The proposed system employs a mixture of compensatory and non-compensatory multi-attribute decision making (MADM) algorithms in selecting networks. Various factors that can impact on the selection of the 'best' network for the end-user are identified as inputs to the decision making process. The non-compensatory MADM is invoked first; its operation is simple and can be described as "the removal of network alternatives from the candidate list that are not suited to the scenario". This approach reduces the number of possible candidate networks to be examined by the more sophisticated compensatory MADM which takes tuneable parameters as some of its inputs.

The following steps are involved in a compensatory MADM algorithm

1) Identify all alternatives and compensatory MADM attributes that impact on the decision making process
2) Assign a relative importance in the decision making process to each of the identified attributes
3) Use a compensatory MADM algorithm to develop ranking for the alternative networks

In this work the TOPSIS algorithm is employed as the compensatory MADM algorithm in order to arrive at the 'best' solution. The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) algorithm is based on the assumption that the solution arrived at is the 'shortest distance' from the ideal best solution but is the 'longest distance' from the worst possible solution. Data collection and all analysis take place on the network, only the network rankings are transmitted to the client device. The client device selects the network to use based on the rankings received; this reduces the amount of work that needs to be done on the client and has advantages for underpowered devices.

In [55] Lee et al. proposed the implementation of their vertical handover decision making algorithm using multiple Vertical Handoff Decision Controllers (VHDC). VHDCs provide the vertical handoff decision function for areas covered by single or multiple Wi-Fi access points (APs) or cellular base stations (BS). The authors envisage the use of the Media Independent Handover Function (MIHF) of IEEE 802.21 to facilitate message exchanges between access networks. These messages would carry information regarding prevailing link-layer conditions, traffic loads, network capabilities, etc., with the AP MIHF containing user equipment (UE) battery life information.

The network selection process is set in motion on receipt of a Link Layer Trigger (LLT) at the VHDC. LLTs regarding user equipment typically indicates one of the following conditions:

- While connected to an AP the RSS at the UE dropped below a pre-defined threshold value
- While connected to a cellular BS the RSS at a particular AP exceeded some specified trigger level

The basic concept is that if a UE can be supported in a location by either a BS or an AP then the UE is switched to the AP.

Ting Bi, et al., proposed in [56] a reputation based network selection mechanism. The network characteristics experienced by a mobile user vary depending on time and on user location within the wireless network. For example, users located closer to the AP or BS will generally have a better connection than a user at the edge of the coverage area and users connecting to the network at a 'quiet' time of day will have experience better performance. User mobility within a heterogeneous wireless environment means that the user will frequently require network selection and handover procedures in order to maintain satisfactory connectivity to the Internet. Due to the impact of user positioning on the level of performance experienced it is difficult to determine the best available network based on data supplied by a single user.

The proposed Reputation-based Network Selection (RNS) solution addresses this issue through the use of the IEEE 802.21 MIH standard mechanism. RNS supports gathering data on network delivery characteristics from multiple users located throughout the wireless environment. Using a MIH Information Server the information supplied by the collaborating

UEs is aggregated and disseminated to users upon request. Having data from multiple locations within a wireless network enables better network selection decisions to be made. The RNS decision making process uses user location, signal strength and delay information as inputs.

In order to be able to select an appropriate network to connect to a UE requires a list of candidate networks and their calculated quality ratings. IEEE 802.21 MIH provides a mechanism that supports gathering and exchanging information between candidate networks, the MIH Information Server and the UE. The Network Reputation Algorithm (NRA) computes a network reputation value based on data supplied by multiple users. The NRA is calculated at the level of the network sector and as a result it provides higher precision within that sector. An Overall Network Reputation Algorithm (ONRA) amalgamates the various sectoral reputation values to generate an overall network reputation value. A Localisation Prediction Algorithm (LPA) employs user location information to estimate the user's route and the user's future positions relative to network coverage areas. Based on the NRA and the LPA the Network Selection Algorithm (NSA) determines the best candidate network for the user to connect to.

The Access Network Discovery and Selection Function (ANDSF) [57] was first defined in 3GPP Release 8 in 2008. ANDSF enables mobile operators to define how the Evolved Packet Core (EPC) of a cellular network can be accessed via non-3GPP networks.

Three types of information that can be transmitted to UEs are defined:

Inter-system Mobility Policy (ISMP) which specifies the network type to be connected to when only one wireless interface is to be used. In reality this will be either LTE or Wi-Fi

Inter-system Routing Policy (ISRP) which in the event that multiple interfaces can be used simultaneously specifies which type of traffic should be sent over each network

Discovery Information, a list of non-3Gpp networks that details the availability of non-3GPP networks in a devices location

Release 8 provides mobility policies to enable the operator to guide the UE in selecting the appropriate Access Point in a given location at a given time. However, Release 8 does not support simultaneous connections to multiple networks. As a result, the Inter-System Mobility Policies (ISMP) was defined to be independent of the type of traffic being transmitted.

If a UE is connected to a Wi-Fi AP all traffic is sent over Wi-Fi and if the UE is connected to a base station all traffic is sent over the cellular network.

Release 10 saw the implementation of Multi-Access PDN Connectivity (MAPCON), IP Flow Mobility (IFOM) and non-seamless Wi-Fi offloading. These techniques enable multiple simultaneous connections to be established to multiple radio access technologies. In order to accommodate these changes, the ANDSF framework was extended to include the Inter-System Routing Policy (ISRP). The ISRP enables operators to indicate preferred or forbidden radio access technologies depending on the type of traffic the UE is transmitting.

An ISRP can be based on:

- The Packet Data Network (PDN) identifier used by the UE for a particular connection
- The destination IP address of traffic sent by the UE
- The destination port of traffic sent by the UE
- A combination of the previous three

Service providers make network selection decisions for the UE based on the operator's requirements e.g. in order to reduce congestion on a particular network or to prioritise certain classes of traffic.

Trestian, et al., in [58] propose a network selection algorithm which bases its decisions on the estimated energy consumption in a bid to be energy efficient while providing a satisfactory level of service to the end user. The proposed solution selects the best "value" network capable of meeting the user's needs based on the device type, application requirements, user preferences and prevailing network conditions. Results from testing showed that the proposed solution achieved a good trade-off between energy consumption, monetary cost and network load while acting in the user's best interests.

3.4 Utility Functions

Utility functions are well-known multi-criteria decision making methods. A utility function summarises the preferences of a consumer in terms of how utility they get from consuming the goods or services in the utility function. [59] Utility can be described as a measure of the value an individual derives from some good or service. For many years, utility functions were the domain of economists but they have long been used in various areas of network research, in particular the area of network selection.

Ormond et al. [60] introduce an intelligent utility- based strategy for network selection in multi-access network environments. They examine a number of utility functions which explore different user attitudes to risk for cash and delay preferences related to the application currently in use and demonstrate that risk takers willing to spend more money get better service.

The authors of [61] propose a user centric intelligent solution for network selection in multi-network environments. A utility function is used during the process of selecting inputs into a classic multiple attribute decision making system. The authors state that the proposed solution achieves a well-balanced trade-off among user preferences, network conditions and service application to the benefit of both the end users and the Radio Access networks.

Liang et al [62] propose a utility-based network selection scheme for use in integrated CDMA cellular/WLAN networks. The proposed solution takes into account the network resource, QoS requirements of applications, user mobility, and vertical handoffs between networks. The goal of the proposed scheme is to guide users to select the most suitable access network, where the users' QoS requirements can be satisfied with the lowest cost to the user. Simulation results demonstrate the effectiveness of the proposed network selection scheme in integrated CDMA cellular/WLAN networks

The authors of [63] focus on the Network Selection problem in integrated heterogeneous wireless environment consisting of various Radio Access Technologies (RATs). The problem is one of how to allocate terminals to most appropriate RATs by jointly examining both users and providers' preferences. They introduce a utility-based optimization function and formulate the terminal assignment problem as an optimization problem which is recognized as NP-hard. A Greedy heuristic is proposed which exploits a metric that measures the utility gained versus the capacity resource spent for each allocation. Testing is conducted to confirm its superior performance against three Bin-Packing heuristics.

3.5 Energy Conservation

Every mobile or portable device requires some form of energy supply. At present the most commonly used type of portable power supply takes the form of a rechargeable battery. Research into alternative, more efficient portable power supplies is very active but it may take

many more years before they become commercially available. Batteries being used at present are of finite capacity and this factor places a huge constraint on the operation of mobile devices.

The period of time during which a device can carry out useful work is determined by the energy reserves contained within the battery and the rate at which the energy reserves are being consumed. Energy consumption rates depend on the number and types of tasks being carried out by the device. The relationship between the energy reserves and the consumption rate is a very simple one, the higher the consumption rate the quicker the energy reserves are depleted.

This presents a very serious problem for users of mobile devices. Battery capacity has evolved very slowly in comparison to developments in other areas of computing and communications. R.A Powers in [64] states that battery capacity had been doubling every 35 years while in comparison, processor capabilities have been almost doubling every 18 months. Batteries have been improving at approximately 5% per year and unfortunately do not follow Moore's Law as have processors. The materials used in battery development are dangerous to store and work with, and development and testing of new batteries takes years instead of months and significant advances are difficult to achieve. Unlike CPUs badly designed or manufactured batteries are liable to explode or catch fire with serious consequences which also slows down development.

It is clear that the functionality and complexity of mobile devices has outstripped the capabilities of the batteries that power them. The use of wireless network interfaces to enable network connectivity has only exacerbated this situation. In addition, the current trend towards thinner, lighter devices restricts the size of the battery that can be used which further limits the battery capacity.

Mobile devices are commonly equipped with multiple wireless adapters in order to enable them to take advantage of the connectivity opportunities offered by heterogeneous multi-network wireless environments. In order to be able to take advantage of the opportunities available the interfaces must be activated and in this state they continuously consume energy reserves.

A great deal of research has been conducted into the amount of energy consumed by the various components and subsystems that make up mobile devices and in particular

smartphones. Carroll and Heiser in [65] conducted a detailed analysis of the power consumption of an early model smartphone. The amount of energy consumed by the devices main components under various conditions was measured and recorded. The author's approach was to take the power measurements at the component level on real hardware. In the suspended state, in which the device was on but not being actively used, the GSM subsystem was seen to dominate power consumption accounting for approximately 45% of the overall amount of power used. This is a direct result of the need for the device to remain connected to the network in order to be able to receive SMS messages or calls. In the suspended state the Wi-Fi interface consumed about 9% of the overall power used.

In the idle state, in which the device is no longer suspended but no applications were in active use the display subsystems consumed the greatest amount of energy at 50% of overall consumption. The GSM subsystem remained a large consumer at 22% while the Wi-Fi interface consumption remained steady at 9%.

Both Wi-Fi and GPRS (provided by the GSM subsystem) were actively tested with the test consisting of downloading a file over HTTP using the wget utility a total of 10 times. Results showed that Wi-Fi provided a throughput of 660 KB/s and GPRS provided a throughput of 3.8 KB/s. Power consumption during the tests was higher for Wi-Fi at 720 mW and 620 mW for GPRS. Moreover, there was also an increase in power consumption for both the CPU and RAM during the Wi-Fi test due to the greater amount of data that needed to be handled.

Carroll and Heiser [65] also tested a wide range of components and activities such as emailing and power consumed by Wi-Fi. In the web browsing test GSM/GPRS consumed 2.5 times the amount web browsing. During the email test GSM/GPRS consumed almost 3 times the amount of power consumed by Wi-Fi for the same activity. The authors analysed the results of their tests and one of their conclusions was that the most effective power management approach on a mobile phone was to shutdown unused components.

Perrucci, et al., in [66] conducted a survey of energy consumption of smartphone components with a focus on radio communications. During testing the smartphone was controlled through the use of Python scripts. The device under test was equipped with several radio technologies including Bluetooth, Wi-Fi and cellular (UMTS). Energy consumption was measured using an energy profiler application and the results obtained were verified through

the use of a multimeter. The authors recorded and tabulated the amount of energy consumed during various activities.

In the context of downloading data, the following results were recorded:

- 2G downloading at 700 KB/s consumed 500 mW
- 3G downloading at 1 Mbps consumed 1400 mW
- Wi-Fi in infrastructure mode downloading at 4.5 Mbps consumed 1450 mW
- Wi-Fi in ad-hoc mode receiving 1375 mW
- Wi-Fi infrastructure mode idle consumed 58 mW
- Wi-Fi ad-hoc mode idle consumed 979 mW

From the results presented by the authors it is clear that the most energy efficient way in which to download data was to use Wi-Fi where available. Although it is more than 4 times faster than 3G resulting in far shorter download times it only consumes 50 mW more than 3G. In the context of Wi-Fi having the interface in idle mode while using ad-hoc networking uses far more energy that idle in infrastructure mode.

The authors also captured the energy cost per bit for downloading using various wireless technologies including HSDPA, GPRS, Wi-Fi and Bluetooth. In this series of tests Wi-Fi was seen to have the highest data download rate and the lowest energy cost per bit. Perrucci, et al., demonstrated that the wireless components of the device under test consumed the most energy and not the display or CPU. Energy can be saved by shutting down interfaces when not in use.

In more recent work Tawalbeh, et al., [67] examined the energy consumption of various smartphone components using two different devices from two different manufacturers. Energy consumption was measured on the devices under test using two different energy profiling applications and the results compared. As might be expected the amount of energy consumed by the display had shown a marked increase over the energy consumed by the screen of earlier smartphones. This increase is due to the increase in both screen size and resolution. Results of the measurements indicated that the 3G subsystem and the display were responsible for consuming the greatest amount of power. The Wi-Fi interface was also seen to consume a considerable amount of power in both devices under test. Between them the 3G subsystem and the Wi-Fi interface were responsible for approximately 50% of all power

consumed. It is interesting to note that the percentage of the power consumed by the cellular subsystems and Wi-Fi has remained reasonably steady over the years regardless of the advances in manufacturing and design.

It has been clearly demonstrated that networking operations in mobile devices can account for a significant proportion of the total power consumed by the system. Kravets and Krishnan in [68] state that a wireless network interface card can be responsible for more than 50% of the total system power consumption of a hand-held devices and for up to 10% of the total system power consumption for high-quality laptops.

In order to reduce energy consumption in mobile devices equipped with wireless networking technology Kravets and Krishnan advocate a software based approach. Their work presents "the design and implementation of an innovative transport layer protocol capable of significantly reducing the power usage of the communications device". Power savings are achieved through the strategy of selectively suspending the communications process for short durations and shutting down the communications device.

During these suspension periods the protocol is responsible for buffering data intended for transmission, at the end of the suspension period the communications device is reactivated and any buffered data is sent. The protocol also determines when to reactivate the communications device. This solution was tested by the authors and was found to deliver power savings of 6%-9% in terms of total system power for laptops and up to 40% of total system power for hand-held devices. However, a serious drawback to this strategy is that, in the case of a mobile device, there is no guarantee that a wireless link will be available when the period of suspension ends possibly resulting in incomplete data transfers.

The IEEE 802.11 standard [69] introduced its own energy conservation mechanism; called Power Save Mode (PSM). The 802.11 standard defines two modes of operation for wireless network interfaces, a wireless interface can be in active mode or sleep mode. In the sleep mode a wireless interface cannot send or receive data and only a subset of its components, e.g. the clock used to synchronise the wireless card with the Access Point, require a supply of power from the system. The active mode enables the wireless interface to send and receive data; in addition, the active mode is divided into three sub-states.

These sub-states are the transmit state, the receive state and the idle state, while in the idle state the wireless interface only monitors the channel. Regardless of the mode or sub-state

that the wireless interface may be in it continuously consumes power, with the difference between the nodes being the rate at which the power is consumed.

Work carried out by Chakraborty in [70] demonstrates that placing a wireless card in sleep mode whenever possible enables a dramatic reduction in power consumption to be achieved.

The rate at which energy is consumed in the various modes or states is well illustrated by measurement work carried out in [71] by Shih, Bahl and Sinclair. Shih et al highlight the fact that Wi-Fi adapters can consume the equivalent of 50% or more of their transmit power levels while in the idle sub-state. Since wireless interfaces can spend a great deal of their operational time in the idle state, this situation can lead to significant energy consumption during which no useful work is being carried out. The focus of the work is on reducing energy consumption by the wireless interface.

The authors propose a refinement to the strategy described in [70], instead of simple putting the wireless into the sleep sub-state they propose shutting it down completely. Once shut down, it is necessary to have the capability to reactivate the wireless interface when required. The authors achieve this by implementing a second low-power channel that remains active when the primary wireless interface is shut down.

Out-of-band control information is carried in this channel and it is used to maintain connectivity and to reactivate the main interface when required. A practical implementation of this system was constructed and tested; the results obtained from the practical implementation showed that it was indeed capable of reducing energy consumption. However, this system requires additional network resources and increases complexity of the handheld device through the introduction of a secondary radio.

Some period of time is required for the wireless interface to make the transition from sleep to active mode and this transition time is in the order of milliseconds, as demonstrated by Krashinsky and Balakrishnan in [72]. Also during this transition period energy consumption by the wireless interface increases and it is considered to be almost equal to the energy consumption of the active mode.

The 802.11 Power Save Mode (PSM) exploits the difference in power consumption that exists between the wireless interface operating modes. The strategy is to allow the mobile device to be in the active mode only for the time required for data transfer to take place. When the wireless interface switches to the idle sub-state, PSM puts it into sleep mode. The

Access Point within an 802.11 infrastructure wireless network is used to carry out this strategy. When a mobile node within a Wi-Fi hotspot wishes to use PSM it informs the Access Point to which it is associated of its decision. All traffic within a Wi-Fi hotspot is routed through the Access Point, this enables the Access Point to buffer all data destined for a mobile node using PSM while the mobile node is sleeping. At regular intervals, called a Beacon Interval and which are usually every 100 milliseconds, the Access Point broadcasts a special beacon frame that contains management information. One component of this information is a Traffic Indication Map (TIM) that indicates which mobile device, if any, has data buffered at the Access Point and every mobile device within the network is synchronised with the Access Point. The Access Point broadcasts beacon frames at regular intervals, with a typical default interval of 100 milliseconds between beacon frames. Any mobile node in sleeping mode wakes at the time the beacon frames are broadcast in order to check for any data intended for it that may be buffered at the AP. If a mobile device using PSM is specified in the TIM it sends PS-Poll packets to the Access Point to retrieve its buffered data. A drawback to using PSM is the fact that a mobile node must send a PS-Poll for every frame of data buffered at the Access Point. This obviously creates a huge amount of additional traffic within a wireless network of limited capacity with the potential to degrade network performance for other mobile nodes within the network.

The performance of the 802.11 Power Saving Mode was examined in depth by Anastasi et al through the development and use of an analytical model described in [73]. The authors validated their analytical model against the output from simulations. The motivation behind the work came from the observation that, although PSM had been analysed on previous occasions, this analysis only dealt with specific applications. The impact of other factors, such as traffic types and the number of concurrent users within a hotspot, had not been considered. Anastasi et al state that they have focused on best-effort Internet applications, (Web browsers, email and file transfer), that do not have a real-time requirement. These applications, the authors claim, are the most commonly used application in Wi-Fi hotspots. The model demonstrates that Power Save Mode performance is dependent on key parameters including traffic profile, Internet throughput and MAC-protocol parameters. The impact on PSM of other users within a cell is highlighted by the authors. The tests carried out showed that as

the number of users increased the performance of PSM degraded. This degradation of performance by PSM was graceful and the authors identified a critical point for the number of users (approximately 35 users) beyond which PSM was found to be ineffective.

Flinn and Satyanarayanan in [74] use measurement and experimentation to demonstrate that the development of a collaborative relationship between the operating system and the applications could be used to reduce energy consumption so that it would be possible for a user to specify goals for battery life duration. The paper describes how energy consumption can be reduced through having applications modify their behaviour dynamically. It is claimed that an acceptable trade-off between energy consumption and application could be achieved. In order to achieve this state of affairs it is necessary to constantly monitor both energy reserves and the rate at which they are being depleted. The system also requires three tasks to be carried out at regular intervals, (1) Determine the remaining energy reserves, (2) Predict future energy demands, (3) Make a decision, (based on steps 1 and 2), on whether or not an application should change fidelity. The authors define changing fidelity to mean that a change is made to the way in which an application presents data. Reducing fidelity can take various forms such as reducing the window size, changing to monochrome colour values and reducing audio quality e.g. using mono instead of stereo. This early work introduces the concept of applications and system components that support wireless operations also being significant consumers of energy. It showed that changes to the application behaviour could have a positive impact on power consumption. However, this strategy would be unacceptable to the majority of present day mobile device users. To these users of mobile devices, the multimedia capabilities of their devices are at least as important as the devices networking capabilities. They would not countenance the drastic reduction in media presentation quality this strategy would impose. If these reductions in quality were a result of decreasing power reserves, then the quality would not improve until the device battery was recharged. In fact, quality would be continuously degraded in line with reducing energy levels. It must be remembered however, that at the time this strategy was formulated user expectations were not as high as they are in the present day.

The IEEE 802.11 standard defines two basic wireless network structures; these are Base Station mode and Ad hoc mode. In Base Station mode every mobile node within the network must be in range of a Base Station or Access Point, all traffic within the network is routed through the Access Point and then on to the destination. As we have seen in [4] the role of

the Access Point can be exploited to gain reductions in energy consumption by the mobile nodes using PSM. In ad hoc wireless networks, on the other hand, no Access Point or Base Station is used. Instead, an informal peer-to-peer network structure evolves.

Feeney and Nilsson [75] investigate the energy consumption of a wireless interface operating in ad hoc mode. Mobile nodes operating in ad hoc mode communicate directly with other nodes sending point-to-point traffic; mobile nodes can join or leave the network at will. For communications to occur between two mobile nodes in an ad hoc network they must be in range of each other. Nodes operating in this environment must be ready to receive data at all times from neighbouring devices since no Access Point is available to buffer data. This means that wireless interfaces operating in ad hoc node cannot enter sleep mode to conserve energy and as a result they consume energy constantly. Even discarded packets have an energy cost associated with them in an ad hoc network.

A mobile node within an ad hoc environment expands energy sending and receiving data, such a device may also receive traffic from many other nodes in the environment. It is only by examining each packet received that a node can determine which packets are for it and which should be discarded. Since it is the case that most of the traffic received in this situation will not be for the receiving node, a great deal of unnecessary energy may be expanded.

Because the ad-hoc networking mode can result in a high rate of energy consumption without any direct benefit to the mobile node it should be avoided where possible in order to conserve energy. In Wi-Fi only wireless environments a node that loses connectivity to an Access Point may attempt to establish a connection using ad hoc networking. However, in a heterogeneous, multi-network wireless environment a mobile node equipped with appropriate wireless interfaces can avoid using ad hoc mode and establish a link using an alternative technology.

Due to the constraint placed on mobile devices by battery capacity, careful consideration should be given as to how energy is consumed by a device. Wireless adapters consume energy during operation [76] and continue to do so even while in standby mode; the amount of energy consumed obviously rises with the number of active wireless interfaces. It has also been demonstrated that transferring data over a mobile phone network consumes more energy than transferring the same amount of data over Wi-Fi [77].

A different approach is taken by the authors of [78] present a data driven strategy for tackling the problem of Wi-Fi being least power efficient in the idle state and causing highest energy overhead when scanning for networks. Bandara and Caldera focus on Wi-Fi usage from the user's perspective and model the Wi-Fi usage of mobile users based on their past usage to predict usage requirements. This approach enables intelligently switching on the Wi-Fi interface only when the user context requires it, thereby reducing long periods of time in the idle state and significantly lessens the number of unnecessary network scans. The authors built their prediction model on trace data collected from the Rice-Livelab study, extracting temporal, application usage, operational state and location context data in order to do so. Their study includes a systematic feature engineering process followed by the deployment of machine learning algorithms on the target dataset. Sampling, Ensemble and Hybrid techniques were used to mitigate the class imbalance problem of the prediction model. Evaluated metrics indicated that the decision tree based classification algorithms perform well with the available dataset and were suitable for working with mobile usage data, which are mostly conflated with noise and data imbalance.

Deogun et al. in their paper [79], proposed a modified scanning algorithm to reduce the number of unnecessary scans. In their proposed scheme, the UE schedules its scanning operations with assistance from the cellular network on the operator deployed WLAN network. They conducted simulations using NS-3 [143], with additional modules implemented for ANDSF and IEEE 802.11u for 3GPP-WLAN interworking. The results showed significant improvements in energy consumption along with better association time for the users.

However, this approach relies on service provider controlled Wi-Fi networks when in reality Wi-Fi deployments are largely isolated, independent networks. Even if service provider Wi-Fi networks are available this strategy fails as soon as the mobile user travels beyond the coverage areas of such networks.

In the context of Wi-Fi based localization which has proven to be a compelling alternative to GPS for mobile devices the authors of [80] propose a novel scanning strategy. Instead of the energy inefficient full, all channel scans that are the norm the authors propose a novel, incremental approach that reduces the energy consumption of Wi-Fi localization by scanning just a few selected channels. This incremental scanning approach was evaluated on eight Android devices using traces from five test subjects. Results showed that, compared to full scans, incremental scanning can reduce the energy consumption between 20.64% and

57.79%. The modern smartphones included in the study all show an energy reduction of at least 40%. Only scanning a subset of the channels used by Wi-Fi presents the problem of how the user knows which channels will be in use to avoid missing active channels.

Lim and Rhee in [81] propose an intelligent power management scheme which uses a reinforcement learning algorithm to optimize the sleep/awake period of a mobile station in various environments. By dynamically adjusting the Listen Intervals of mobile stations, their energy consumption can be optimized, while the trade-off between the energy consumption and the transmission delay is efficiently managed. The authors tested an implementation of their power management scheme in a simulated environment using NS-3 [132]. The simulation results demonstrated that the proposed scheme could improve both power consumption and delay performance.

Rattagan in [82] presents a monitoring method and power management policy for Wi-Fi when several background apps are utilizing the Wi-Fi interface for data transfers. Instead of tackling the energy consumption of Wi-Fi directly this approach modifies the behaviour of applications using the Wi-Fi interface. The proposed method can non-intrusively monitor the Wi-Fi usage of these background apps at the application framework level (API) in runtime without modification of the smartphone operating system. With the API-level monitoring data, the author applied it to efficiently manage the Wi-Fi power consumed by specific applications. Experimental evaluation showed that by applying the proposed method to only one background app, the application which made most use of the Wi-Fi network, resulted in an 8% improvement of the total energy consumption, compared with the default Wi-Fi power management policy

All wireless networking operations consume energy and once the radio hardware has been activated, the actual amount of data transferred has little further impact on energy consumption. Therefore, to make the best possible use of a wireless link it is important to send as much data as possible over the link since the energy cost associated with that network operation has already been incurred. For energy conservation, it is better to transmit data in large bursts and then not transmit anything allowing the interface to sleep, rather than transmitting little over longer periods of time and reducing the sleep periods.

3.6 Data Offloading

The continued rise in the amount of video content being delivered over cellular networks has placed severe pressure on network resources. This is not a new phenomenon, as the capacity of any network grows the amount of traffic traversing the networks rises to match and then exceed capacity. Cellular networks are being touted as a means of enabling M2M communications between IoT devices, this coupled with the projected growth in the number of IoT devices that will want to communicate and transfer data this situation will only worsen into the future.

Service providers are caught between a rock and a hard place, they need to increase revenue and develop new revenue streams by growing their customer base but in doing so they risk degrading the network performance enjoyed by early adapters. Dissatisfaction with the service provided results in customer churn, service providers can address the reducing network performance by increasing their capital expenditure (CAPEX) through deploying new systems and upgrading existing ones. This has the knock-on effect of increasing operating expenditure (OPEX) due to increased staffing and maintenance requirements. Even for those service providers willing to increase spending in order to improve their networks the process is not straight forward due to the mature nature of the market in developed countries. Additional spectrum is difficult and expensive to acquire, suitable sites for new cell towers are scarce and it is difficult to arrive at co-locating agreements with rival operators. Service providers are taking measures to reduce the load on their cellular networks by retreating from truly unlimited data plans, introducing and enforcing data caps and by offloading data from the cellular networks to alternative data transport systems such as Wi-Fi. Research into data offloading has been ongoing for many years.

General strategies for data offloading from cellular networks include:

- Indoor offloading to Femto cells and APs
- Outdoor offloading to Wi-Fi
- Ad-hoc peer to peer offloading

From a temporal aspect offloading can be divided into delayed offloading and non-delayed offloading. Essentially the difference between delayed offloading and non-delayed offload-

ing is the amount of latency in delivering the data that can be tolerated by the various applications. With non-delayed offloading there is no additional delay incurred by sending traffic over the secondary (assumes that we consider the cellular interface to be the primary). Delayed offloading introduces additional delays due to the time required to move within range of an appropriate AP or transfer data through an opportunistic D2D contact. All data offloading strategies depend completely on the mobility of the user; mobility enables an end user to move into range of an AP or other device to which it can offload data. However, for non-delayed offloading mobility can create problems by moving a user out of range of an AP.

Dimatteao, et al., in [83] propose an integrated architecture named the Metropolitan Advanced Delivery Network (MADNet). They employ a Delay Tolerant Network (DTN) approach that takes advantage of the fact that much of the data traffic sent over mobile networks is delay tolerant in nature. The proposed architecture is based on the idea of using the cellular network for signalling and a combination of cellular and alternative technologies such as Wi-Fi to carry data. Location services such as GPS, cell tower triangulation, etc. are employed to aid users in selecting a location for delivery. The authors tested MADNet using simulations and to promote realistic evaluation they modelled the behaviour of the simulated mobile nodes on a real data set of 500 taxi cabs in San Francisco that had been captured over a period of 30 days. They demonstrated that the deployment of several hundred Wi-Fi APs over a metropolitan area of approximately 314 square kilometres would enable 50% of delay tolerant data traffic to be offloaded from the cellular networks.

Currently many AP deployments are un-coordinated and haphazard which results from many being setup by individuals and businesses without any planning or central control. In fact, the ability to cheaply, easily and quickly setup a wireless network using Wi-Fi has been the driving force behind the technology's widespread adoption and popularity. However, if a service provider wants to be able to reliably offload data from their cellular network to an alternative network there needs to be some rational method applied to AP deployment.

The problem of how to effectively deploy APs has been the focus of much research. Ristanovic, et al., in [84] and Bulut and Szymanski in [85] developed Wi-Fi AP deployment algorithms that are intended to facilitate offloading as much traffic from cellular networks as possible. Their approaches are similar and they propose heuristic solutions to the problem of finding the optimal AP deployment. The aim of the deployment algorithms is to place the APs in close proximity to the locations having the greatest number of users or mobile data

requests. Simulations were used to test the algorithms and the results showed that depending on the AP density reductions in the overall amount of data traffic on the cellular networks of between 20% and 70% were achievable.

In order to gain an understanding of how much data might be offloaded from a 3G cellular network to Wi-Fi the authors of [86] conducted a quantitative survey. They recruited 97 iPhone users from metropolitan areas of South Korea and had them install a custom application named Delay Tolerant Application (DTAP). DTAP recorded the Wi-Fi connectivity statistics in the background for two and a half weeks periodically uploading the collected data to a server.

Lee, et al. [86], examined data offloading from the point of view of both on-the-spot offloading and delayed offloading. On-the-spot offloading uses Wi-Fi connectivity to transfer the file immediately and for as long as the connection lasts or until the file transfer completes. With delayed offloading each data transfer is associated with a deadline, if the data transfer is not completed to a Wi-Fi AP it is resumed whenever Wi-Fi connectivity is regained. In the event that the file transfer does not complete before the deadline expires the cellular network is used to finish the transfer. The authors note that both on-the-spot and delayed offloading reduce the load on cellular networks.

From the data uploaded by the DTAP application the author's found that mobile users were within coverage of a Wi-Fi AP approximately 70% of the time. They typically remained in Wi-Fi coverage areas for 2 hours at a time and after leaving the coverage area they returned to it after about 40 minutes.

Building on the data received from the DTAP application the authors designed and ran trace-driven simulations to measure the efficiency of both on-the-spot and delayed offloading. Some of the key findings from the simulations were that on-the-spot offloading could offload about 65% of the total cellular traffic load. On-the-spot offloading also delivered an energy saving of approximately 55% for data transfers. Because Wi-Fi connections had higher data transfer rates than the 3G data connections transmission times were greatly reduced with the reduction in transmission times translating directly into energy savings.

In the case of delayed offloading the authors found that improvements over on-the-spot offloading only became substantial when long delays were introduced. Short delay deadlines

of 100 seconds showed little impact, in order for delayed transfers to achieve substantial gains the deadlines needed to last for several tens of seconds.

Balasubramanian, et al., in [87] investigate whether or not access to Wi-Fi can be used to augment 3G cellular networks. A detailed study was carried out into 3G and Wi-Fi accessibility for user traveling in vehicles. The study was conducted in three different cities, the authors found that across all three cities 3G availability was 87% and Wi-Fi availability was 11%. Throughput achieved over Wi-Fi was lower than the throughput achieved over 3G and the packet loss in Wi-Fi was higher. A system named Wiffler was developed that uses fast switching and delay tolerance to overcome poor Wi-Fi availability and performance. In the case of delay tolerant applications Wiffler uses a simple model of the environment to predict Wi-Fi availability. Data transfers are delayed based on the prediction of future Wi-Fi connectivity but only if the delay will reduce the load on the 3G cellular network and the transfer can be completed within a set period of time known as the applications tolerance threshold. In the case of applications such as VoIP that are sensitive to delay Wiffler fast switches to 3G if there is any delay on Wi-Fi. The authors implemented the Wiffler application and tested it on a vehicular testbed. Results from the tests showed that Wiffler could provide a significant reduction in data transfers over 3G. A reduction of 45% in traffic was achieved with a tolerance threshold of 60 seconds.

Trestian, et al. in [87] propose a novel cellular network architecture using targeted infrastructure upgrades. Selected locations known as Drop Zones are identified by an algorithm developed by the authors. The Drop Zones are the targets for the infrastructure upgrades and fall within the movement patterns of large numbers of users. The idea is that the mobile users delay data transfers until they enter the Drop Zones at which time they send/receive bulk data transfers. The authors highlight the fact that many users already exhibit this pattern of behaviour and suggest that the service providers further encourage delayed transfers through pricing incentives. The Drop Zone algorithm aims to reduce the number of APs required and the average data transfer delay. There is of course the danger that frequently the situation will arise in which multiple users arrive in a Drop Zone simultaneously and begin bulk data transfers overwhelming the system.

Not all data offloading strategies employ APs, Device to Device (D2D) or Terminal to Terminal (T2T) offload data transfers to other UEs. One use case is for a single device to download content from a server and to then redistribute it to other devices within the group. D2D

communications can take place out of band using alternative technologies such as Bluetooth or Wi-Fi or in band using the same frequencies as those used for cellular transmissions. Regardless of whether the communications are in band or out of band the behaviour of the devices within the group are conceptually similar to that of an ad-hoc Wi-Fi network in that communications take place between nodes without the use of an AP or base-station.

In comparison to standard data distribution in cellular networks D2D real-time offloading offers some advantages in terms of coverage, energy consumption and average throughput but at the cost on introducing complexity. Link quality can change rapidly if participating nodes change position making it difficult to guarantee QoS. If data transfers can be deferred, then delayed offloading is a far better choice for data distribution.

Stiemerling and Kiesel in [88] focus on D2D media streaming in the context of high UE mobility. The basic idea proposed by the authors is that each video segment is downloaded once over the cellular connection and is the redistributed to the other group members through short range D2D communications. There has to be coordination between participating nodes to decide which node will provide the point of attachment to the cellular network. Having only a single point of attachment reduces the load on the cellular network itself. One node acts as a central controller who coordinates content retrieval between participating nodes. The authors conducted simulations to determine the minimum number of cooperating nodes required to meet throughput and reception delay targets.

Cooperative data offloading faces a number of challenges including how to discover neighbouring nodes, how to coordinate actions between nodes, ensuring service continuity, etc. Andreev, et al., in [89] propose a network driven approach to these issues in which an intelligent network architecture assists connected users in both the content discovery and connection establishment phases. The authors tested their proposal through simulations and demonstrated a 2.5 increase in throughput as well as offloading approximately 30% of data traffic.

3.7 Patterns of Movement in Urban Environments

In order to be able to build simulation scenarios that reflect the real world with a reasonable level of accuracy we must first understand user behaviours, specifically in the area of user

mobility. A mobile entity is one that can move from one location to another and in urban environment mobile entities are either pedestrians or passengers in a vehicle of some type. Regardless of whether a mobile entity is a pedestrian or a passenger in a vehicle they share some major common characteristics:

- The entity can be either stationary or in motion
- If an entity is in motion, then it must be moving in some direction that is constrained by the physical environment
- The velocity at which the mobile node moves reflects human behaviour, crowding and environmental conditions in the case of pedestrians. In the case of passengers their speed is constrained both by local speed limits and traffic congestion levels

In order to understand the behaviour of mobile entities we examine the characteristics of these entities under the following headings:

- User velocity – both pedestrian and vehicular speeds in an urban environment
- User behaviour while in motion – path selection, movement patterns
- User behaviour in a deployed wireless network

3.7.1 User velocity

The speed at which vehicular traffic can travel through urban or city streets is determined by local speed limits and to a large extent, by the level of congestion. High congestion levels reduce traffic flow and therefore speed, while low levels of congestion allow an increased traffic flow resulting in higher speeds.

In general, all vehicular traffic within an urban traffic system, with the obvious exception of emergency and police vehicles, moves at approximately the same speed. This is due to the reduced space available for overtaking slower vehicles or carrying out other manoeuvres intended to increase speed. Speed restrictions may also be imposed by the prevailing physical conditions of narrow streets and traffic management measures that may be in place. In many urban areas part of the traffic management plans has been the introduction of bus lanes and corridors which are laneways dedicate to the use of mass transit vehicles such as buses [90]. The idea is that by dedicating part of the roadway to buses only the average speed

achieved by public transport will increase and more people will be encouraged to use it. A reduction in available parking spaces is also used to force commuters and others out of their cars in a bid to reduce congestion and pollution. Bus lane may also be used by taxis, cyclists and emergency vehicles including the police.

Much of the work in the area of vehicular traffic speeds has been carried out with reference to public transportation and in particular to the speeds achieved by public buses. This work spans decades but it remains relevant to present day conditions since congestion levels in many urban centres has been close to maximum for many years. Vucan and Vuchic [91] state that typical bus operating speeds in American cities in the 1980's as being 15 to 20 km/h during off-peak hours. During peak hours, when congestion levels were higher, reduced speeds of only 8 – 14 km/h were typical. According to figures from 2005 the average bus speed in comparable European cities, provided by the UK Commission for Integrated Transport in [92] and presented in Table 1 range between 13 km/h and 20 km/h with a mean operating speed of 16.5 km/h (4.58 m/s) for the major European cities considered.

Table 1 Comparison of Average Operation Speeds

City	Average operating speed (km/hr)
Rome	20
Berlin	19
Madrid	19
Athens	18
London	18
Dublin	13

It is difficult to see any improvements in traffic speeds in urban centres in the future, this view is supported by the case of the city of Dublin (Republic of Ireland), presented in a 2007 report [93], commissioned by the Dublin Bus Authority. The report is a comprehensive review of bus operations, congestion, speeds achieved and forecasted changes in Dublin and shows increased congestion leading to slower average bus speeds and that congestion is predicted to worsen in future years. In 2000 the average speed of the Dublin Bus fleet was 14.6

km/h, in 2003 it was 13.5 km/h and in 2005 it was 12.9 km/h, which was 36% slower than comparable cities. In his Opening Statement to the Joint Oireachtas Committee on Transport, Tourism and Sport, on the 5th of October 2016 Mr Ray Coyne (CEO Bus Átha Cliath – Dublin Bus) stated that "Our network speed at peak times is the region of 14kmph, with substantial variances on all transport corridors" [94]. It is worth noting that there has been little or no improvement in the average operating speeds for public buses over a 26-year period.

The study of pedestrian walking speeds in urban or city environments has been ongoing for several decades. In his important work in the early 1970's Fruin [95] describes and quantifies the space people require to walk, queue, crowd and wait. He also discusses the space people need to access transportation systems, elevators and escalators as well as the number of people that can walk on stairs and through corridors. In order to develop this information Fruin carried out a series of studies on the behaviour of pedestrians within transportation terminals. He carried out two important studies in New York City in 1971, at the Port Authority Bus Terminal and Pennsylvania Train Station, allowing him to observe pedestrian walking speeds under free-flowing conditions. He observed that the mean pedestrian walking speed to be approximately 1.35 m/s or 4.86 km/h.

In 1998 Young conducted a study of the walking speeds of passengers in various airport terminals [96]. The effect of automated pedestrian movement systems on pedestrian walking speeds was the focus of the study. Empirical observations, similar to those carried out by Fruin, were taken of pedestrian movements. Analysis of these observations revealed that, under free-flow conditions airport pedestrians behave in a manner similar to those in Fruin's studies. Young's work at two different sites provided information on free-flow walking speeds, the observed mean pedestrian speed was approximately 1.34 m/s or 4.82 km/h. The previous works studied pedestrian walking speeds in transportation terminals, which although useful, does not provide an indicator of how pedestrians might behave in an urban area or city centre. It should be remembered that pedestrians in transport terminals might be encumbered with luggage and may be moving in at a different rate than normal in order to meet boarding times, etc.

It is necessary to examine the behaviour of pedestrians in a street setting in order to get a sense of how a pedestrian should behave in a simulated environment. Much of the work in the area of pedestrian walking speeds deals with the amount of time required by a pedestrian

to cross a roadway. These studies are undertaken since it is necessary to allow the pedestrian enough time to safely cross but it is important not to delay traffic more than is necessary.

The average pedestrian walking speed is used to determine the timer settings for crossing lights. Knoblauch, Pietrucha, and Nitzburg [97] compared the differences in walking speeds of older and younger pedestrians, both male and female. Unsurprisingly, it found that older pedestrians tended to walk at a slower rate than younger pedestrians. The team observed over 7000 pedestrians, of which 3,365 were 65 years or older. These observations showed among many other things that the average pedestrian walking speed for older pedestrians was approximately 1.22 m/s (4.39 km/h), and 1.51 m/s (5.53 km/h) for younger pedestrians. There were also observed differences in walking speed between males and females. This is a somewhat artificial walking environment since the pedestrian starts from a dead stop and moves a relatively short distance under a strict time constraint. In addition, the fear of being trapped in approaching traffic must have a motivating effect on the pedestrians. However, even in this situation the pedestrian walking speed is close to that observed by Fruin and Young in their studies. A similar rate of speed is presented in the Manual on Uniform Traffic Control Devices [98] which gives a guide pedestrian walking speed of 1.2 m/s (4.32 km/h). This guide is employed nationally in the United States in order to standardise the behaviour of traffic control devices in North America.

The authors of [99] provide a comprehensive study of the requirements and behaviour of pedestrians in an unconstrained urban walking environment. It provides a very clear insight into pedestrian behaviour in an urban setting. They found a very large spread in pedestrian walking speeds but the work shows an average pedestrian walking speed of approximately 1.3 m/s (4.68 km/h converted from feet/s). This average value is very close to the walking speeds observed in other situations.

For their 2013 study [100] Chandra and Bharti captured data related to pedestrian speeds using video cameras at 7 locations in 3 different cities in India. Walking and crossing speeds were analysed on the basis of gender and location type. The location types as described by the authors were restricted to “sidewalks, wide sidewalks, precincts and carriageways”. Analysis of the collected data demonstrated that overall Indian men walked at a slightly slower pace that reported by other researchers while female pedestrian speeds closer to previously observed results. Average walking speeds across both genders were 1.25 m/s (side-

walk), 1.36 m/s (wide sidewalk), 0.97 m/s (precincts) and 1.23 m/s (carriageways). The pedestrian speeds observed in the Indian study are broadly in line with the results seen in many previous studies. It is therefore reasonable to say that, in an urban environment the average walking speed of a pedestrian is 1.3 m/s.

3.7.2 Mobile entity behaviour

In order to understand how mobile users, and pedestrians in particular, behave we must also understand how they navigate through their surroundings. Within an urban environment it is not possible to move in totally random patterns, physical objects such as buildings, fences, hoardings or conceptual and cultural barriers such as road markings force the pedestrian into constrained routes. It is always possible to ignore the logical barriers and enter a roadway but in this case pedestrians generally cross in the most direct route i.e. a straight line.

Vehicular traffic is far more restricted in an urban area than pedestrians and the directions the traffic can move is tightly controlled. But not all urban spaces are as constrained as the streets; many urban areas and city centres have open areas such as squares or parks where pedestrians have greater freedom of movement, yet even here pedestrians do not move in totally random ways.

The idea that pedestrians will travel in a straight line if possible is raised by the work of Golledge [101] who carried out a series of experiments in 1995 using both laboratory environments and tests in a physical environment. In the physical tests 32 subjects were placed in an environment with which they were familiar during daylight hours, they were then given a destination to which they had to travel. The routes selected by the subjects were recorded by researchers and then plotted on maps of the area. These plotted paths revealed that the majority of the routes followed were as straight as was practical in the environment.

Later work by Ruth Conroy Dalton in [102] presents the results of an experiment in which route-choice decisions made by subjects at road junctions were recorded. The experiment was carried out in a simulated urban environment, the subjects navigated through the simulated environment using a virtual reality headset. The group of subjects consisted of 30 members (68% male, 32% female), with an average age of 28. An urban area of 650 x 650 metres was modelled, with this area containing objects that represented urban “block” structures in the shape of squares and triangles. Virtual streets ran between the shapes and intersected at junctions where the test subjects were required to make a decision.

All test subjects started from the same position within the simulated environment and were tasked with navigating to the opposite corner by the most direct route. The subjects were immersed in the virtual environment for a maximum of ten minutes and moved through it at a walking speed that approximated real world pedestrian walking speeds. During the experiment the position of each subject was recorded ten times per second and the routes they followed were plotted onto a plan of the simulated environment. The author observed that the subjects appeared to be selecting the straightest possible routes through the environment. This finding supports the hypotheses made by Hillier [103] in which he states that people tend to follow the longest line of sight that approximates their desired heading. They also reflect the results of the earlier field trials carried out by Golledge [101].

Helbing, et al. in [104] show that although each pedestrian will have their own destination, aims or personal preferences, the dynamics of pedestrian crowds is surprisingly predictable. It is only at low crowd densities that pedestrians can move freely and as crowd density levels increase their behaviour is affected.

Pedestrian behaviour is affected by repulsive interactions with other pedestrians which results in self-organising phenomena. Self-organisation of pedestrian crowds is seen in the development of separate lines of uniform walking direction in crowds of pedestrians moving in opposite directions. It is also observed in the oscillation of passing direction that occurs at bottlenecks such as doorways.

3.7.3 Mobile entity behaviour in wireless environment

Studies carried out into the behaviour of mobile users within deployed wireless networks provide a picture of the way in which users exploit the networks. They also show the movement patterns of mobile users in a variety of settings. Tang and Baker in [105] analysed a seven-week trace of mobile user activity within a metropolitan-area wireless network in order to investigate how such users take advantage of the wireless environment. The wireless network in question was deployed in the San Francisco Bay area and of a substantial size.

Among their findings they observed that users typically used the network during the day and evening and that the radios were most active during non-work hours. The authors also found that over 50% of the network users moved location and that the more locations a mobile user visited on a daily basis the closer together on average the locations were. From

these findings we see that almost half of the network users did not move location and of those that do move location do not move very far.

Working in a campus environment Henderson, Kotz and Abyzov analyse a network trace for a mature wireless network for their work in [106]. They compare this trace with a trace that was taken soon after the network was deployed, a gap of approximately two years. This provides the authors with a unique opportunity to study the changes that have taken place in user behaviour.

A change in the type of applications being employed by users of the network was observed, as was a diversification of device types being used. The increased heterogeneity of devices manifested itself in a greater number of Mobile devices and VoIP enabled devices, reflecting an increased availability of such devices. The users of Mobile devices and VoIP enabled devices were found to have different mobility characteristics to laptop users. Also of interest was an observed increase in the amount of traffic per device, although the number of devices on the network had increased the amount of traffic per device had doubled in the two years between traces.

Henderson, Kotz and Abyzov [106] were interested in user mobility characteristics such as how often and how far a user moved during a session. It was not possible for the authors to directly measure physical mobility; they had to infer mobility from user roaming patterns. Roaming by a wireless interface between Access Points does not imply physical mobility, cards were seen to "ping pong", associating and re-associating with several Access Points (APs) many times in succession.

Kotz and Essien in [107] define a "mobile session" as one where a card visits APs in more than one building; however, it was found that a stationary card may ping-pong between APs located in different buildings. The authors found that users spent almost all of their time in their home location. The home location definition was based on a modified version of the home location defined in [94]. In this case the home location is defined as the AP at which the user spent more than 50% of their time, with the modification that all additional APs within a 50 metre session diameter are included.

Users of the network were found to be relatively immobile, 95.1% of all users had a home location and of this number 50% spent almost all their time (98.7% of connection time) connected to the network at their home location. These figures may be skewed due to the

fact that halls of residence were included but if overnight periods (12 am to 6 am) are removed, the figures show that 50% of users spend 69.2% of their time associated with a single AP. The lack of mobility may be due to the use of laptops on the network; however, it is unlikely that laptops will be used by mobile users in urban areas due to their relatively large size. It is probable that the devices used by mobile users will be handheld devices.

In [108] McNett and Voelker analysed the mobility patterns of mobile device users in a campus-wide wireless network. A trace of wireless network activity that covered an eleven-week period was used for the analysis, the trace contained data from approximately 275 users. The authors observed that mobile device users were almost twice as mobile as laptop users. In this study the authors define mobility as the number of different APs with which a device is associated. This makes it difficult to determine exactly how physically mobile users were since Henderson et al, in [106], shows clearly that roaming between APs does not imply mobility.

McNett and Voelker do not give any indication of the distances travelled by network users or of the distances separating APs. From the paper [108] it seems that network users formed clusters within buildings such as halls of residence and libraries when associated with an AP. The authors found that the average pedestrian walking speed of network users was 1 m/s, which is 50% less that the average for younger pedestrians [97] and is very close to the observed average pedestrian walking speed of older pedestrians.

Schwab and Bunt in [109] analysed the trace of network usage for an 802.11b wireless network, the wireless network was deployed on a college campus and consisted of 476 APs spread over 161 buildings and structures. Although coverage of the grounds outside the buildings was not an aim of the network, the compact nature of the campus meant that a usable signal could be received in many outside locations.

The network activity was captured in the trace for a period of eleven weeks. A majority of the wireless enabled devices on the network were laptops and a total of 1706 unique MAC addresses were recorded. Analysis of the trace showed that few of the devices that connected to the network were very mobile, that is, the locations at which the mobile device connected to the network did not change very often. The average number of locations at which mobile devices connected to the network was five buildings and nine APs visited during the trace period. No single device was found to visit even half the network and 18% of all devices

only associated with an AP in a single building for the duration of the trace. It was clear from the analysis that most users limited their network activity to a few sites key to their daily routine. The authors do not mention that roaming in a wireless network is not an indication of physical mobility. While the average of nine separate APs to which users associated may be inaccurate the physical separation of buildings means that their assumptions on user mobility based on the buildings they visited must be accurate.

A similar study was conducted in a business setting by Balazinska and Castro [110]. It was carried out to gain an understanding of the way in which the network was being exploited by the users of a corporate, wireless local area network. They observed that distinct differences existed in mobility levels among different users and that some of the differences were quite large. However, most network users were found to spend a large fraction of their time at one location.

Over the years the capabilities of the devices carried by users have changed almost beyond recognition having powerful processors, various sensors and a multitude of applications. Although the characteristics of the devices may have changed the behaviour of the users with regard to mobility has changed little. Tossel, et al., in [111] studied the mobile web usage of 24 students who were on average 19 years of age for a period of 1 year. Each student was supplied with an iPhone and unlimited data plans along with unlimited text messaging and 450 phone minutes per month. The authors gathered data on various aspects of the subject's behaviour including their locations, browsing habits, calls, etc. Each location was logged and was based on unique Cell IDs. It was found that the 24 participants frequently returned to the same locations to use their phones, the granularity of the location information was coarse as the users could have been using their phones anywhere within the identified cell. Never the less, there were a large number of visits to the user's top three to five locations. This matches the behaviour observed by Henderson, Kotz and Abyzov [106] in a similar setting i.e. college campus using a different technology. This would indicate that human behaviours are consistent regardless of the technology available.

In the world outside the college campus Yang, et al., [112] collected one week's worth of continuous HTTP traffic from a cellular network between December 28th 2013 and January 3rd 2014 within a large metropolitan area in mainland China. The authors studied user behaviour from three aspects 1) data usage, 2) mobility patterns and 3) application usage. They

used cell towers to define the user's location which they extracted from the data trace. Although the granularity of the location information was quite coarse it was adequate to map user's daily movement patterns within the metropolitan area. Typically, within urban areas cell sizes are small to facilitate efficient frequency reuse and to boost the overall capacity of the system, small cell sizes would also improve the accuracy of the location data. On the 28th of December 2013 approximately 35% of the users tracked visited only one cell (these users can be considered to be stationary) and overall 90% of users visited less than 10 cells that day. The author's point out that non-mobility does not mean that users do not move, it just means that they might not use the data service or that they go online from one location (cell) for example their place of work or study. Over the course of the week that was studied it was seen that 50% of the users visited less than 10 distinct cells and that 90% of users visit less than 10 cells in a day. The authors conclude that the range across which the users moved was very limited. No information is given as to whether the users commute to work or study or if they live a short distance away. This would obviously have an impact on the number of distinct cells visited.

From the studies of pedestrian behaviour and rates of movement, urban transportation trends and the behaviour of mobile users in deployed wireless and cellular networks the following conclusions can be drawn:

- Pedestrians exhibit an average walking speed of 1.3 metres per second in a variety of environments
- Passengers in vehicular traffic in urban areas move at an average rate of 4.6 metres per second
- Vehicular movements within a built environment cannot be random as they must follow streets
- Pedestrian movement patterns are not random; they move in straight lines as much as possible
- Crowds of pedestrians develop self-organising behaviour if crowd density is high enough and that the behaviour of large groups of pedestrians is not random

Studies of user behaviour in wireless networks show that, in fact, most users are not very mobile at all. They move location and then remain motionless for various periods of time. Even when they do move location they do not move very far, favoured locations tend to be close together

3.8 Quality of Experience

Quality of Experience (QoE) is a subjective metric and measuring it typically involves using panels of human assessors making it both a time consuming and expensive activity. In addition, the need to use human assessors makes it impossible to measure QoE in real-time for video content streamed over a network. A great deal of research has been focused on the possibility of developing an objective way in which to measure QoE. Quality of Experience is difficult to measure using objective metrics since parameters used to define QoE can differ from service to service

There is more to QoE than the quality of the video and audio components. According to the authors of [113] the parameters that affect QoE can be broadly divided into three groups;

- The quality of the video and audio content at source
- Quality of Service (QoS) in the context of content delivery over the network
- End user perception including user expectations, context in which video is viewed, users emotional state, etc.

Content quality is related to the type of codec used that is whether the codec is lossless or lossy. QoS parameters that have the greatest impact on the quality of streamed content are packet loss, jitter, delay and the amount of available bandwidth. The first two categories are reasonably easy to quantify but the third category is not. As previously stated a panel of human assessors is typically used to subjectively capture QoE using the Mean Opinion Score (MOS) system [114]. The Mean Opinion Score is well understood in the context of telecommunications having first been used to subjectively determine the quality of voice calls and later being applied to the evaluation of video content. MOS scores are expressed as values between 1 and 5 with the minimum threshold for acceptable quality corresponding to a MOS of 3.5. The MOS scale is a five-point scale where 1 = bad, 2 = poor, 3 = fair, 4 = good and 5 = excellent.

Various QoE measurement methodologies exist including;

- No reference model - in this methodology there is no knowledge of the original media stream or source files. It attempts to predict QoE by monitoring QoS parameters in real-time

- Reduced reference model – this model has a limited knowledge of the original stream and attempts to combine this limited knowledge with some real-time measurements to predict QoE
- Full reference model – this methodology assumes full access to the reference video, may also combine this knowledge with real-time measurements
- For the best accuracy the full reference model should be used but this approach requires control of both ends of the system. The no reference model is easier to adapt but may not always produce the most accurate results.

3.8.1 QoE of Video

For end users viewing video the most important factors relating to the quality of the video include viewing distance, brightness, resolution, contrast, sharpness, etc. It should be noted that fidelity and quality are not the same thing, fidelity relates to how closely the processed video matches the original source video. In the case of a low quality original video a high fidelity reproduction will still be low quality. The perceived quality will be affected by visual distortions such as jerky motion, aliasing, blockiness, blurring, staircase like slanted lines, etc. Another important perceptual factor for video containing audio is synchronisation between the video content and the audio content. Errors also occur during transmission of the multimedia content over the network. Packet loss, delay and jitter are the three most important influences on video quality. The visual impact of packet loss depends on the codec used and the type of information lost.

Various methods for measuring video quality have been developed. Peak-Signal-to-Noise-Ratio (PSNR) give the ratio in decibels (dB) between the signal power of the original signal versus the power of a reconstructed compressed signal. PSNR is a commonly used metric for video quality. However, Huynh-Thu and Ghanbari in [115] demonstrated that it does not accurately reflect QoE, despite this it continues to be a popular method for evaluating quality differences among videos.

Pinson and Wolf in [116] introduce the Video Quality Metric (VQM), a software tool developed by the Institute for Telecommunication Science (ITS) for objectively measuring perceived video quality. It is used to measure the perceptual impact of video impairments such as block distortion, colour distortion, blurring, etc., and it combines them into a single metric.

Early work by Khirman and Henriksen [117] examined the relationship between Quality of Service of a network and the Quality of Experience as seen by the end-user. When the authors examined user satisfaction with HTTP services with a focus on web browsing they found that the user's QoE was strongly influenced by the available bandwidth on the network and latency.

In [118] Mok, et al., conducted a series of subjective experiments to evaluate the relationship between application QoS and QoE. They adopted both analytical and empirical approaches to the study of the correlation between network QoS and application QoS. The authors propose three application performance metrics (AMPs) to quantify the application QoS for HTTP streaming. The three metrics represent the temporal structure of video playback regardless of content. The three AMPs are;

1) Initial buffering time – the period of time between the start of downloading and the start of content playback
2) Mean buffering duration – the average duration of buffering events
3) Re-buffering frequency

From analysis of the results of their experiments the authors found that re-buffering was a major factor impacting on QoE. In tests high packet losses lead to lowered network throughput which increased the re-buffering frequency and thus reducing QoE.

Kim, et al., in [119] propose a QoE assessment model for video streaming services using QoS parameters. The authors identify delay, jitter, packet loss rate and available bandwidth as QoS parameters having the greatest impact on QoE. QoE related QoS parameters have differing levels of influence on QoE and should therefore be treated differently. The authors assign what they term Relative Importance Degree values to the QoS parameters identified earlier Packet Loss 58.9%, Jitter 15.1%, Delay 14.9% and Bandwidth 11.1%.

Yoon, et al., in [120] propose a system for video and web QoE assessment in LTE networks. The system, QoE Analytics, estimates QoE metrics per UE in real-time by inspecting user plane data without any QoE input from the UE. HTTP Live Streaming protocol (HLS) was used for testing, HLS is an adaptive streaming protocol. In HLS streaming over TCP the perceived user quality was affected by start-up delays and frequent rebuffering. Despite having sufficient bandwidth for the video bitrate in use stalling events and rebuffering can affect QoE, the stalling events and rebuffering occur due to network impairments such as delays

and packet loss. Results for testing showed that the QoS metrics reported by the QoE Analytics system were very close to the QoE metrics reported by the UEs.

Sideris, et al., in [121] investigated how the duration of the video segments selected might affect a MPEG-DASH user's QoE. The authors designed a set of experiments in which two MPEG-DASH users resided on the same access network and requested the same content. In each experimental round the users viewed the same content but using different segment durations during which they were asked to rate their QoE. Following analysis of the results of the experiments the authors found a clear correlation between the segment duration and the users reported QoE level. Users utilising a longer segment duration achieved a higher QoE level than the one utilising the shorter segment duration.

The authors of [122] propose a predictive packet drop technique to maintain a certain level of QoE based on predictive PSNR value without users' feedback. The proposed mechanism identifies packets as part of video frames and predicts the impact of their delay and loss on the resulting video performance. This reduces the need for client feedback and optimizes the resulting QoE of the delivered video. Based on this information, the proposed algorithm can prioritize the video frames (I-Frames, P-Frames or B-Frames) and decide whether to queue or drop them in scenarios where bandwidth is limited. This approach does not consider the impact on HD video content of dropping frames.

Seufert et al. in their work [123] conduct a simulative performance evaluation of the impact of Wi-Fi offloading for a mobile end user of a HTTP adaptive video streaming (HAS) service depending on availability and range of the Wi-Fi hotspots. This is done in the context of the strategy to lessen the load on cellular networks in cities by offering users the opportunity to offload mobile connections to lower cost Wi-Fi networks. The simulations were based on connectivity measurements from a German city and evaluates the key performance indicators for the QoE of HAS, i.e., initial delay, stalling, and quality adaptation. In addition, a smartphone energy model was applied to assess the energy consumption during the streaming. The results indicated that Wi-Fi offloading of HAS connections to public Wi-Fi hotspots is not attractive for end users both in terms of QoE and energy consumption. However, it was shown that Wi-Fi offloading can be beneficial for end users in cases where high bandwidths can be received via Wi-Fi.

The authors of [124] address the challenges faced by users of mobile devices that are connected to cellular networks. Connection quality may vary greatly over the course of a video and high QoE video streaming can be a challenge as the user data volume is metered and eventually limited. Prefetching videos is proposed as an approach to mitigate this issue. Here, videos that the user is likely to watch are prefetched or downloaded in advance on the user's smartphone, e.g., while he is connected to Wi-Fi. However, this approach can only be efficient if only the videos that the user will watch are identified in advance. This constitutes a major estimation and prediction challenge. To address this challenge, the work presents three contributions: 1) a user study over multiple months that draws valuable insights on the user video request behaviour. 2) a novel privacy-preserving prefetching framework denoted vFetch that prefetches videos based on the user's preferences. 3) a trace-based evaluation and parameter study that demonstrates vFetch's efficiency with a hit rate of ~50% for a 50 GB cache.

Park et al. in [125] examine the impact of Multipath TCP (MPTCP) on QoS and the end-user's QoE while streaming video over wireless links. Modern computing devices such as smartphones and tablets are multi-homed with Wi-Fi and 4G/LTE wireless interfaces and Multipath TCP (MPTCP) is a commonly proposed method of aggregating the bandwidth of these interfaces. However, the effect of MPTCP on the quality of experience of existing services, especially as a result of variances in bandwidth and latency among the individual paths over the wireless networks is not well-known. The authors explore the quality of service (QoS) and quality of experience (QoE) of adaptive video streaming using MPTCP over wireless networks. They conduct systematic measurements over three mobile network operators, AT&T, Verizon Wireless (VzW), and T-Mobile, along with Wi-Fi. Based on extensive measurements, they demonstrate that MPTCP can improve the QoS and QoE of video streaming only if the network interfaces have the roughly similar bandwidth and latency. The studies also show that MPTCP can perform worse than TCP in case of extreme differences between the network interfaces

3.9 Adaptive Bit Rate Algorithms

Adaptive bitrate (ABR) algorithms are used by modern video streaming clients to adapt the bitrate (i.e. quality) of the video being streamed to match as closely as possible the available bandwidth of the communications link. ABR algorithms make use of the MPEG-DASH standard first published in 2012 and revised in 2014 as MPEG-DASH ISO/IEC 23009-1:2014 [126]. In the event of a reduction in available bandwidth the ABR algorithm will request content at a reduced bitrate (and therefore quality) in order to avoid stalling events. Conversely, when link conditions improve the ABR algorithm will request video at a higher bitrate (and quality). This adaptive approach increases user QoE levels in general, but when the switches in quality are performed repeatedly may result in reductions of QoE for the end user, which needs to be avoided. Many ABR algorithms have been proposed, common ABR strategies that employ MPEG-DASH include buffer-based algorithms, throughput based algorithms and dynamic algorithms that include elements of both buffer and throughput approaches. Game theory based ABR strategies have also been developed.

Buffer based algorithms such as BOLA [127] and BBA [128] use the current level of the playout buffer to select the bitrate of the next segment to download. The buffer level provides an indirect view of network throughput; a high buffer level indicates that network throughput is good while a low buffer level indicates that network throughput is poor. This category of ABR selects high bitrate content on high buffer levels and low bitrate content on low buffer levels. BBA requires a relatively large buffer capacity to ensure stability during operations. BOLA also uses some throughput information to reduce oscillation between bitrates thereby enabling stable streaming at lower buffer capacities. Huanget al [129][139] demonstrated that buffer level was a good proxy for network throughput and had the benefit of being simple.

ABR algorithms such as FESTIVE [130] and PANDA [131] rely on estimating the available throughput on the link and then selecting an appropriate bitrate based on the estimated value. [129] also demonstrated that accurate estimation of available bandwidth can be difficult. Hybrid ABR algorithms such as ELASTIC [132] and ABMA+ [133] use a mixture of buffer based and throughput based algorithms in an attempt improve user QoE.

ELASTIC makes use of a feedback control system that tries to hold the buffer level at or close to a predefined level. It can, however, be slow to react to changes in throughput. ABMA+ starts with the aim of keeping the probability of re-buffering events occurring to a

minimum. It depends on complex estimation calculations that are computer offline ahead of time. The use of these offline computations makes the implementation of ABMA+ complex. The authors of [15] have developed three ABR algorithms BOLA-E that builds on the original buffer based BOLA, DYNAMIC and a FAST SWITCHING algorithm. The FAST SWITCHING algorithm can replace already downloaded video segments with segments of a higher bitrate if link conditions improve. The three algorithms presented in [134] have been implemented in the official DASH IF Reference player [135] and are being used in production environments.

Bentaleb et al have adopted a completely different approach to DASH ABR strategies in [136]. They have developed a game theoretical ABR algorithm that they have named GTA. This algorithm aims to maintain a high playback bitrate, reduce start-up delay while minimizing switches in bitrate and the occurrence of stalling events.

The ultimate goal of ABR algorithms is to improve or maintain QoE for the end-user. Timmerer et al [137] identified what they considered to be the most important QoE factors impacting DASH clients. These were the initial or start-up delay, stalling events, switches in quality and media throughput. The authors state that users experiencing stalling events report a very low QoE and that stalling events should be avoided at all costs even if this meant increasing the start-up delay. The position within the video stream at which a stalling event occurred had a significant impact on user QoE. Stalling events towards the end of the video have a higher impact on QoE than stalling events occurring at the beginning or middle of the stream.

Table 2 Summary Table Part 1

Area	Summary	Implementation
Network Detection	Wi-Fi scanning delays shown to be most significant contributor to Wi-Fi connection delays. Handoff delay can be separated into 3 sub-delays – Probe Delay, Authentication Delay & Association Delay. The Probe Delay accounts for more than 90% of overall delay and hardware selection can also affect delay. Little difference in connection time patterns between Wi-Fi networks using 2.4GHz and 5 GHz. Also little difference in connection time costs for users in enterprise Wi-Fi networks and users in home Wi-Fi networks. Reduce scanning delay to reduce overall connection delay. A simple but effective scanning strategy is to only scan a sub-set of channels e.g. only scan channels 1,6, and 11 which are non-overlapping channels that are likely to be in use.	Connection delay plays a vital role as an input to the SONS utility function (Chapter 4) as part of the total delay calculation
Energy Conservation	All mobile devices are constrained by the capacity of their battery and all active hardware components consume energy. Networking operations in mobile devices can account for a significant proportion of total power consumed. Wi-Fi scanning operations consume energy even when no APs are detected. Full, all channel scans are energy inefficient. All wireless networking operations consume energy, once an interface has been activated the actual amount of data transferred over the interface has little impact on energy consumption. For energy conservation it is better to transfer as much data as possible over the activated link since the energy cost associated with the network operation has already been incurred. Energy consumption can be reduced by reducing the number of wireless scanning operations conducted. Energy can also be saved by shutting down wireless interfaces when not in use.	Energy conservation through shutting down interfaces when not in use or reducing scans is handled in SONS(Chapter 4), MDF (Chapter 5) and AIS (Chapter6)
Data Offloading	Growth of video traffic creates pressure on network resources, projected growth in IoT traffic is expected to worsen the situation. Mobile phone service providers offload data from to Wi-Fi in order to protect their cellular networks. Service providers impose data-caps on pay-as-you-go subscriber accounts to control the amount of data downloaded over their cellular networks. In order to protect their data-caps mobile users offload data transfers to Wi-Fi whenever possible	Data offloading to Wi-Fi networks is an important goal of MDF (Chapter 5) and AIS (Chapter 6)
Patterns of Movement in Urban Environments	Pedestrian average walking speed is 1.3 metres per second in a variety of environments, passengers in vehicles in urban areas move at an average rate of 4.6 metres per second. Vehicular movements within built environments cannot be random as they must follow streets. Pedestrian movement patterns are not random; they walk is straight lines as much as possible. Crowds of pedestrians develop self-organising behaviour if crowd density is high enough and the behaviour of large groups of pedestrians is not random. Most mobile users are not in fact very mobile, they tend not to move far and favoured locations tend to be close together.	User movement patterns and behaviours informed the NS3 simulations described in Chapters 5 & 6

Table 3 Summary Table Part 2

Area	Summary	Implementation
Quality of Experience (QoE)	QoE is a subjective metric typically measured using expensive panels of assessors. The need for human assessors make it impossible to measure QoE in real-time for streamed video content. User QoE is strongly influenced by available bandwidth and latency. Key performance indicators for QoE are initial delay, number of stalling events and frequent changes in video quality. The number and duration of stalling events is important as is the location in the video stream at which the stalling event takes place. Stalling events towards the end of the video stream have a greater impact on user QoE than stalling events that occur at either the start of the video stream or in the middle of the stream.	Optimising QoE is a primary objective of both MDF (Chapter 5) and AIS (Chapter 6)
Adaptive Bit Rate Algorithms (ABR)	ABR algorithms are used by steaming clients to adapt the bitrate of streamed video to match the available bandwidth of the communications link. In general, this adaptive approach increases user QoE but repeated switches in video quality reduces user QoE. Many ABR algorithms have been proposed including buffer based, throughput based and dynamic algorithms. Buffer based algorithms use the current playout buffer level to select the bitrate of the next segment to download, buffer level provides an indirect view of link throughput. Throughput based algorithms rely on estimating throughput on the link but estimation of throughput can be difficult. Dynamic ABR algorithms use a mixture of buffer based and throughput based algorithms in a bid to improve QoE. All ABR algorithms use one interface at a time.	AIS (Chapter 6) uses the playout buffer level as a metric to decide which action to take.
Utility Functions	Utility functions are well known multi-criteria decision making methods. They summarise the preferences of the user in terms of how much utility the user gets from consuming goods or services or from selecting one network over another. Utility functions are frequently used in complex decision making scenarios. They take multiple, separate input criteria and produce a single value output that simplifies the decision making process.	SONS (Chapter 4) employs a novel utility function to determine when the user device should scan for available Wi-Fi APs.

CHAPTER 4 SCAN-OR-NOT-TO-SCAN (SONS)

This chapter focuses on the Scan-Or-Not-to-Scan (SONS) framework. The SONS framework is discussed in detail and the SONS utility function is described. Verification of SONS through modelling and simulations is described and the results presented and discussed.

4.1 Motivation

For many people around the world their mobile device is their only gateway to the data and services hosted on the Internet. This is not just the case in developing countries, a 2019 study by Pew Research [5] shows that 37% of adults in the USA mainly use their smartphones to access the Internet. In the 18 to 29-year-old age group 58% of people stated that their smartphones are their primary means of connecting to the Internet. Additionally, this cohort of younger mobile subscribers typically uses pre-paid, fixed data allowance plans and want to reduce the amount of data downloaded over cellular links whenever possible in order to protect their data-caps. For mobile users in urban HetNet environments the opportunity exists for them to change their point of attachment from one type of network to another in order to improve their connectivity, to improve download speeds or protect data-caps. Typically, in currently deployed wireless environments, this will involve switching the point of attachment from a cellular network connection to a Wi-Fi AP and vice versa. To connect to the wireless network that best meets their needs a user must first detect available networks, then select the most suitable network. However, it may not always be in the best interest of the user to invoke a network detection and selection strategy in an attempt to improve connectivity. Scanning for available Wi-Fi APs consumes energy and when there is no realistic chance of establishing a useful connection to a Wi-Fi network this is a waste of the already limited energy resources of mobile devices. Under certain conditions it may also be detrimental to the end user's Quality of Experience (QoE) to switch networks, even when scanning has detected an apparently 'better' network.

When switching the Point of Attachment from a mobile phone network to a Wi-Fi AP the cellular connection can be maintained until the connection to the Wi-Fi AP has been established. This 'make-before-break' strategy means that the node is rarely in a completely disconnected state for prolonged periods of time. However, when connectivity to the Wi-Fi AP

is lost due to the user moving out of range of the AP there is a period of time during which the node has no active data connections to any network. This disconnected state is a result of the unavoidable delay experienced in establishing a data connection to the mobile network and while in this state the node relies on the contents of the playout buffer to maintain video playback. In this situation the playout buffer can become depleted before a new data connection is established leading to stalling events in the video playback. Stalling events, especially when they occur towards the end of the video, have a negative impact on user QoE. This becomes particularly problematic when the mobile node is travelling at a speed such that connections to Wi-Fi APs with durations of only a second or two are established. Automatically connecting to Wi-Fi AP without regard to the prevailing circumstances will have the user's device continuously establishing Wi-Fi connections over which little or no useful data can be transferred but which result in complete loss of connectivity on each occasion. In this context there is a need for a solution which determines when it is feasible to perform network detection scans and invoke network selection algorithms such as to achieve good QoE given finite device energy resources.

This chapter presents the novel Scan-Or-Not-to-Scan (SONS) framework. The SONS framework decides, based on device sensors and other inputs, when it is appropriate for the user to conduct network detection scans and execute network selection strategies and when it is not. The use of the SONS framework enables the user to conserve energy by shutting down unused interfaces and minimising the number of potentially unsuccessful scans for Wi-Fi APs. By reducing both the number of connection events in which little or no data can be received over Wi-Fi and the number of unnecessary handovers SONS helps maintain the mobile user multimedia QoE at high levels. SONS employs the user's remaining data cap (RDC) as a significant input in its decision making process. To the best of our knowledge SONS is the first scan or no scan decision making process that considers the user's data cap.

4.2 The SONS Framework

The SONS framework is not a network detection mechanism, rather it seeks to reduce the number of unnecessary scanning operations by determining when successful scanning operations might be carried out and when they might fail. Other systems for reducing the number of scanning operations carried out by a node have been proposed such as iScan. The authors of [140] proposed iScan, a scheme that modifies the node's Wi-Fi scanning behaviour based on network load and node speed over the ground. If the node is stationary iScan conducts scans based on the network load only. Unlike SONS iScan does not consider AP coverage area or the users remaining data-cap and does not prevent scanning when the node speed over the ground is too high to enable a useful connection to an AP to be established.

Another approach to reducing the number of scans carried out by a node is proposed in [141]. Han et al. propose SplitScan which uses Bluetooth to exchange information regarding detected Wi-Fi APs with nearby nodes in an attempt to reduce the need for scanning. However, this approach does not work if there are no adjacent nodes or the nodes in the vicinity are beyond range of the Bluetooth adapter. In addition, SplitScan does not take user speed over the ground or the users remaining data-cap into consideration.

The SONS framework is designed to be generic and to work with any network detection and selection algorithm specified by the user. Figure 12 depicts a typical urban HetNet wireless environment in which SONS would operate, an environment where both cellular networks and Wi-Fi APs are deployed. The mobile phone networks provide coverage over relatively large geographic areas and independent Wi-Fi APs within the cellular coverage areas provide isolated, possibly non-overlapping islands of Wi-Fi connectivity. As the user travels through the urban environment they pass through both cellular and Wi-Fi coverage areas.

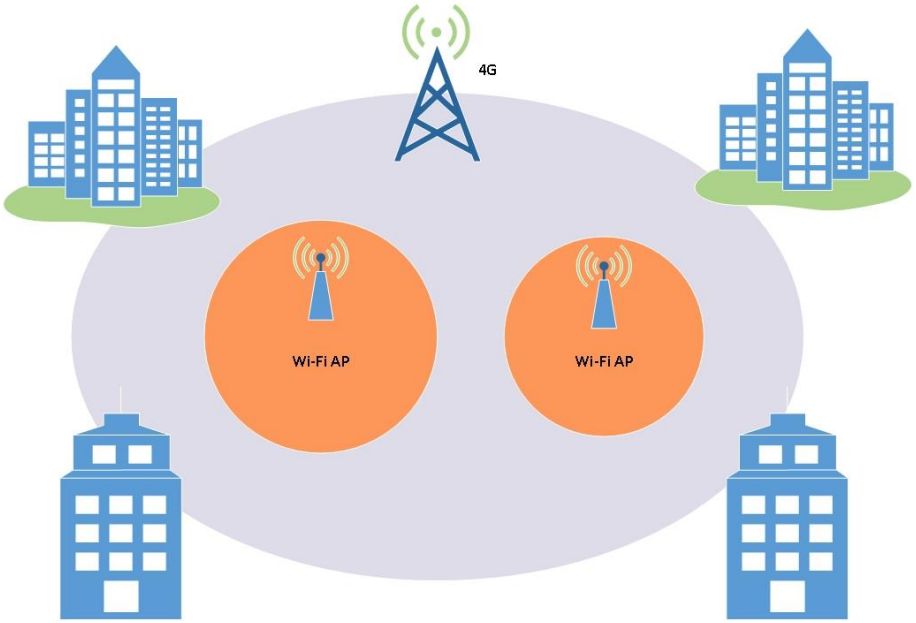

Figure 12 Urban HetNet containing 4G and Wi-Fi APs

The SONS framework takes into consideration the fact that the speed at which a mobile node is travelling has a significant impact on whether or not a useful connection can be established to a Wi-Fi network. SONS uses the average speed at which the mobile node is travelling and the end-users remaining data-cap as inputs into its decision-making process. The end-user's remaining data-cap (RDC) plays a vital role in the decision making process, the lower the user's cellular RDC the more useful it is to the user to establish a connection to a Wi-Fi AP. It is important to bear in mind that SONS is not a network selection algorithm, but rather a decision-making mechanism for determining when to execute a network selection strategy.

4.2.1 Mobile User Maximum Speed to Support a Useful Wi-Fi Connection

When establishing a connection to a wireless network the user will invariably experience delays due to the scanning and connection processes. These delays can have a significant impact on the amount of time that the user device is connected to the network. The amount of time a mobile user remains in range of a Wi-Fi AP depends on the coverage area of the AP and the speed at which the mobile user is travelling. To establish a connection this dwell time, that is the amount of time spent within range of a Wi-Fi AP, must be greater than the inevitable scanning delay and connection delay combined.

The dwell time is calculated using the formula:

$$\textbf{dwell time} = \frac{APC}{SM} - (\textbf{sd} + \textbf{cd}) \quad (4.1)$$

where *APC* is the diameter of the Access Point coverage area in metres, *SM* is user speed in metres per second, *sd* is the scanning delay, *cd* (connection delay) is the time taken to establish a connection to an AP.

Clearly if a mobile user is travelling at too high a speed then they will not remain in range of a Wi-Fi AP for a period of time sufficient to enable a useful connection to be established. In the context of SONS, the threshold speed is the speed at or below which a user must travel in order to be able to establish a useful connection to a Wi-Fi AP. To calculate a reasonable threshold speed, we require realistic Wi-Fi AP coverage areas and delay times.

4.2.2 Wi-Fi Access Point Coverage Areas

To develop a realistic understanding of real-world Wi-Fi deployments for use in the design of SONS, a survey of Wi-Fi APs was conducted in an area of Dublin city centre during late August 2015 (Figure 13). The survey was conducted using a Nexus 7 tablet and the Wigle [142] wardriving application which employed the tablet's on-board GPS unit to map the locations of detected Wi-Fi APs. Many buildings in the survey area are old, solidly constructed with brick and stone structures and are typical of the type of buildings found in many urban environments. The solid nature of their construction results in walls that very effectively block radio signals, with a severe limiting effect on the coverage area of the detected Wi-Fi APs. This blocking effect is even more pronounced with higher frequency systems such as 5GHz Wi-Fi and 5G cellular networks.

The O'Connell Street area of Dublin city, which lies on the north side of the river Liffey, was selected for the survey as it contained a good mixture of transport links such as bus and tram lines, coffee shops, businesses, etc. It is also one of the main thoroughfares of the city and has a large number of pedestrians at all times of the day. The survey area was restricted to the east side of O'Connell street between 53.350386 degrees N, 06.260354 degrees W and 53.349817 degrees N, 06.260067 degrees W, a distance of approximately 80 metres. These positions correspond to the junction of O'Connell Street and Cathedral Street at the northern

end of the survey area and the junction of O'Connell Street and Earl Street North at the southern end of the survey area outlined in red in Figure 13.

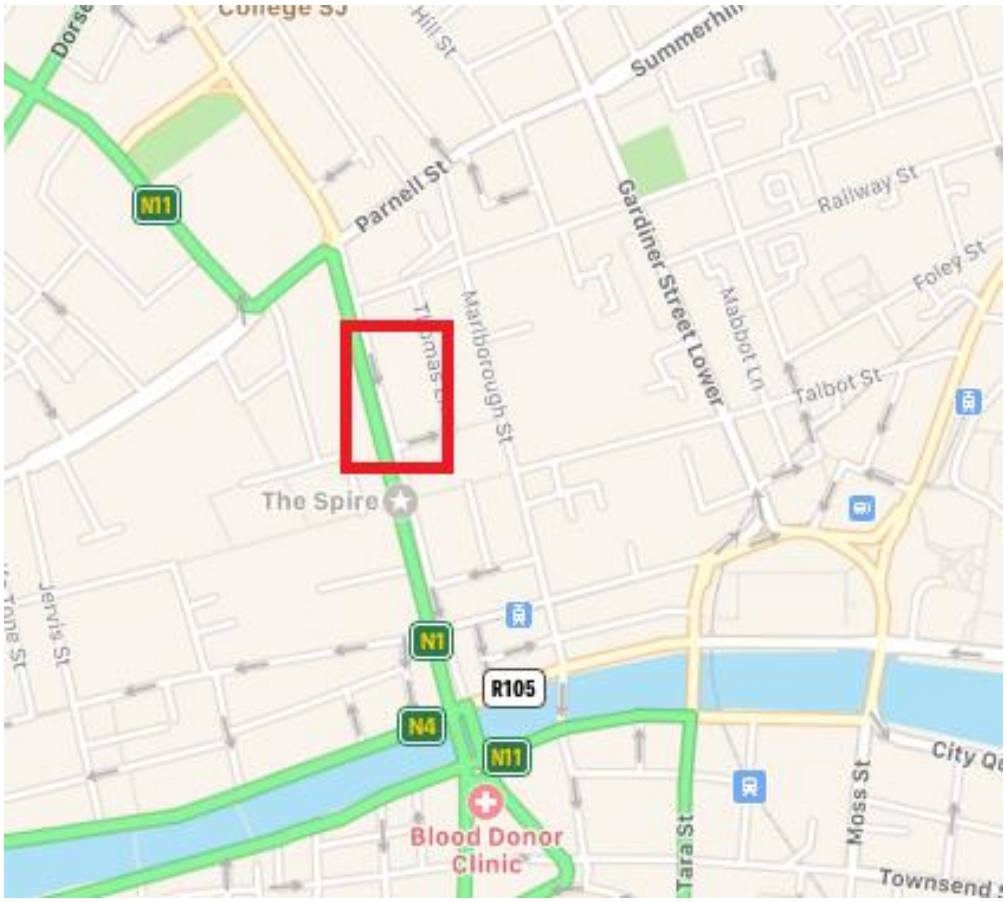

Figure 13 Wi-Fi AP Survey Area

An initial scan of the survey revealed a potential problem, namely a very large number of Wi-APs that appeared to be positioned on the roadway and footpaths as indicated by blue marks in Figure 14.

Figure 14 Results of initial scan of survey area

Closer examination of the initial scan results (Figure 15) revealed that some of the reported Wi-Fi APs were in fact mobile devices. It is assumed that these devices had been operating as mobile broadband hotspots and that users forgot to turnoff this functionality when in motion.

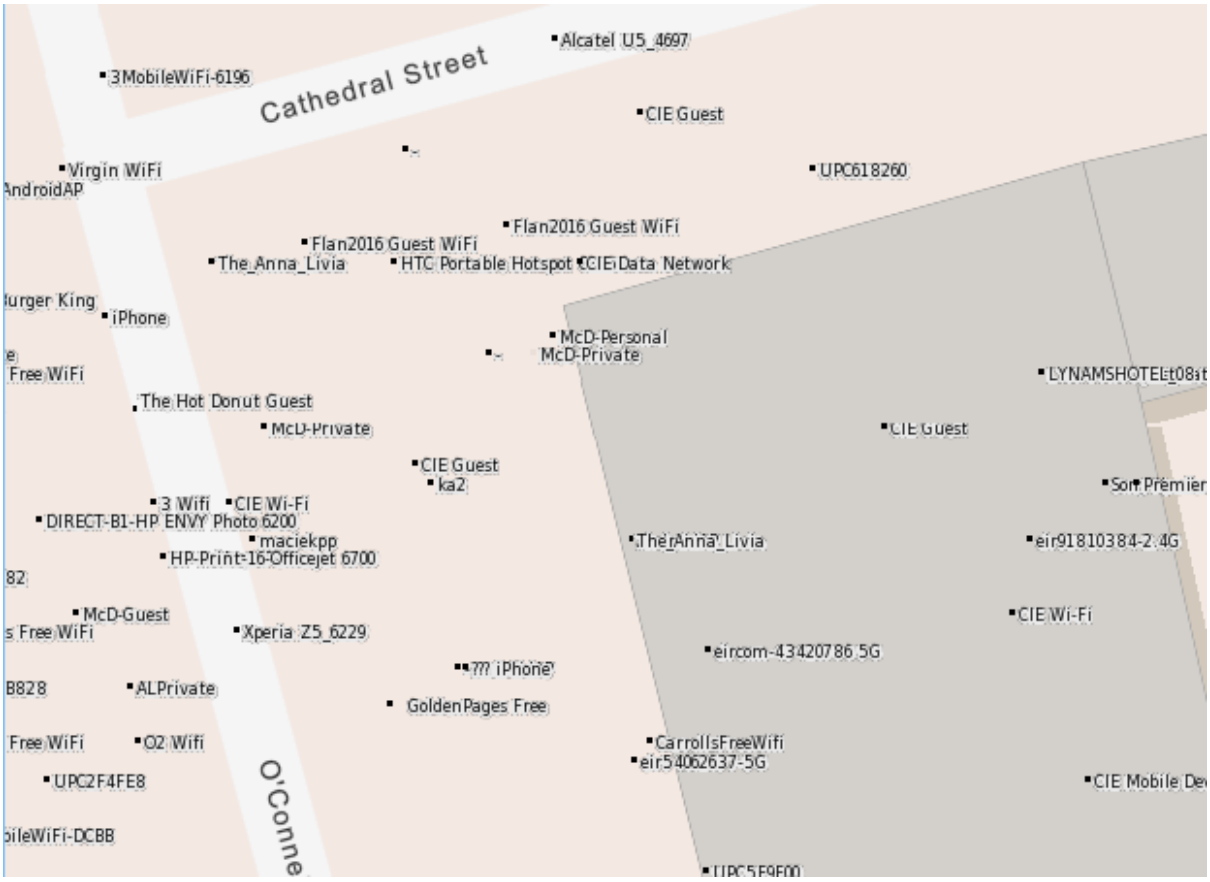

Figure 15 Detailed View of Initial Scan Results

Other reported APs were office equipment such as printers and it was obvious that the positioning information was inaccurate. It appears that the inaccuracies in GPS positioning were as a result of partial blocking of the GPS signals by the surrounding built environment (Figure 16). To calculate its position a GPS-enabled device measures its range or distance from multiple GPS satellites. If the device has a very accurate clock it requires signals from 3 GPS satellites to calculate an accurate position. A device that does not have an accurate clock will require signals from 4 satellites to calculate an accurate position. GPS-enabled smartphones are typically accurate to within a 4.9 metre radius under open skies. The GPS satellites broadcast their signals in space with a certain level of accuracy but what the receiver gets depends on multiple, additional factors including receiver design features and quality, atmospheric conditions, satellite geometry and signal blocking.

Many factors can degrade GPS positioning accuracy, including:

- Satellite signals being blocked by buildings and other infrastructure such as bridges
- Signals being reflected off buildings and walls (“multipath”)

- Poorly designed GPS hardware

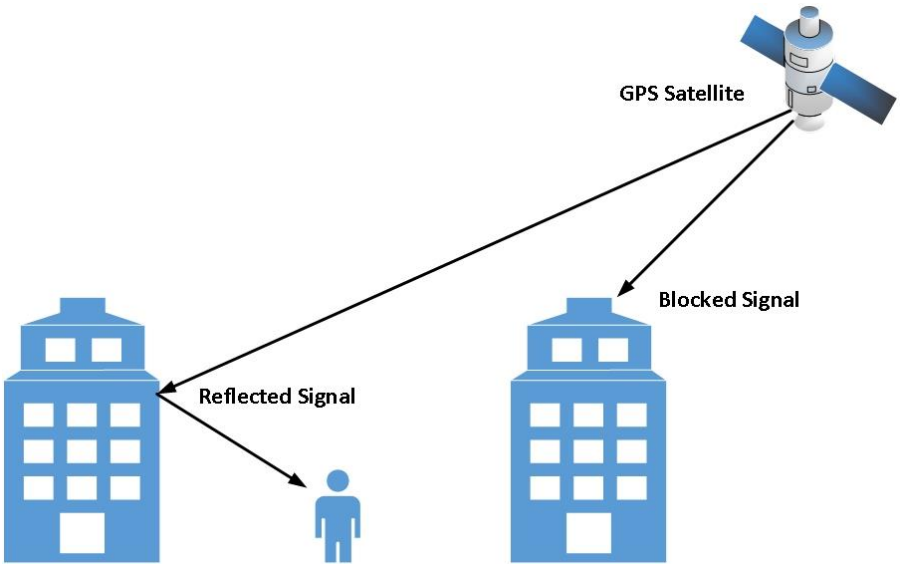

Figure 16 GPS – Impediments to GPS signal reception

Often the GPS hardware is operating correctly and the problem lies with the mapping software. Some of these issues include:

- Incorrectly drawn maps
- Missing roads, buildings, etc.
- Incorrectly estimated street addresses
- Mislabelled businesses and other objects of interest

It was assumed that the confused results of the initial scan were the result of a combination of issues involving the GPS positioning systems, possibly due to multipath and signal blocking effects, and the mapping software. A selection strategy was required to winnow out inappropriate APs from the multitude returned by the initial scanning operations. Many of the SSIDs seen in the scanning results contained information that identified businesses in the area and it was decided that only APs that could be mapped to the physical location of a business in the area would be selected.

Figure 17 Wi-Fi APs selected during the survey

Using this strategy 20 SSIDs that could be mapped to physical location were selected and the process of estimating the coverage areas of the associated APs was begun. The procedure for estimating the approximate diameter of each of the coverage areas was as follows; for APs 1 to 10 shown in Figure 17 the mobile device used in the survey was moved from north to south parallel to the block of buildings housing the Wi-Fi APs. When an AP was first detected the position was noted and this position was considered to mark the northern edge of the coverage area. This procedure was repeated for APs 1 to 10 but moving from south to north, when an AP was detected the position was recorded and this marked the southern boundary of the APs coverage area. The same strategy was employed for APs 11 to 15 on Earl Street North and APs 16 to 20 on Cathedral Street. Having discovered the approximate boundaries for the selected APs a diameter for each coverage area was calculated, the results are presented in Table 4.

Table 4 Estimated Diameters of AP coverage areas

Estimated Coverage Area of Survey Access Points			
AP Number	Est. Diameter (metres) & Direction of Travel	Start Position	End Position
1	20m N ⇔ S	53.350473 N 6.260140 W	53.350273 N 6.260140 W
2	40m N ⇔ S	53.350534N 6.260121 W	53.350130N 6.260119 W
3	30m N ⇔ S	53.350450N 6.260100 W	53.350150N 6.260080 W
4	45m N ⇔ S	53.350469N 6.260070 W	53.350029N 6.260060 W
5	35m N ⇔ S	53.350407N 6.260045 W	53.349947N 6.260030 W
6	40m N ⇔ S	53.350322N 6.260022 W	53.349922N 6.260017 W
7	20m N ⇔ S	53.350183N 6.260009 W	53.349983N 6.260007 W
8	80m N ⇔ S	53.350430N 6.259971 W	53.349630N 6.259962 W
9	50m N ⇔ S	53.350204N 6.259954 W	53.349700N 6.259950W
10	60m N ⇔ S	53.350245N 6.259933 W	53.349645N 6.259929 W
11	40m E ⇔ W	53.349953N 6.260175W	53.349953N 6.259615W
12	25m E ⇔ W	53.349972N 6.260010W	53.349972N 6.259610W
13	35m E ⇔ W	53.349981N 6.259984W	53.349981N 6.259484W
14	20m E ⇔ W	53.349999N 6.2599775W	53.349999N 6.2599500W
15	40m E ⇔ W	53.350024N 6.2599813W	53.350024N 6.2599253W
16	45m E ⇔ W	53.350375N 6.260450W	53.350375N 6.259750W
17	40m E ⇔ W	53.350392N 6.260247W	53.350392N 6.259687W
18	35m E ⇔ W	53.350411N 6.260110W	53.350411N 6.259620W
19	13m E ⇔ W	53.350427N 6.259853W	53.350427N 6.259653W
20	45m E ⇔ W	53.350451N 6.259943W	53.350451N 6.259313W
Mean APC	**37.9 metres**		
SD	**14.859**		

The largest AP coverage area observed during the survey had a diameter of approximately 80 m and the smallest was approximately 13 m in diameter with many diameters clustered around 35 to 45 metres. As each AP was an independent, standalone device there was no standardised distribution of coverage area sizes. Even in scenario in which all the APs were part of the same organisation each coverage area size would vary due to device placement within a building, the size and internal layout of each building and the thickness of external and internal walls.

In this context it was decided to determine the mean coverage area diameter and standard deviation for the surveyed APs and an average coverage area diameter of 37.9 metres with a standard deviation of 14.859 was calculated.

The Empirical Rule (68-95-97.5 Rule) states that 68.27% of results will fall within the mean plus or minus one standard deviation. This rule in conjunction with the APC mean and standard deviation informs the SONS-related research in terms of setting default values for APCmin and APCmax which are defined as follows:

APCmax = μ + 1 standard deviation (4.2)

APCmin = μ - 1 standard deviation (4.3)

APCmax = 37.9 + 14.859

APCmax = 52.759

APCmax = 53m

APCmin = 37.9 - 14.859

APCmax = 23.041

APCmin = 23m

Because the Standard Deviation was quite large in relation to the mean AP Coverage area it was decided to use 1 SD in calculating both APCmax and APCmin. An initial calculation of APCmin using 2 SDs resulted in an APCmin value of approximately 8.2m which was completely unrealistic and the decision was taken to only use 1 SD in the calculations of APCmax and APCmin.

During operations SONS will use the APC_{max} and APC_{min} values calculated above as its default values. Over time SONS builds a historical record of AP coverage areas which are mapped to the AP's geological location for future use. In the absence of this historical data SONS reverts to the APC_{max} and APC_{min} default values.

4.2.3 Threshold Speed

Threshold Speed(TS) is the maximum speed at which a mobile node can establish a useful connection to a detected AP. In order to calculate the Threshold Speed, we require the AP coverage area diameter and the total delay(td). For the purpose of calculating user Threshold Speed, we assume an AP coverage area diameter equivalent to the mean AP coverage area diameter plus one standard deviation based on the results from the Wi-Fi survey (Table 4). Table 5 presents the results of tests for scanning delay and Wi-Fi connection delay observed at various locations using both a smartphone and a laptop computer

The average duration of the observed Wi-Fi scans was 3.59 seconds with a standard deviation of 0.464344, similar to the scan delays reported in [106], and an average delay in connecting to an AP of approximately 4.498 seconds (standard deviation 0.286813) was experienced. Both the scanning delay and the connection delay were measured manually using an electronic stop watch.

Table 5 Wi-Fi Scanning, Connection and Page Load Delay Times

Location	**Device Type**	**Scanning Delay (seconds)**	**Connection Delay (seconds)**
Bus [onboard]	Smartphone	3.6	4.8
Train [onboard]	Smartphone	3.8	4.7
Train [onboard]	Laptop	3.9	4.6
Coffee Shop	Laptop	3.6	3.98
Coffee Shop	Smartphone	3.2	4.1
College Reception Area	Laptop	3.6	4.3
College Reception Area	Smartphone	3.3	4.8
Bus Stop	Smartphone	4.0	4.7
Train Station	Smartphone	3.3	4.5
	Mean	**3.59**	**4.498**
	Std Dev	**0.264344**	**0.286813**

We define the maximum total delay (tdmax) and the minimum total delay (tdmin) as follows:

Td(max) = (msd + 2 SDs) + (mcd + 2 SDs) (4.4)

Where td is total delay, msd is mean scan delay, mcd is mean connection delay and SD is Standard Deviation, as the Standard Deviation is quite small in relation to the mean Scanning and Connection delays it was decided to use 2 SDs in calculating td(max).

The Threshold Speed (TS) is calculated using the mean AP coverage area diameter plus 1 standard deviation and the maximum total delay td(max). We first calculate td(max):

td(max) = (msd + 2 SDs) + (mcd + 2 SDs)

msd = 3.59, msd SD = 0.264344

mcd = 4.498, mcd SD = 0.286813

td(max) = (3.59 + (2 * 0.264344)) + (4.498 + (2 * 0.286813))

td(max) = 4.1187 + 5.072

td(max) = 9.19 seconds

Threshold Speed is the maximum speed at which a mobile node can travel and continue to detect a Wi-Fi AP, connect to the detected AP and download a webpage.

The Threshold Speed (TS) is calculated as follows:

TS = AP Coverage Area Diameter / maximum total delay

TS = (mean APC + 1 SD) / td(max) (4.6)

From Table 1 we get a mean APC coverage area diameter of 37.9 metres and a SD of 14.859

TS = (37.9 + 14.859) / 9.19

TS = 52.76 / 9.19

TS = 5.74 metres per second

Therefore, a mobile node traversing a heterogeneous wireless network would need to be moving at a speed of approximately 6 metres per second or less in order to make a useful connection to a Wi-Fi AP. In light of this a value of 6 metres per second was selected as the default threshold speed for the SONS framework.

4.3 SONS Utility Function

A novel SONS utility function is introduced to indicate the mobile user's preference regarding attempting to scan for alternative wireless networks. This utility function simplifies the assessment of complex situations that involve uncertainty or risk and enables us to formalise the decision making process. Cardinal Utility is the assignment of a numerical value to utility, and models that incorporate cardinal utility use the theoretical unit of utility, the util, in the same way that any other measurable quantity is used. In this work utility is expressed simply as a number and the output from the SONS utility function, the utility score, is used to decide whether or not to activate the Wi-Fi interface and begin scanning operations. Because the user speed, AP's coverage area, total connection delay and the amount of data remaining in the user's data allowance have a significant impact on the decision making process, SONS includes these parameters in its utility value calculation.

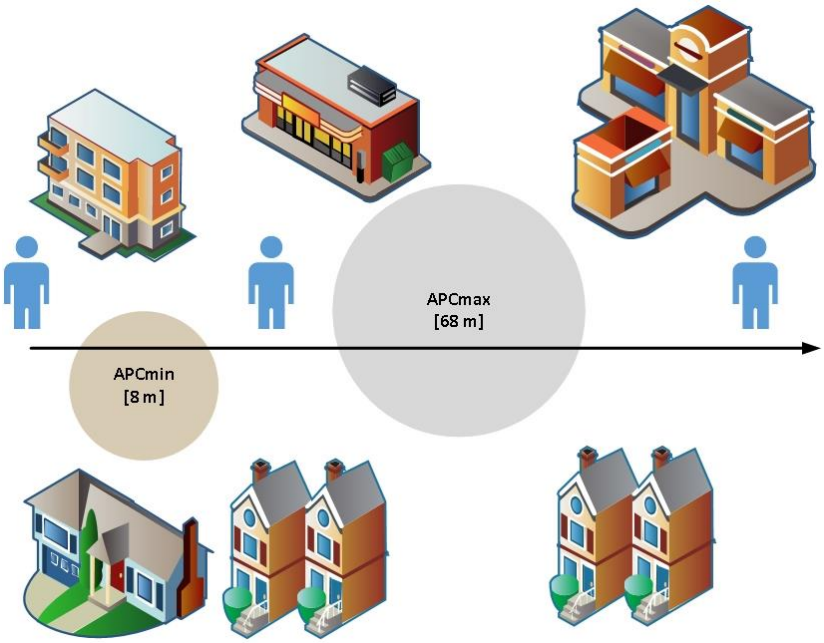

Figure 18 APCmin and APCmax

Without loss of generality we assume the coverage area of a Wi-Fi AP to be a circle with diameter APC. The AP's coverage area is defined as *APC (diameter)*, users speed in metres/second is defined as *SM*, combined connection delay is defined as *td*.

The user's maximum daily data allowance in MB is calculated by dividing the total data allowance by the number of days in the billing period and the result is defined as RDC_{max}. The user's remaining data-cap is defined as *RDC* and it is the remaining number of MB of allowance from the current date until the end of the billing period divided by the number of days remaining in the billing period. On day one of the billing period RDC_{max} will be equal to *RDC*, RDC_{max} will remain constant throughout the billing period but *RDC* will reduce in value over the billing period as the user consumes their data allowance. The data reserve limit set by the user is referred to as RDC_{min}. When initialised, SONS calculates the maximum and minimum utility scores s_{max} and s_{min} and any subsequent utility scores will be evaluated to see if they fall between these upper and lower boundaries.

The SONS utility function is presented in (4.7) and lower bound and upper bound utilities are calculated with formulas (4.8) and (4.9).

$$\mathbf{s} = \left(\frac{APC}{SM} - \mathbf{td}\right) * \frac{1}{RDC} \qquad \mathbf{(4.7)}$$

$$\mathbf{s_{min}} = \left(\frac{APCmin}{SMmax} - \mathbf{td}_{\mathbf{max}}\right) * \frac{1}{RDC\ max} \qquad \mathbf{(4.8)}$$

$$\mathbf{s_{max}} = \left(\frac{APCmax}{SMmin} - \mathbf{td}_{\mathbf{min}}\right) * \frac{1}{RDC\ min} \qquad \mathbf{(4.9)}$$

SMmin is 1 for all speeds less than 1 since a mobile node at rest will have a speed of 0mps and we cannot divide by zero

$$\mathbf{U(s)} = \begin{cases} \mathbf{1}, \boldsymbol{smin} < \boldsymbol{s} \leq \boldsymbol{smax} \\ \mathbf{0}, \boldsymbol{otherwise} \end{cases} \qquad \mathbf{(4.10)}$$

Output of the SONS utility function, the utility score U(s), is a dimensionless quantity and is expressed as a whole number either 1 or 0.

4.4 SONS Illustrative Example

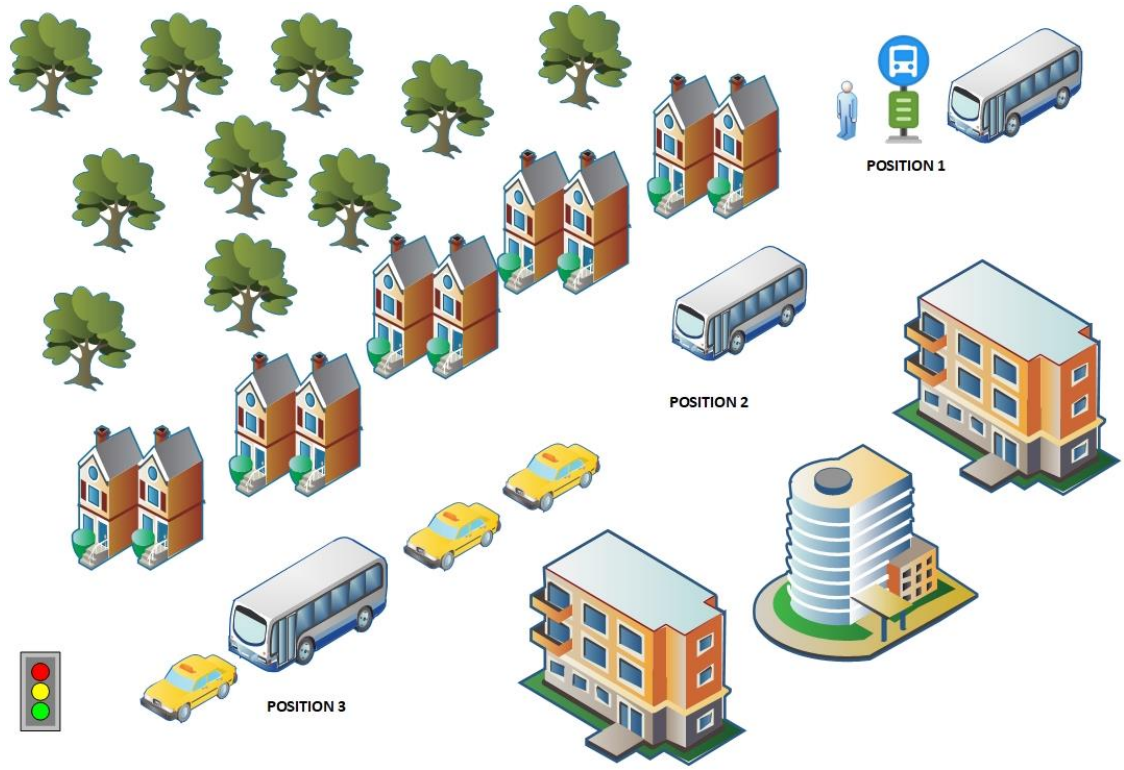

Figure 19 Bob's Commute

Bob is an urban commuter who travels between home and work by bus and on foot (Figure 19). Each workday Bob leaves home on foot and walks to the nearest bus-stop (Position 1). When Bob reaches the bus-stop he usually has to wait for the bus to arrive and he passes the time by watching some video clips on his smartphone. Bob's smartphone is a multi-homed device equipped with an interface for connecting to mobile cellular networks and an IEEE 802.11 interface for connecting to Wi-Fi APs. In this scenario the device has automatically established a connection to Bob's mobile phone provider's network and the Wi-Fi interface is de-activated.

Bob wants to protect his cellular data-cap and reduce energy consumption but he is also prepared to use his mobile data allowance whenever necessary to ensure that he can enjoy his videos with as few stalling events as possible. In order to achieve this objective, Bob has installed the SONS framework on his device.

The SONS framework on Bob's smartphone runs in the background as a service and monitors user speed, data-cap, battery level and the MPD-Cache directory.

Each time Bob's smartphone wakes from sleep mode SONS calculates his maximum Remaining Data Cap (RDC_{max}), this is done by dividing Bob's data allowance for the billing period by the number of days in the billing period. For a user with a 50GB data allowance **(50GB = 51200MB)** and a 28-day billing period the RDC_{max} on day 1 of the billing period is:

$$\boldsymbol{RDC_{max} = Total\ Data\ Cap\ /\ Days\ in\ Billing\ Period}$$

$$\boldsymbol{RDC_{max} = 51200\ MB\ /\ 28 \quad = 1892\ MB}$$

SONS then generates the lower bound utility value which is calculated using the lower bound formula (4.8).

Default values for various inputs such as delay, AP coverage area, etc. are employed by SONS for its calculations in the absence of real world information, particularly during start-up and when the device exits sleep mode. The default values are presented in Table 6 below.

Table 6 Default Values for Use in SONS Utility Score Calculations

Default Values Used in SONS Utility Functions				
Metric	**Mean**	**Std. Dev**	**Formula Used**	**Default Value**
td_{max}	3.59 4.498	0.264344 0.286813	5.4	9.2 sec
td_{min}	3.59 4.498	0.264344 0.286813	5.5	7 sec
APC(max)	37.9m	14.859	5.2	53m
APC(min)	37.9m	14.859	5.3	23m
SM_{max}	NA	NA	5.6	6 mps

In order to determine the maximum total delay td_{max}, SONS monitors the delay for each streaming session and records the longest total delay.

In the event that there is not any entry associated with td_{max} , typically at system start-up, a default value of 9 seconds is used (Table 6). APC_{min} is the smallest diameter of Wi-Fi AP coverage areas for recently detected Wi-Fi APs. In the event that there has been no recent detection of APs a default value of 23m is employed (Table 6). A default value for SM_{max} of 6 metres per second, the maximum user speed over the ground at which a useful connection to a Wi-Fi AP can be established, is used (Table 6).

The lower bound utility value (4.8) in this example is calculated as

$$\mathbf{s_{min} = (23/6 – 9) * 1/1892 = -0.00273079}$$

SONS then generates the upper bound utility value which is calculated using formula (4.9). The RDC(min) value is set by the user and represents the portion of their data allowance that the user wishes to reserve to ensure that they have mobile data available for essential services such as maps/navigation, online searches, WhatsApp, etc. Bob has decided to reserve 1000 MB for essential services. The td(min) value is the minimum total delay, again SONS tracks the total delay for each streaming session and records the shortest total delay in a text file. In the event that the td(min) file is empty a default value of 7 seconds is used (Table 6). APC(max) is the largest Wi-Fi AP coverage area diameter in metres that has been detected, the default value is 53 metres (Table 6). The SM(min) value is the minimum user speed over the ground, for user speeds of less than 1 metre per second the default value is 1 metre per second.

The upper bound utility value (4.9) in this example is calculated as follows:

$$\mathbf{s_{max} = (53/1 – 7) * 1/1000 = 0.0046}$$

SONS now calculates Bob’s utility score using formula (4.7).

Since Bob is stopped at a bus stop SM = 1 metre per second (note SM = 1 mps for all values less than 1), APC = 80m, td = 8 seconds and RDC = 1600 since we assume that some portion of the daily data allowance has already been consumed by various activities.

$s = (80/1 – 8) * 1/1600 = 0.045$

The overall utility is computed as in (4.10) and since $-0.00273079 < 0.045 \leq 0.046$:

$$\mathbf{U(s) = 1}$$

An overall utility score of U(s) = 1 in this example implies the Wi-Fi interface can be activated and a user specified network detection and selection algorithm implemented.

When Bob selects a video that he wishes to view the video's MPD is downloaded to a temporary directory named MPD-Cache. The MPD-Cache directory is empty by default and SONS checks the directory size once per second. When SONS detects that the folder size has changed it assumes that a manifest file has been downloaded and it invokes the Resolution Discovery Module (RDM) which is described in Chapter 5.

4.5 SONS Principle, Architecture and Algorithm

For the mobile user wishing to connect to Wi-Fi networks, time is of the essence since the window of opportunity for making decisions regarding network selection may be very short. Complex algorithms require more time to execute than do simple algorithms. Consequently, a very simple decision-making process in which there are only four inputs that enables decisions to be made quickly is considered.

Modern urban HetNets provide commuters, equipped with multi-homed devices, with an opportunity to be always "best connected". These multi-homed devices are typically equipped with both a cellular interface and a Wi-Fi interface. For the user, being best connected means connecting to the network that best meets their current needs and in order to do so the user must be able to detect wireless networks in their proximity, determine which network is the "best" according to some selection process and connect to it. Typically, on initialisation a mobile user's device will first establish a connection to the cellular network of their mobile phone service provider. If the user wishes to change their point of attachment to another network, they must scan for available Wi-Fi networks and the Wi-Fi scanning process may be started manually by the user or an application may be running automatic scans in the background. Automatic scanning for Wi-Fi APs is problematic since the scans may continue to be conducted even when there is no reasonable chance of connecting to detected APs due to the user's speed over the ground, thus consuming energy resources for no real gain. User initiated scanning operations require the user to manually enable the Wi-Fi interface to begin scanning, again without regard to the chances of successfully connecting to a detected network. SONS abstracts from the user the decision making process around when to scan for available Wi-Fi networks, removes the need to manually initiate scans and

it also ensures that automatic scanning only occurs when there is a realistic prospect of a useful connection to Wi-Fi being established.

The architecture of the SONS framework consists of two main components, a Movement Analysis Unit (MAU), the Decision-Making Unit (DMU) and several other modules. A block diagram presented in Figure 20 shows the relationship between the GPS/accelerometer inputs, MAU, DMU, MPD-Cache Monitor, Bus/Train sub-module and the user defined network selection algorithm.

4.5.1 Movement Analysis Unit (MAU)

The Movement Analysis Unit (MAU) takes input from the mobile device's on-board accelerometer and GPS unit to calculate the user's average speed over the ground on continuous basis. Mobile device on-board GPS units typically receive updates once per second and based on this update cycle the MAU samples the user speed once per second. The average speed is calculated once every 10 samples, calculating average speed approximately every 10 seconds allow large fluctuations in speed to be dampened. When the average speed has been calculated and converted to metres per second (mps) the DMU is called and the average speed is passed to it.

4.5.2 Decision Making Unit (DMU)

The decision-making process takes place within the DMU (Figure 20); we present the decision making logic in Figure 21, the decision process flowchart and we present the pseudo code for the decision-making process in Algorithm 1.

The user's average speed as calculated by the MAU, the subscriber's Remaining Data Cap (RDC) and the current battery level act as inputs into the Decision-Making Unit (DMU) for use in calculating the utility scores (5.7, 5.8, 5.9, 5.10). When the DMU has calculated a utility score it sets a "freshness" countdown timer to 5, this countdown timer is decremented by 1 every second until the next utility score calculation. If an external module such as MDF (Chapter 5) or AIS (Chapter 6) requests the latest utility score the remaining "freshness" countdown timer value is passed to it as well as the utility score. The user can set reserve values for both the battery level and RDC and if either of these reserve values are reached the DMU will not calculate the utility scores. The DMU also contains 2 sub-modules, the MPD-Cache Monitor (MCM) and the Bus/Train sub-module (4.5.2).

4.5.3 MPD-Cache Monitor (MCM)

Recall that the MPD-Cache directory is a temporary holding area for downloaded manifest files or MPDs associated with video content prepared for use with DASH enabled media player. The MPD-Cache Monitor checks this directory once per second for changes in directory size, since this directory is empty by default a change in size indicates the presence of a downloaded MPD file. When a change in directory size is detected the MCM decreases the check interval and rechecks the directory size every 0.5 seconds and compares the current directory size with the previous directory size. The download is considered to be completed if the file size does not change for a period equal to 4 rechecks or 2 seconds. When the file has completed downloading the Resolution Discovery Module (RDM) is initialised. The RDM is responsible for analysing the MPD to determine the resolution of the target video and for invoking either MDF (described in Chapter 5) or AIS (described in Chapter 6) as appropriate.

4.5.4 Resolution Discovery Module (RDM)

When the user selects a video to stream the associated MPD file is downloaded to a temporary holding directory named MPD-Cache. SONS monitors this directory for downloaded MPDs by checking the directory size once per second. This directory is empty by default and when SONS detects a change in the directory size it is assumed that an MPD file has been downloaded and the Resolution Discovery Module(RDM) is initialised.

The MPD file is an XML document and the RDM processes the file to determine the associated video's maximum resolution by examining the "height" attribute in the MPDs "Representation" tag. If the maximum value of the height attribute in the file is less than 720p Standard Definition (SD) video is assumed and MDF (Chapter 5) is initialised. On the other hand, if the maximum value of the height attribute is greater than or equal to 720p a high definition (HD) video stream is assumed and the AIS module (Chapter 6) is initialised.

4.5.5 Pedestrian Mode

Commuters in urban environments move from location to location under their own power by walking or cycling, by some mechanical means such as bus or train or by some combination of all these forms of transportation. SONS recognises this through its use of two modes

of operation, pedestrian mode and bus/train mode. Pedestrian mode is the default mode for SONS and is used when a mobile user is travelling at 5 metres per second (the Threshold Speed) or less and bus/train mode is selected when the user is travelling at speeds in excess of 5 metres per second. The pedestrian mode encompasses all commuter travel behaviours at low speeds and while at rest. Bus/train mode addresses those situations in which the commuter might be travelling at speeds in excess of 5 metres per second as indicated by the user device's on-board systems.

4.5.6 Bus-Train Mode

In certain circumstances, it is possible that a mobile user may travel at a relatively high speed and be capable of establishing a connection to a Wi-Fi AP. For example, many public transport systems provide free on-board Wi-Fi for passengers. In this scenario the mobile user is potentially travelling at a high speed yet always remains within the coverage area of the on-board Wi-Fi AP and therefore the Wi-Fi interface should be enabled to facilitate network detection and selection.

The Bus/Train sub-module is a component of the DMU and is activated when the mobile user is found to be travelling at speeds above the threshold value of 5 metres per second. When the Bus/Train sub-module is invoked it activates the Wi-Fi interface and initiates a scan for available Wi-Fi APs. The results of the scan are stored in memory and the Wi-Fi interface is disabled. A 60 second Countdown Timer is started and when it expires the Wi-Fi interface is re-activated and a second scan takes place. The result of the second scan is compared with the stored results from the first scan. If a Wi-Fi AP is seen in the results from both scans the user is considered to be stationary in relation to the AP and the DMU calls the appropriate network selection algorithm. If the same AP is not seen in both scan results the stored result from the first scan is replaced with the result of the second scan. A timer with a duration of 60 seconds is started and when the timer expires the Retry Counter is checked, if the Retry Counter is greater than 0 the Wi-Fi interface is activated and the Retry Counter is decremented by 1. A scan is carried out and the results are again compared with the result held in memory. The process repeats itself until the Retry Counter reaches 0 at which point the Wi-Fi interface is de-activated. In this scenario SONS concludes that no on-board Wi-Fi networks are available and that it is a waste of resources to continue to scan for them.

A default Countdown Timer value of 60 seconds is selected since a vehicle travelling at a speed above the Threshold Speed of 5 mps will have moved beyond the coverage area of a Wi-Fi AP in 60 seconds. Recall that APCmax is 53 m, a vehicle moving at 5 metres per second for 60 seconds will have travelled 300 meters, far outside the average coverage area. A Retry Counter default value of 4 is selected for use since we require at least 2 scans to confirm the presence of an onboard Wi-Fi AP. Wi-Fi scans are not guaranteed to detect an AP on each scan therefore conducting a total of 4 scans is a reasonable compromise between reliability and conducting an excessive number of scans. If an on-board Wi-Fi AP is detected a user speed of 1 metre per second is passed to the DMU and is used in the calculation of the utility score.

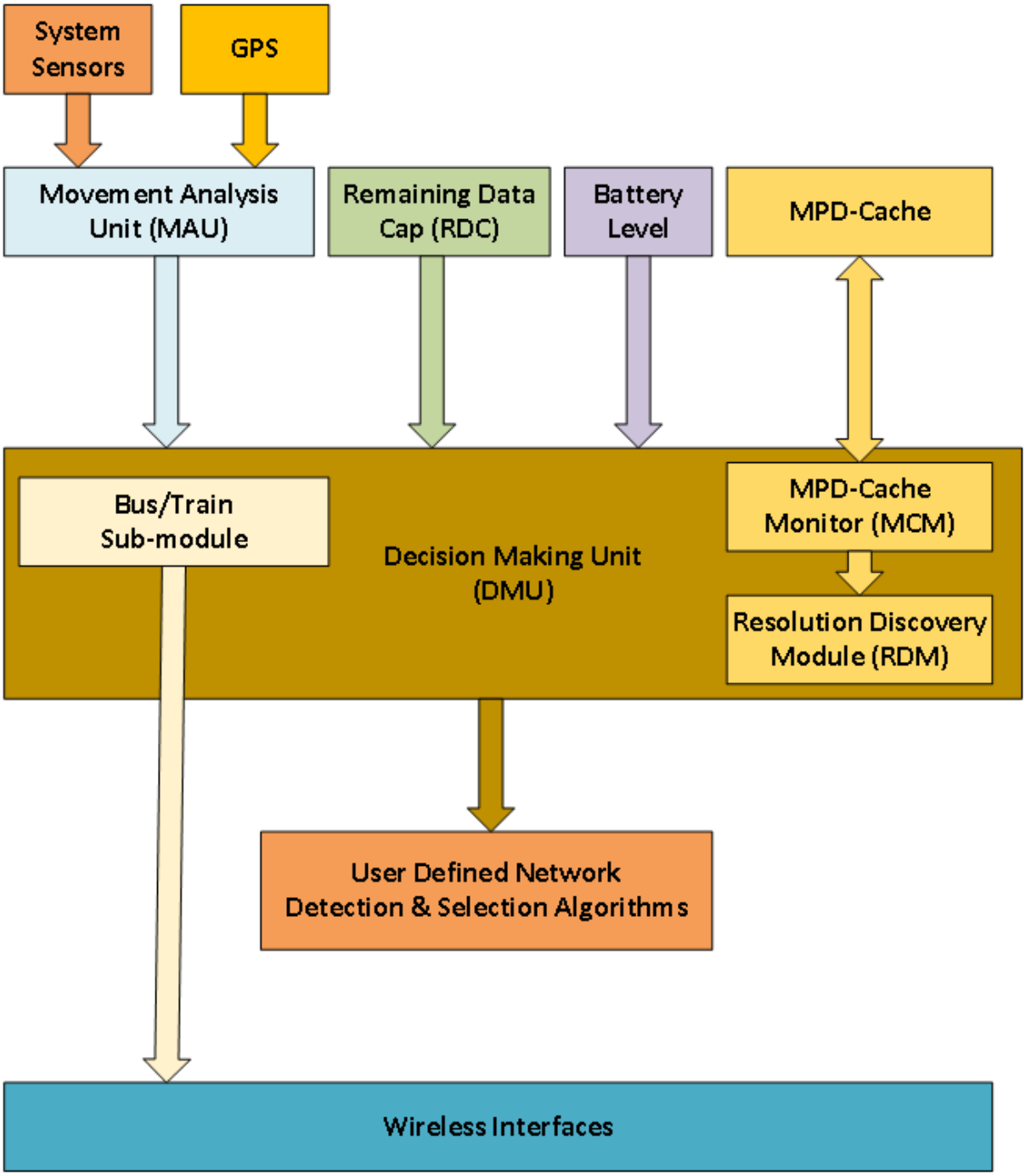

Figure 20 SONS System Component Block Diagram

The decision-making process takes place within the DMU (Figure 20); we present the decision making logic in Figure 21, the decision process flowchart and we present the pseudo code for the decision-making process in Algorithm 1.

The SONS MAU samples the user's speed over ground once per second and calculates the average speed every 10 samples or every 10 seconds. When the average speed has been

calculated the MAU writes the value to a text file for use by the DMU. The DMU checks the speed file every 10 seconds to retrieve the latest average speed and if the speed is greater than the Threshold Speed of 5 metres per second the Bus/Train module in activated. If the Bus/Train module determines that the device is within range of an on-board Wi-Fi AP it returns a user speed of 1 metre per second to the DMU for use in the utility score calculations. When the speed value retrieved by the DMU is less than the Threshold Speed the DMU ignores the Bus/Train module and calculates the current utility score using the available values (4.10). If the DMU calculates a utility score of 1 it activates the Wi-Fi interface and invokes the users preferred network detection and selection algorithm.

The MCM operates within the DMU and monitors the MPD-Cache directory, when it detects the presence of a downloaded MPD file it activates the Resolution Discovery Module (RDM). RDM's role is to analyse the MPD file and to determine whether the target video content is Standard Definition (SD) or High Definition (HD). If the MPD refers to SD content RDM invokes MDF (Chapter 5), on the other hand if RDM determines that the MPD file refers to HD video content it invokes AIS (Chapter 6).

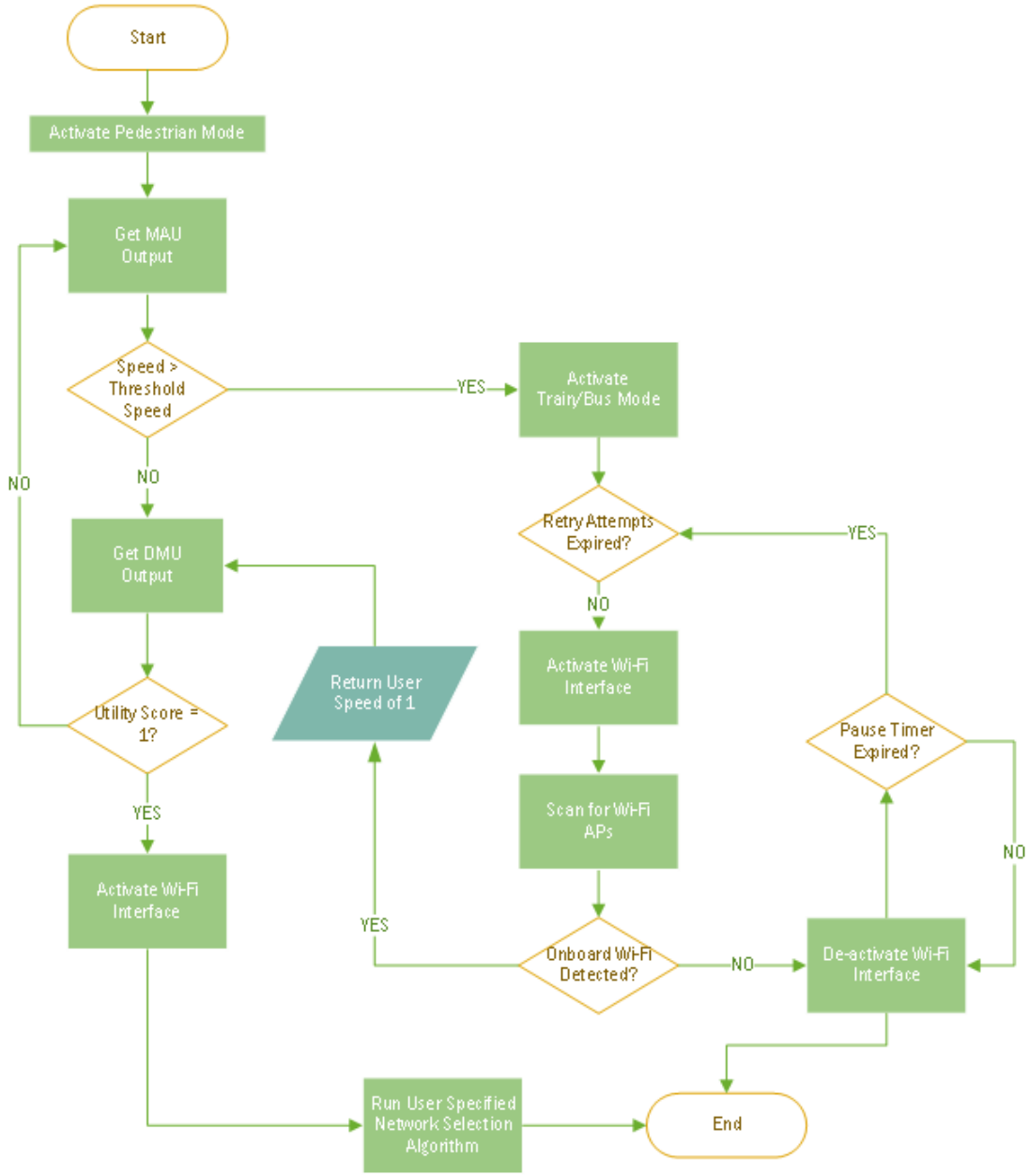

Figure 21 SONS Decision Making Process Flowchart

Algorithm 1 SONS Scan or No-Scan Algorithm

INPUT:

Remaining Data-Cap (RDC);

User speed over ground;

Battery Level;

PROCEDURE:

SONS Movement Analysis Unit (MAU)

Input:

User speed over ground;

Procedure:

Calculate average user speed over ground

Convert average user speed to metres per second

Output:

Metres Per Second;

SONS Decision Making Unit (DMU)

Input:

APC or default APC value;

Metres Per Second;

RDC;

Battery Level;

Procedure:

Decision Making

if *Battery Level <= Reserve Battery Level* **then** *shutdown SONS*

if *RDC Level <= Reserve RDC Level* **then** *shutdown SONS*

if *Metres Per Second > threshold value* **then** *invoke Bus/Train Mode*

else:

Calculate Smin and Smax according to (4.8) and (4.9)

Calculate current utility score based on (4.10)

if U(s) = *0* **then** *return to MAU*

elseif U(s) = *1* **then**

invoke user defined network detection & selection algorithm

Output:

U(s);

BUS/TRAIN Mode

Input:

Metres Per Second

Procedure:

Decision Making

if *number of retry attempts = 0* **then** *get MAU input*

else:

activate Wi-Fi interface

scan for Wi-Fi APs

if *on-board Wi-Fi AP detected* **then**

invoke user defined network detection and selection algorithm

else:

de-activate Wi-Fi interface

wait for pause timer to expire

if *pause timer expires* **then** *check number of retry attempts*

if *number of retry attempts > 0* **then**

activate Wi-Fi interface

scan for Wi-Fi APs

if *on-board Wi-Fi AP detected* **then**

invoke user defined network detection and selection algorithm

elseif *number of retry attempts = 0* **then** *get MAU input*

OUTPUT:

U(s)

Note: The MAU has its own procedure because it runs continuously monitoring the user's speed over the ground. The MAU monitors the user speed once per second as this is a typical update frequency for phone GPS modules and it records the observed speed. Every 5 samples the MAU calculates an average speed and converts the speed to metres per second.

4.6 Modelling and Simulations Overview

The behaviour of the SONS framework was modelled using the utility function described in section 4.3 and two simulation scenarios which will be discussed in sections 4.6.4 and 4.6.5. Simulations were conducted to evaluate SONS and demonstrate its benefits. A two-part approach was taken as follows:

Part I Utility Score Generation – Part I was concerned with generating a set of utility scores using the SONS utility function for various Wi-Fi AP coverage area diameters and for various user speeds. The utility scores are used to indicate when conditions are suitable for attempting to establish useful connections to Wi-Fi APs for mobile users.

Part II NS-3 Simulation – Part II was concerned with constructing a realistic simulation model informed by the real world survey results (Table 4), the Wi-Fi delay factors presented previously in Table 5, observed data transfer rates for Wi-Fi (Table 8) and the cellular network performance metrics discussed in Section 4.6.1 (Table 6 and Table 7). The NS-3 [143] simulation model was used to test the veracity of the SONS utility scores.

4.6.1 Cellular Network Connection Delay

The amount of time needed to establish a useful connection to a Wi-Fi AP using both a laptop and a smartphone at various locations has been examined in Section 4.2.3 Table 5. This section examines the delays experienced in establishing a data connection to a 4G mobile phone network located in the Republic of Ireland for use in the NS-3 [143] simulations. The cellular network of Three Ireland(Hutchison) Limited was selected for use in this work. The company is a telecommunications and internet service provider who operate 3G and 4G/LTE mobile phone services throughout the Republic of Ireland [144]. They launched their first Irish 4G/LTE network in Dublin City in January 2014. Information on the download and upload link speed, connection type, ping times and jitter were gathered using the Speedtest application by Ookla [145]. The Speedtest application is available for various platforms and enables the results of multiple speed-tests to be stored for later examination. Both cellular and Wi-Fi connections can be tested using the application.

The data transfer rates for the Three Ireland mobile phone network are presented in Table 6 and the percentage of time that 4G connections were available is also shown. When 4G service is not available the network falls back to providing 3G services. The results in Table 6

were obtained using the Ookla Speed Test application [145] at various locations in Dublin City. Table 8 presents the examples of the delays experienced while connecting to the data services provided by Three Ireland Ltd, these connection delays were observed while obtaining the results shown in Table 6. Connections were established to various public Wi-Fi networks in Dublin City centre and the download speeds were measured using a connection speed measurement tool [145]. The Wi-Fi speed-test results are shown in Table 8 and the delays experienced in establishing connections to the various Wi-Fi APs were presented previously in Table 5. These results inform the Wi-Fi characteristics used in the NS-3 simulations.

Table 7 Three Ireland 4G/LTE Network Performance

Three Ireland 4G/LTE Data Network Performance				
Type	**Download (Mbps)**	**Upload (Mbps)**	**Ping (ms)**	**Jitter (ms)**
3G	1.88	0.86	41	8
3G	1.34	0.22	47	1
LTE	4.77	2.77	36	3
LTE	6.28	5.98	39	2
LTE	2.67	0.67	21	28
LTE	3.61	1.15	31	7
LTE	1.87	0.86	38	8
LTE	2.69	1.53	31	8
LTE	2.57	1.95	28	6
3G	4.83	0.07	39	11
LTE	3.20	1.64	45	7
LTE	2.34	1.51	35	7
3G	3.49	1.51	40	46
LTE	2.16	0.87	31	15
3G	3.23	2.72	31	15
LTE	1.70	1.86	31	15
LTE	2.41	3.73	27	2
Mean	**3.002353**	**1.758824**	**34.76471**	**11.11765**
Std. Dev	**1.2256**	**1.394725**	**6.575618**	**10.81329**
LTE/4G Availability (%) 70%				

Table 8 4G Connection Delays

Connection Delays to 3 Ireland 4G Data Network [in seconds]					
Inside	7.69	7.28	7.13	6.91	7.11
Outside	7.48	7.28	10.6	7.17	7.35

Table 9 Observed Download/Upload Rates at Public Wi-Fi APs

Dublin City Centre Wi-Fi Performance				
Type	**Download (Mbps)**	**Upload (Mbps)**	**Ping (ms)**	**Jitter (ms)**
802.11n	8.2	3.2	17	6
802.11n	9.4	4.6	39	11
802.11n	7.6	3.12	22	9
802.11n	5.4	3.4	46	4
802.11n	6.8	4.5	29	19
802.11n	4.3	2.1	28	27
802.11n	8.7	2.9	34	16
802.11n	5.4	1.6	41	12
802.11n	4.6	4.3	27	9
802.11n	3.5	2.8	31	7
802.11n	7.3	3.5	40	23
802.11n	8.1	2.1	29	17
802.11n	6.3	3.0	37	39
802.11n	10.2	2.4	46	10
802.11n	7.6	3.6	28	44
802.11n	9.5	4.4	33	26
802.11n	8.9	1.2	25	42
Mean	**7.165**	**3.1**	**32.47**	**18.88**
Std. Dev	**1.92**	**0.98**	**7.9**	**12.43**

4.6.2 Simulation Exercise Overview

The simulation exercises described in this section were carried out to validate SONS's utility function. A set of utility scores were generated for the two different simulation environments described here (Table 9, Table 11). Recall that SONS first generates an upper and lower bound for the utility score and any subsequent utility scores that lie within the upper and lower bound indicate that a useful Wi-Fi connection might be established.

The simulation software selected for use in this work was NS-3 Network Simulator [143], a discrete-event network simulator that is intended primarily for research and educational use. Although many other network simulation packages such as OMNET++ [146] and NetSim [147] are also available NS-3 was selected as it is free software, licensed under the GNU GPLv2 license, and is publicly available for research and development use. It is also updated on a regular basis with new stable versions being released every three months, these releases include new models which are developed, documented, validated and maintained by a large and engaged community. Open validation of the models by third parties is actively encouraged to ensure that the models are of high quality. There is also extensive documentation for the software with many examples and tutorials.

The results presented in this section were generated using a series of NS-3 [143] simulations informed by the following data:

- Real world AP coverage areas (Table 4)
- Real world Wi-Fi scanning and connection delays (Table 5)
- Observed 4G/LTE Download speeds (Table 7)
- Measured 4G/LTE connection delays (Table 8)
- Wi-Fi download speeds (Table 9)

The simulations performed were concerned with models of the real-world data rates for Wi-Fi and LTE. In the simulations the average observed Wi-Fi AP coverage area (Table 4) formed the basis for the node layout for the NS-3 simulations. The observed Wi-Fi download speeds (Table 9) and Wi-Fi scanning and connection establishment delays (Table 5) were used to determine the number of data packets received from each network AP at the mobile node.

The simulation environment consisted of three nodes representing Wi-Fi APs with non-overlapping coverage areas and a single node that represented the LTE network. During each simulation the mobile node travelled in a straight line through the coverage areas of the APs between the 3 Wi-Fi APs and the LTE eNodeB. (Figure 22)

Two simulation environments were created; in the first simulation environment the diameter of the Wi-Fi APs coverage area was 40m (Figure 22) and the coverage pattern of the LTE network was replicated by permitting omni-directional connectivity to the node representing the LTE network. In the second simulation environment (Figure 23) the 3 Wi-Fi APs had coverage areas with a diameter of 80m. The NS-3 simulation parameters used are presented in Table 10.

Table 10 NS-3 Simulation Parameters

NS-3 Simulation Parameters	
Parameter	**Value**
eNodeB Configuration	1 eNodeB, single cell, UL Bandwidth 5 Mbps, DL Bandwidth 5 Mbps
Wi-Fi AP Configuration	3 APs, 1 data source per AP, DL Bandwidth 11 Mbps
UE Configuration	1 UE, speed 0 – 10 metres per second, 3 data sinks (1 per AP), 1 LTE interface
All NS3 Helper classes e.g. Wi-Fi, routing	Default values used

4.6.3 Assumptions

For the purposes of the NS-3 simulations, the following assumptions were made:

- The mobile node can connect to any detected Wi-Fi AP
- Connections to either the Wi-Fi APs or the LTE network are never refused
- Connection delays were constant
- Once a connection is established the average data transfer rate for that connection remains constant, Wi-Fi at a rate of 11 Mbps (mean plus 2 standard deviations Table 9) and LTE at a rate of 5 Mbps (mean plus 2 standard deviations Table 7).
- When not in use, a wireless interface is shut down
- When an interface is brought back up there is no delay in beginning operations apart from an appropriate connection delay (including scanning delay Table 5, Table 8)
- It is assumed that the mobile node passes through the widest part of an APs coverage area

4.6.4 Simulation Environment 1

In the NS-3 Simulation Environment 1 a mobile node moves along a straight line route at a constant speed during each simulation run (Figure 22). Four different communication strategies were modelled:

- A mobile node using LTE only
- A mobile node using Wi-Fi only
- A mobile node using a combination of LTE and Wi-Fi without SONS being implemented
- A mobile node using a combination of LTE and Wi-Fi with SONS implemented

For each of the four communication strategies listed above the mobile node travelled the route a total of 6 times, each time at a different speed. Three independent, non-overlapping 802.11 networks exist within the simulation environment and each Wi-Fi AP has a coverage area 40 m in diameter. The simulations were run over a fixed length route and as the speed of the mobile node over ground increased the duration of the simulations decreased. In each

simulation the mobile nodes starting position was approximately 30 metres from the centre of the coverage area to the left of Access Point A.

For the LTE only simulation the mobile node established a connection to the LTE network when the simulation began and maintained this connection for the duration of the simulation ignoring the Wi-Fi APs completely. In the second simulation the mobile node only connected to Wi-Fi APs and did not connect to the LTE network at any time during the simulation. During the third simulation the mobile node establishes a connection to the LTE network when the simulation begins and scans for Wi-Fi APs whenever it is not connected to one. When the mobile node changes position and moves within range of a Wi-Fi AP it begins the process of establishing a connection to the AP. SONS is not implemented in this simulation and the node attempts to connect to detected Wi-Fi APs regardless of how fast it is moving. While the connection to the AP is being established the mobile node retains its connection to the cellular network and if a connection to an AP is established the connection to the cellular network is dropped. As the mobile node moves out of range of a Wi-Fi AP it experiences a delay of 7 seconds before the LTE connection is established. In the fourth simulation SONS is implemented, the mobile node uses both LTE and Wi-Fi to establish connections but on establishes connections to Wi-Fi when conditions are suitable. During this set of simulations, the mobile node only scans for Wi-Fi APs when the utility score is between s_{min} (4.8) and s_{max} (4.9) and it is not connected to a Wi-Fi AP, thus simulating the functionality provided by the SONS framework. If conditions do not support Wi-Fi communications the node employs LTE only.

Note: As the mobile node moves along a straight line path from left to right through the simulation area it is assumed to pass through the widest part of each of the three APs. In Figure 22 the node's path is shown as being slightly off centre, this is to facilitate the text placement within the AP coverage area.

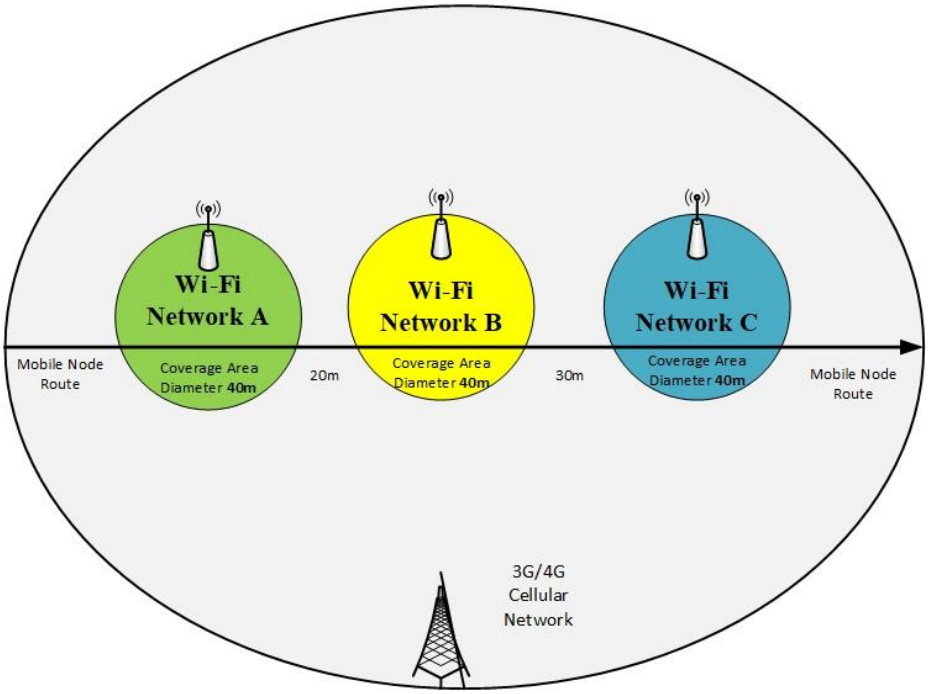

Figure 22 Scenario 1 Simulation Environment, APC 40m

4.6.5 Simulation Environment 2

The four communication strategies simulated in Simulation Environment 1 are also investigated in Simulation Environment 2. In the Simulation Environment 2 simulations the mobile node again moves from left to right along a straight line route through the AP coverage areas in the NS-3 environment travelling at a constant speed during each simulation run. Wi-Fi APs in this environment have coverage areas 80 m in diameter and the mobile node travels the route a total of ten times, each time at a different speed. Again in this set of simulations when SONS is implemented the mobile node only scans for Wi-Fi APs when the utility score is between s_{min} (4.8) and s_{max} (4.9) and the mobile node is not already connected to a Wi-Fi AP.

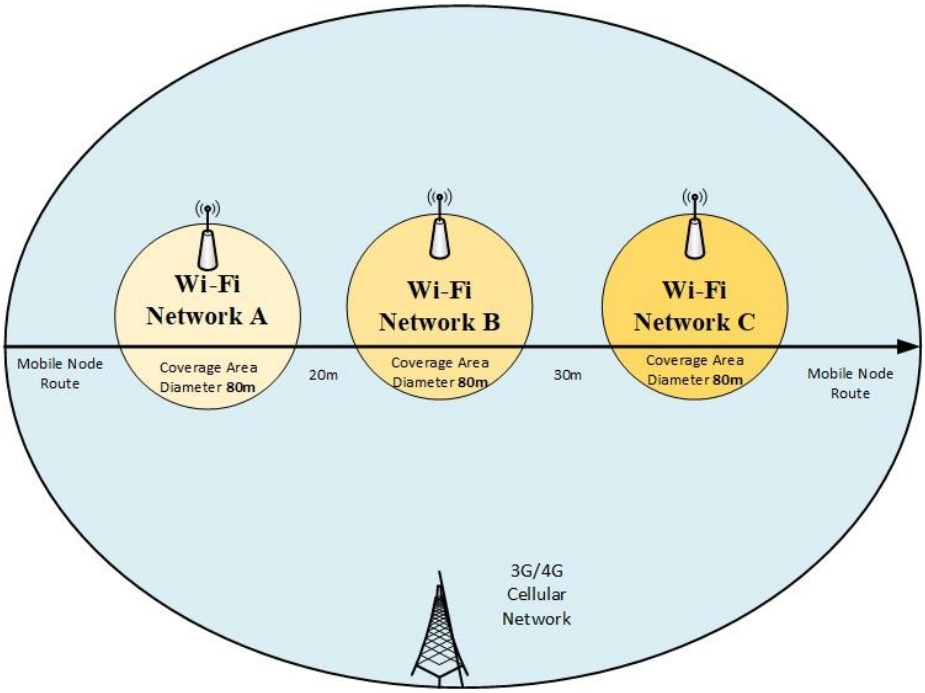

Figure 23 Simulation Environment 2, APC 80m

In each of the simulations conducted the mobile node moved from the left hand side of the simulated environment to the right hand side without stopping. When the mobile node reached the left hand edge of a Wi-Fi AP's coverage area a delay of 8 seconds was implemented before data was received at the mobile node from the AP. A delay of 8 seconds was selected as a reasonable compromise between tdmin (7 seconds) and tdmax (9 seconds). During the delay period the connection to the 4G network, if one existed, was maintained. Once a connection to the Wi-Fi AP was established the 4G data connection was shut down to conserve energy. When the right hand side of the AP's coverage area was reached the transfer of data between the AP and the mobile node was halted immediately.

Following this loss of connectivity to the Wi-Fi AP a 7 second delay in establishing a connection to the 4G network was introduced. This 7 second delay replicates the real-world delay observed during the establishment of a connection to a 4G data network. During the delay caused by the 'break before make' handover from Wi-Fi to cellular no data is transferred between the server and the mobile node. In the event that another Wi-Fi AP was detected and could be connected to before the 4G link was established the new Wi-Fi link was established and the 4G connection attempt aborted.

4.6.6 Result Analysis

Table 11 presents the pre-calculated utility scores for Simulation Environment 1 for the various speeds and RDC levels used in the simulations with the AP coverage area set at a diameter of 40 metres. As the speed of the mobile node over the ground increases the utility score decreases indicating a reduction in the utility to the user of attempting a connection to a Wi-Fi AP. However, we can also see that as the user's data-cap reduces the attractiveness (utility) of connecting to a Wi-Fi AP increases regardless of the mobile node's speed. For example, at a speed of 2 metres per second and an RDC of 1800 (full data allowance) the utility score is 0.0066667 but at the same speed (2 metres per second) with the reserve RDC level of 100 the utility score has increased to 0.2057143. As the mobile user's speed approaches 5 metres per second the utility score becomes very low and at 5 metres per second the utility score is 0. A utility score of 0 indicates that no useful connection can be established to a Wi-Fi AP and therefore no user defined network detection and selection algorithm should be invoked.

Table 11 Calculated Utility Scores for APC of 40m

APC 40	Remaining Data Cap					
Speed	1800	1200	800	400	200	100
1.4	0.011429	0.017143	0.025714	0.051429	0.102857	0.205714
2	0.006667	0.01	0.015	0.03	0.06	0.12
3	0.002963	0.004444	0.006667	0.013333	0.026667	0.053333
4	0.001111	0.001667	0.0025	0.005	0.01	0.02
5	0	0	0	0	0	0
6	0	0	0	0	0	0
U(smax) = 0.46			U(smin) = 0.00244			

The purpose of the NS-3 simulations described in this section is to test whether or not the SONS utility scores are a reliable indicator of when a mobile user should attempt to connect to Wi-Fi APs in their proximity. From Table 11 we see that the lower bound utility score is 0.00244 and the upper bound is 0.46, utility scores falling between these bounds should indicate that conditions are such that a connection to Wi-Fi should be attempted. It is important to note that a utility score greater than the lower bound value is no guarantee that a connection to an AP can be established or that if such a connection is made that a useful

amount of data can be transferred. On the other hand, a utility score that is less than the lower bound but greater than zero does not indicate that a connection to a Wi-Fi AP could not be made rather it indicates a low level of utility for the user in connecting to an AP. For example, at a speed of 4 metres per second and with an RDC of 1800 the utility score is 0.001111 util, this low utility score is a result of the user enjoying a full data allowance and having no current need to protect it. It is only when the utility score is zero that conditions are such that attempting to connect to a Wi-Fi AP is futile.

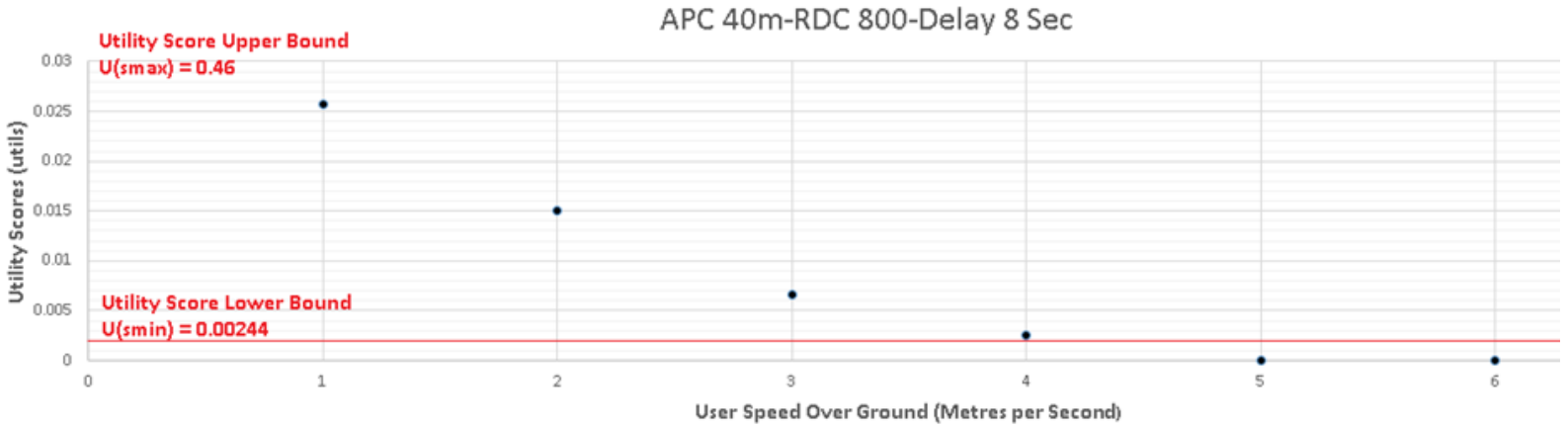

Figure 24 Utility Scores for APC 40m RDC 800 Delay 8 sec

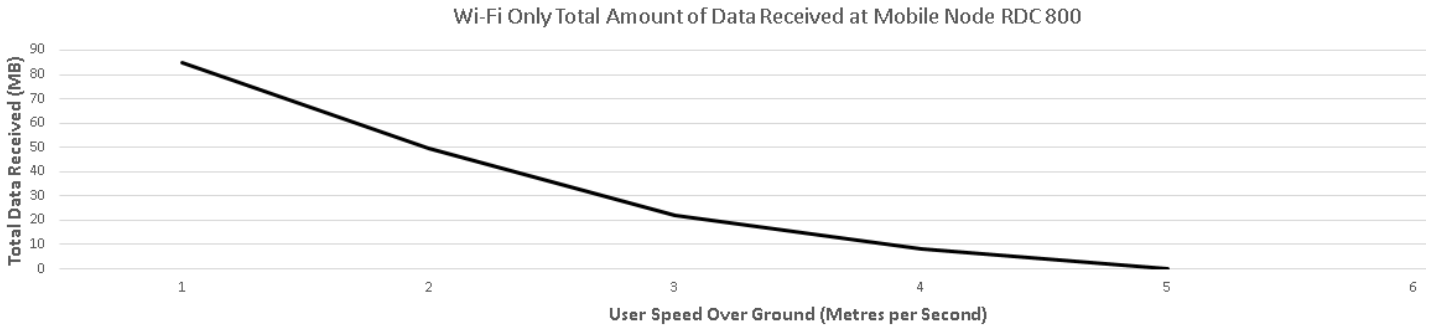

Figure 25 Total Data Received at Node Wi-Fi Only RDC 800

If we examine the case where the user's RDC is 800 we see that the utility scores for speeds up to 4 metres per second are greater that the lower bound value and indicate that a useful connection to Wi-Fi APs might be made. Figure 24 presents a plot of the calculated utility scores including the lower bound which is indicated by a horizontal red line. At speeds up to 4 metres per second the plotted utility scores are at or above the lower bound while at speeds in excess of 4 metres per second the utility scores fall below the lower bound value.

Figure 25 presents a graph of the total amount of data received at the mobile node using Wi-Fi only. In this set of simulations, a very simple handover decision process is implemented. The mobile node travels a straight line path from the left hand side of the simulation area to the right hand side passing through the centre of the Wi-Fi APs coverage areas. When the node reaches the left hand side of an AP coverage area it begins the connection establishment process, and as soon as it reaches the right hand side of the coverage area the Wi-Fi connection is dropped. The mobile node automatically connects to each AP as it encounters them and signal strength is not considered for the purpose of this work. As the node speed over the ground increases the overall amount of data received decreases, this is due to the reduction in the amount of time that the mobile node remained in range of the AP. However, a comparison between Figure 24 and Figure 25 clearly shows that when the utility score was 0 (Figure 24, Table 9) the amount of data received using Wi-Fi only was 0 (Figure 25).

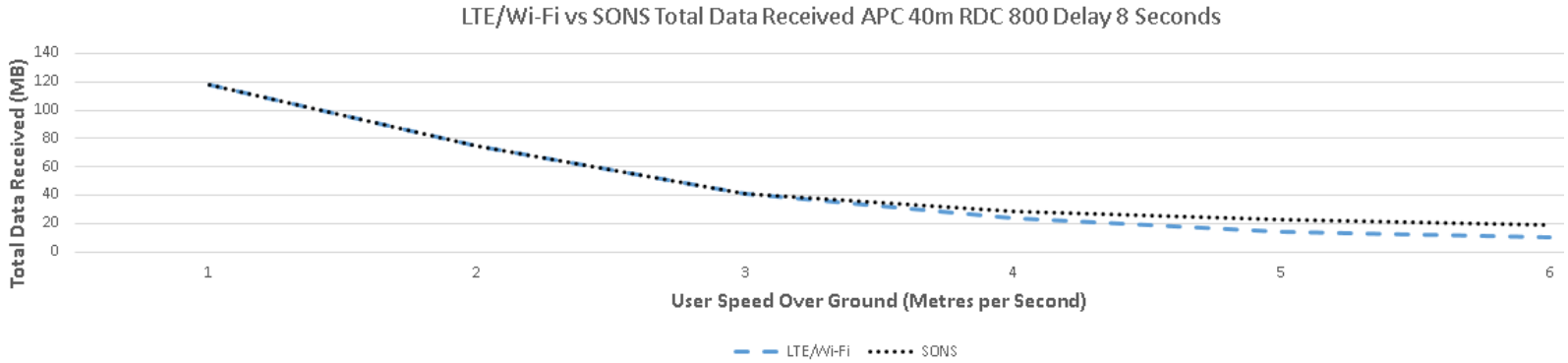

Figure 26 Total Data Received at Mobile Node LTE/Wi-Fi vs SONS APC 40m RDC 800

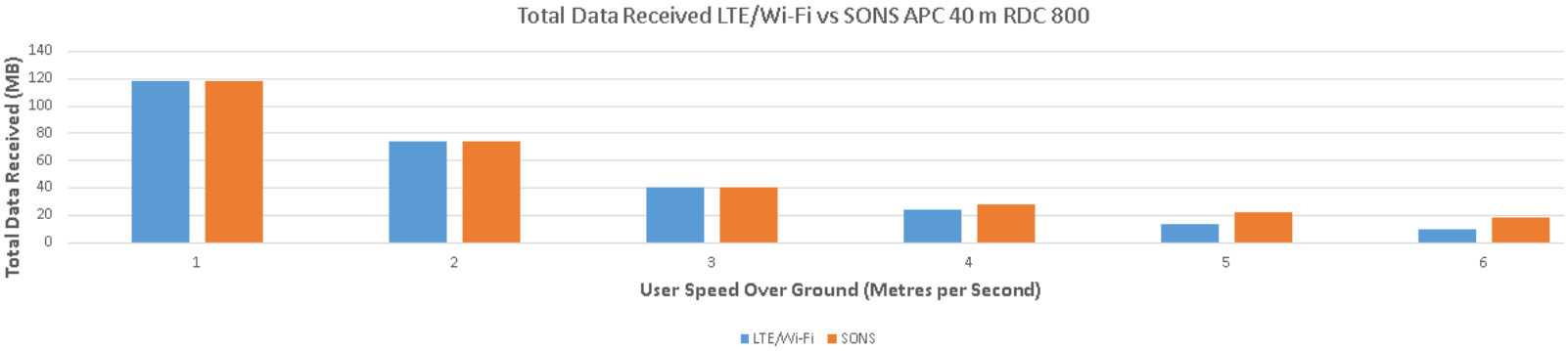

Figure 27 Total Data Received APC 40 RDC 800 LTE/Wi-Fi vs SONS

Figures 26 and 27 present the total amount of data received at the mobile node while employing LTE/Wi-Fi without implementing SONS and the total amount of data received when implementing SONS. The figures show that both strategies result in the mobile node receiving the same amount of data at low speeds (up to approximately 3 metres per hour) but that at higher speeds a greater amount of data is received when SONS is implemented. This is due to the reduction in the total amount of time the mobile node is disconnected from the cellular network during the hand over from Wi-Fi to the cellular network. At the lower speeds the total disconnection period was 14 seconds for each strategy, however at the higher speeds implementing SONS resulted in no periods of disconnection while not implementing SONS resulted in being disconnected for a total of 14 seconds. When SONS was implemented the utility scores calculated at speeds greater than 4 metres per second were zero, as a result no attempts were made to establish connections to the Wi-Fi APs and therefore no handovers from Wi-Fi to cellular occurred.

Table 12 Total Amount of Data Transferred APC 40m and RDC of 800

APC 40 m & RDC 800					
Speed (mps)	**LTE/Wi-Fi No SONS (MB)**	**Percentage of time disconnected (%)**	**LTE/Wi-Fi SONS (MB)**	**Percentage of time disconnected (%)**	**Utility Score**
1.4	118	11%	118	11%	0.025714
2	74.5	15.6%	74.5	15.6%	0.015
3	40.74	22.3%	40.47	23.3%	0.006667
4	23.88	31%	23.88	31%	0.0025
5	13.75	39%	22.5	0%	0
6	10	46.7%	18.75	0%	0

Table 12 presents the total amount of data received by the mobile node when using a combination of LTE and Wi-Fi connections both with and without implementing SONS. While the SONS utility score was greater than the calculated lower bound of 0.00244 and less than the upper bound of 0.46 the total amount of data received at the mobile node was the same

whether or not SONS was implemented. This is a result of the patterns of connecting and disconnecting to and from Wi-Fi APs and the cellular network being the same in both use cases. However, when the utility score was equal to zero a greater amount of data was received at the mobile node when SONS was implemented due to a reduction in the amount of time the node was in a disconnected state. For example, at a speed over the ground of 5 metres per second the node which did not implement SONS experienced a disconnected state 39% of the time whereas the node which implemented SONS was never in a disconnected state (Table 11). Again at a speed of 6 metres per second over the ground the node which did not implement SONS was disconnected 46.7% of the time while the node that did implement SONS was never in a disconnected state.

Table 13 Calculated Utility Scores for APC of 80m

APC 80	Remaining Data Cap					
Speed	1800	1200	800	400	200	100
1.4	0.027301587	0.040952381	0.061429	0.122857	0.245714	0.491429
2	0.017777778	0.026666667	0.04	0.08	0.16	0.32
3	0.01037037	0.015555556	0.023333	0.046667	0.093333	0.186667
4	0.006666667	0.01	0.015	0.03	0.06	0.12
5	0.004444444	0.006666667	0.01	0.02	0.04	0.08
6	0.002962963	0.004444444	0.006667	0.013333	0.026667	0.053333
7	0.001904762	0.002857143	0.004286	0.008571	0.017143	0.034286
8	0.001111111	0.001666667	0.0025	0.005	0.01	0.02
9	0.000493827	0.000740741	0.001111	0.002222	0.004444	0.008889
10	0	0	0	0	0	0
U(smax) = 0.46			U(smin) = 0.00244			

Table 13 presents the pre-calculated utility scores for Simulation Environment 2 for the various speeds and RDC levels used in the simulations with the AP coverage area set at a diameter of 80 metres. Again, as the speed of the mobile node over the ground increases the utility score decreases indicating a reduction in the utility of attempting a connection to a Wi-Fi AP. However, we can also see that as the user's data-cap reduces the attractiveness (utility) of connecting to a Wi-Fi AP increases for any particular speed. For example, at a speed of 2 metres per second and an RDC of 1800 (approximately full data allowance) the utility score is 0.0169133 but at the same speed with the reserve RDC level of 100 the utility score has increased to 0.32. As the mobile user's speed approaches 10 metres per second the utility score becomes very low and at 10 mps the utility score is 0. A utility score of 0 indicates that

no useful connection can be established to a Wi-Fi AP and therefore no user defined network detection and selection algorithm should be invoked.

Table 14 Total Data Transfer APC 80 m, RDC 800

APC 80m & RDC 800					
Speed (mps)	**LTE/Wi-Fi No SONS (MB)**	**Percentage of time discon-nected (%)**	**LTE/Wi-Fi SONS (MB)**	**Percentage of time discon-nected (%)**	**Utility Score**
1.4	235.7	5.32%	235.7	5.32%	0.061429
2	157	9.33%	157	9.33%	0.04
3	95.75	14%	95.75	14%	0.023333
4	65.15	18.67%	65.15	18.67%	0.015
5	46.75	23.3%	46.75	23.3%	0.01
6	40.72	28%	40.72	28%	0.006667
7	25.76	32.66%	25.76	32.66%	0.004286
8	19.19	37.3%	19.19	37.3%	0.0025
9	14.12	42%	21	0%	0.001111
10	9.38	46.7%	18.75	0%	0

Table 14 presents the amount of data received at the mobile node using a combination of LTE and Wi-Fi connections with SONS being both implemented and not implemented. Also shown is the percentage of time that the mobile node was in a disconnected state for various utility scores when the user's RDC was at 800 and the APs had a coverage area 80 metres in diameter.

The total amount of data received at the mobile node and the amount of time that the node was in a disconnected state were the same regardless of whether or not SONS was implemented for speeds under 9 metres per second. At speeds under 9 metres per second the calculated utility score was greater than the pre-calculated lower bound utility score of 0.00244 and connections to Wi-Fi APs were established as was the case in which SONS was not implemented. When the mobile node was travelling at 9 metres per second the utility score was less than the lower bound and in the case in which SONS was implemented no connec-

tions to Wi-Fi APs were attempted and as a result there were no disconnections during hand-overs from Wi-Fi to LTE. This resulted in 50% more data being received at the mobile node than during the equivalent simulation in which SONS was not implemented. The same behaviour is seen when SONS is implemented and the mobile node is travelling at 10 metres per second. A utility score of zero is calculated, no connections to Wi-Fi are made and no disconnections were experienced. In this case there was an increase in the total amount of data received at the mobile node from 9.38 MB to 18.75MB.

When SONS was not implemented and the mobile node was moving at a speed of 10 metres per second the mobile node connected to available Wi-Fi APs. However, there was no transfer of data as the dwell time was equal to the connection delay of 8 seconds, by the time a connection was established the mobile node was already leaving the APs coverage area. This situation resulted in the mobile node being disconnected from the cellular network for 46.7% of the simulations duration.

SONS determines on behalf of the user when conditions are such that attempts to connected to detected Wi-Fi APs might be successful. It does not directly protect the user's data-cap but activates, when the time is appropriate, mechanisms such as MDF and AIS that have the capacity to do so. By invoking MDF or AIS to offload data whenever possible from the user's cellular connection to an available Wi-Fi connection SONS indirectly protects the user's data-cap.

4.7 Energy Consumption by Wi-Fi Scanning Operations

All networking operations consume some amount of energy and Table 15 presents the estimated energy consumption for Wi-Fi scanning operations during the Simulation Environment 1 operations.

Table 15 Energy Consumption Per Wi-Fi Scan (mWatts) APC 40m

Energy Consumption per Wi-Fi Scan (mWatts) APC 40						
Node Speed	Number of Scans (No-SONS)	Scan Energy (150mW per scan)	Number of Scans (SONS)	Scan Energy (150mW per scan)	Energy Saving (mW)	SONS Utility Score
1.4mps	14	2100	14	2100	0	0.025714
2mps	11	1650	11	1650	0	0.015
3mps	10	1500	10	1500	0	0.006667
4mps	9	1350	9	1350	0	0.0025
5mps	8	1200	0	0	1200	0
6mps	6	900	0	0	900	0

In simulations in which SONS was not implemented the mobile node scanned for Wi-Fi networks whenever it was not connected to an AP. When the SONS framework is implemented, there is a significant reduction in scanning activity since the node will only scan for Wi-Fi networks when the SONS utility score falls between the upper and lower bounds. This leads to a reduction in energy consumption, for mobile devices in the simulations in which SONS was implemented this can result in a saving of 10,800mW per hour assuming a scanning frequency of 5 seconds.

4.8 Conclusions

Ubiquitous wireless networks and multi-homed mobile devices make it possible for a mobile user to be always “best connected” and to consume multimedia content on-the-go. Opportunities may exist to seek out and connect to a ‘better’ network than the current connection, but this behaviour consumes energy and mobile users constrained by dependence on a battery must protect their battery resources where possible. It has also been demonstrated that

it is not always appropriate to engage in network detection and selection since, under certain circumstances, this can reduce the total amount of data transferred. Additionally, initiating wireless operations when they are of no benefit, results in the unnecessary consumption of energy for no gain.

This section presents the SONS framework which abstracts from the user the decision of whether or not to carryout network detection and selection operations. Implementing the SONS framework has clear benefits for the end-user as follows:

- It removes from the user the need to make decisions on when to invoke Wi-Fi detection and selection algorithms
- If SONS returns a utility score of 0 no attempt is made to connect to a detected Wi-Fi AP thereby reducing the number of disruptions to data transfer operations due to handoffs
- The amount of time lost to connection delays during handovers from Wi-Fi to cellular networks is reduced
- The reduction in scanning activities reduces the amount of energy consumed, not activating Wi-Fi interfaces unless there is a reasonable chance of establishing a useful connection also helps conserve energy (Table 15)

Testing performed demonstrates that SONS-driven management of network discovery operations can help improve data transfers, reduce energy consumption and protect the user experience by reducing the number of unnecessary handovers.

The improvement in the amount of data received at the mobile node depends on the AP coverage area and the speed at which the mobile node is traveling. For example, in Simulation Environment 1 the APs had a coverage area 40 metres in diameter and the mobile node travelled at various speeds ranging from 1.4 metres per second to 6 metres per second. In these simulations a mixture of LTE and Wi-Fi connections were used and the results from simulation runs in which SONS was not implemented were compared with the results from simulation runs in which SONS was implemented. At a mobile speed of 5 metres per second there was an increase of 63.6% in the amount of data received from 13.15 MB (SONS not implemented) to 22.5 MB (SONS implemented). When the mobile node was traveling at 6 metres per second the increase in total amount of data received at the mobile node was 87.5% from 10 MB (SONS not implemented) to 18.75 MB (SONS implemented).

In Simulation Environment 2 the AP coverage areas had a diameter of 80 metres, when the mobile node was travelling at 9 metres per second there was an increase of 48.7% in the amount of data received at the node up from a total of 14.12 MB (SONS not implemented) to 21 MB (SONS implemented). At a node speed of 10 metres per second the increase was 99.6% from 9.38 MB (SONS not implemented) up to 18.75 MB (SONS implemented).

Energy consumption is decreased by shutting down interfaces when they are not in use. In the simulations carried out when SONS was not implemented the mobile node activated the Wi-Fi interface for scanning operations whenever the mobile node was not connected to an AP leading to unnecessary energy consumption. When SONS was implemented a SONS utility score of zero indicated that establishing a connection to a Wi-Fi AP was not feasible, in this case the Wi-Fi interface was disabled and scanning operations were curtailed. By not activating the Wi-Fi interface when the utility score was zero a total of 1200 mWatts in energy was saved over the course of the simulation when the node was moving at 5 metres per second and 900 mWatts was saved at a speed of 6 metres per second.

The number of handovers was also reduced when SONS was implemented, for each Wi-Fi AP emulated in the simulated environment a handover was required when the node moved into range of the AP and also when it moved out of range of the AP.

SONS fulfils Thesis Objective 1 introduced in Chapter 1, Section 1.6 and reproduced below:

1. To develop a process to decide when to initiate Wi-Fi scanning operations that takes into consideration a user's remaining data-cap, AP coverage areas and the user's speed over the ground

The SONS utility function determines, based on user speed over the ground, the user's remaining data-cap and the AP coverage area when it appropriate to attempt to connect to a Wi-Fi AP and when it is not. Output from the utility function, the utility score, was verified through modelling and simulation.

CHAPTER 5 MPEG-DASH-BASED FRAMEWORK (MDF)

In the previous chapter, the SONS framework and its utility function were introduced. This chapter presents MDF which is initialised by SONS when a SD video stream is detected. MDF is described in detail and its operation verified through a series of simulations. The results of the simulations are presented and discussed.

5.1 Motivation

Subscribers who use prepaid mobile network data plans with limited data allowances wish to protect the data-cap imposed on them by their mobile network provider. In order to do so they will switch their point of attachment from their cellular network connection to alternative wireless networks such as Wi-Fi whenever possible. However, mobile users who stream video content also wish to achieve the best Quality of Experience (QoE) that the prevailing conditions permit. Mobile users frequently stream Standard Definition (SD) video content originally recorded on devices such as smartphones, tablets, etc. that is not of the highest quality. In this context maintaining or improving end user QoE is achieved through reducing stalling events to a minimum since picture quality itself might be compromised. This chapter introduces an innovative MPEG-DASH-based Framework (MDF) for improving end-user video experience in heterogeneous multi-network wireless environments by reducing the number of stalling events. The SONS framework (Chapter 4) determines when a device should invoke user specified network detection and selection algorithms and also initialises MDF when appropriate.

Unlike MDF, other schemes which seek to enhance QoE for mobile users do not differentiate between SD and HD video streams and as a result apply complicated solutions to SD video streams that do not require them. Zhang et al. in [148] propose a scheme for enhancing QoE that uses an environment-aware QoE model. The proposed scheme targets the provision of differentiated HTTP Adaptive Streaming (HAS) services under different environments and the improvement of the QoE associated with HAS for mobile users.

5.2 Overview

This section describes MDF, a generic technology agnostic framework that seeks to optimise the performance of MPEG-DASH enabled clients streaming Standard Definition (SD) video in urban HetNet environments. MDF employs the SONS mechanism described in Chapter 4 to determine when conditions are suitable for a user to attempt to switch their point of attachment from a cellular network to an alternative Wi-Fi networks. Attempting to connect to alternative networks only when a reasonable possibility of establishing a useful connection exists reduces disruption to connectivity and also improves QoE by reducing stalling events. MDF also matches the requested video segments with both connection type and device capabilities (e.g. screen size) to avoid downloading segments of a higher definition than can be utilised and to protect the user's data-cap. The data-cap is protected by downloading as little data over the cellular connection as is practicable.

Figure 28 illustrates the major architectural components of MDF, its interaction with the Scan-or-Not-to-Scan (SONS) module described in Chapter 4. The main components of MDF are the Manifest Manager (MM) used to process the MPD file, the Decision Implementation Unit (DIU) that downloads video segments as directed by the Decision Making Unit (DMU), the Screen component that determines the devices screen size, a Buffer Monitor that checks the playout buffer level and the Interface Monitor that tracks the states of the wireless interfaces.

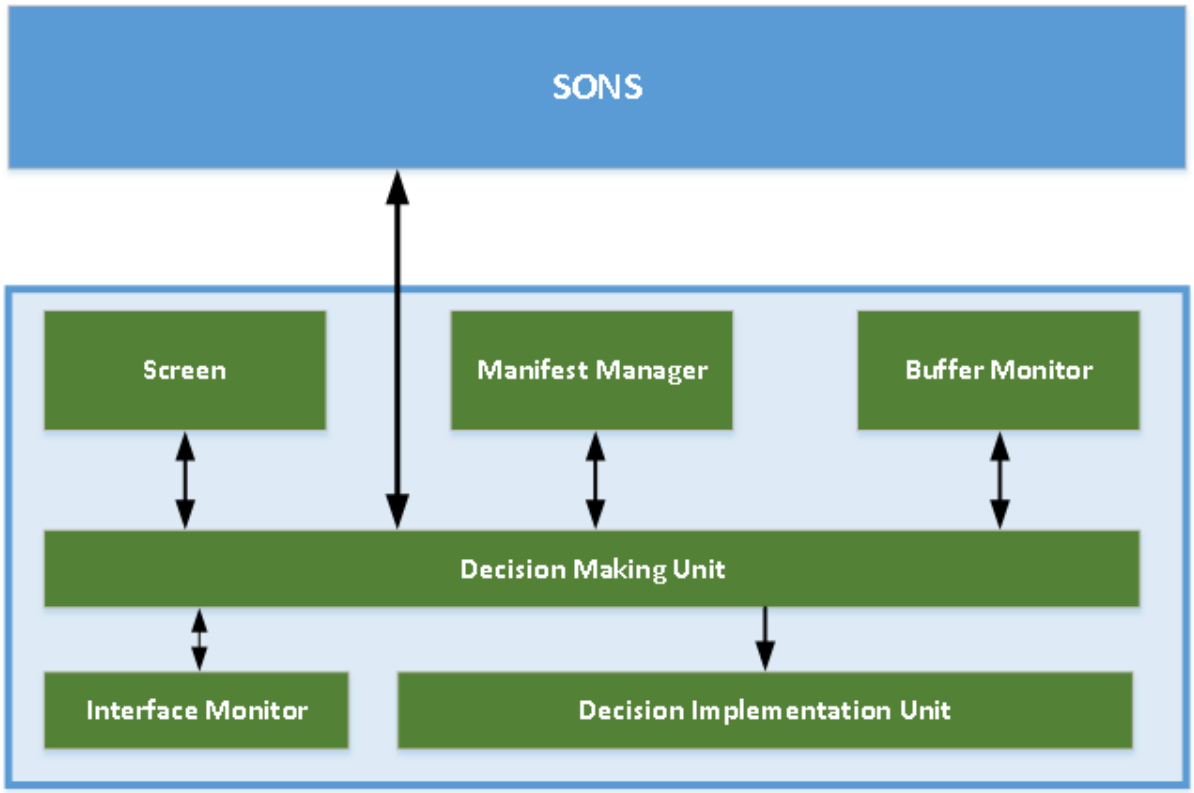

Figure 28 MDF Block Level Architecture

On initialisation MDF carries out the following three actions:

1. MDF's Decision Making Unit (DMU) polls the SONS framework for the current utility score s and the freshness countdown timer value
2. The DMU requests the devices screen size from the Screen component
3. The DMU instructs the Manifest Manager to process the MPD file currently held in the MPD-Cache directory

1) DMU polls SONS for the current utility score - When SONS calculates the utility score a "freshness" timer is set to 5 seconds and is decremented by 1 for every second that elapses, and when the timer reaches 0 the utility score is recalculated. The freshness timer value indicates the period of time in seconds until the next utility score is calculated. MDF takes the freshness counter value and uses it to synchronise its utility score requests with the utility value calculations in order to ensure that it has the most recent utility score. To reduce the initial delay if the freshness timer value is greater than 3 MDF uses the current utility score and sets its update time to match

the freshness timer. If the freshness timer value is less than 3 MDF sets its update timer to match the freshness value and waits until the utility score is updated.

2) In addition to requesting the utility value from SONS MDF also requests the devices screen size from the Screen sub-module, the returned screen size will be used to determine the maximum segment bitrate that MDF will request. The resolution of the requested segments will match as closely as possible the reported device screen resolution.

3) Once the DMU has decided on the video segment bitrate to use based on screen size the information is passed to the Manifest Manager which processes the MPD file in the MPD-Cache directory to generate a list of required video segments. Video segments identified in the MPD as having a two second duration are selected for use. The segment duration of 2 seconds was chosen as a good compromise between encoding efficiency and flexibility based on the work of [31]. In the event that segments of 2 second duration are not available those segments having a duration as close to 2 seconds as possible are selected. The segment list is then returned to the DMU and the MPD file is deleted from the MPD-Cache directory.

The DMU polls the Buffer Monitor to check on the amount of data that is currently being held in the playout buffer. MDF seeks to maintain a stable playout buffer level whenever possible to reduce the possibility of stalling events occurring and the buffer level to be maintained depends on the connection type currently in use. DMU aims to keep the playout buffer as full as possible when connected to Wi-Fi and as low as possible when connected to cellular.

When the SONS utility score is 1 the users preferred network detection and selection algorithms are invoked. If a connection to a Wi-Fi AP is successfully established the playout buffer level is changed to a target level of 8 seconds before the cellular data connection is dropped. The time taken to establish a data connection to a cellular network has been shown to be 8 seconds and in order to be able to bridge the connection delay to cellular the playout buffer level should be maintained at a level of at least 8 seconds. If the level of the playout

buffer is less than that required, the DMU will alert the DIU to download multiple segments from the list to bring the buffer up to the necessary level.

A SONS utility score of 0 indicates that attempting to establish connections to Wi-Fi is not feasible. In this case, since there will be no hard handover from Wi-Fi to cellular the DMU will attempt to maintain a playout buffer level equivalent to the duration of the selected segments, which by default is 2 seconds. Previous work [128] has shown that when the play-out buffer level drops below 1 second stalling events occur more frequently; maintaining a buffer level of 2 seconds assists in reducing the number of stalling events. As there will be no periods of time during which the node will be disconnected due to "break before make" handovers from Wi-Fi to cellular there is no requirement to be able to bridge gaps in connectivity which removes the need for a full playout buffer. The user also wishes to download as little as possible over the cellular link and maintaining low buffer levels in anticipation of establishing a Wi-Fi connection at a later time aids in this objective.

The DMU uses the Interface Monitor to check the status of the device's wireless interfaces. If the Wi-Fi interface changes status from active to shut-down the DMU will alter its playout buffer level requirements from high to low. If, on the other hand, the Wi-Fi interface changes state to up then the DMU will respond by changing the buffer level requirement from low to high. This strategy is adopted to manage scenarios in which the Wi-Fi network is available but the Wi-Fi interface on the device becomes unavailable through malfunctioning or through manual user intervention.

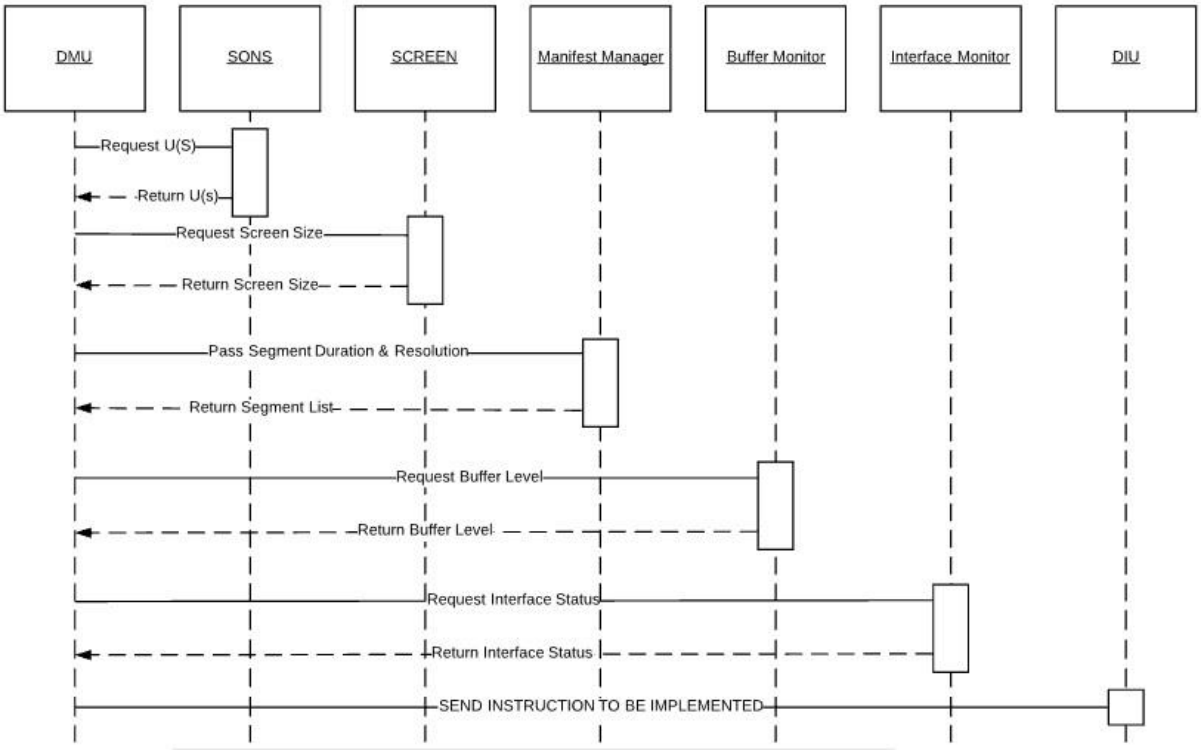

Figure 29 MDF Instruction Sequence

Identifying the screen size of the device on which the DASH client is running is important since it makes no sense to request a high bit rate segment to display on a small sized screen. [149] demonstrated that end users are satisfied with low bit rate segments on smartphones due to the small screen size. Requesting lower bit rate segments makes downloads faster, reduces the load on the access networks and helps reduce the amount of space needed for buffering. It also helps protect the end user's data caps by minimizing the total amount of data downloaded when a user is connected to a cellular network.

The aims of MDF are as follows:

1) Reduce the amount of data downloaded over a cellular connection to protect end user data-cap
2) Reduce the number of stalling events during video streaming to maintain or improve end-user QoE
3) When connected to Wi-Fi networks maintain a playout buffer level greater than 8 seconds to bridge loss of connectivity due to handovers from Wi-Fi to cellular networks and when connected to cellular networks maintain a buffer level equivalent to the selected segment duration e.g. 2 seconds

5.3 Testing the MDF Framework

The MDF test environment consisted of a Lenovo ThinkPad laptop equipped with an Intel i7 processor, 16GB of RAM and an SSD drive running Debian 9 GNU/Linux. MDF's concepts were tested using the NS-3 [143] simulation package and Linux LXC containers [150]. Two Debian based LXC containers were created as well as a set of tap/bridge devices to enable the external containers to send traffic to and from the NS-3 network simulation. One of the Debian containers ran a modified version of VLC [151] and acted as a streaming server while the other container hosted a VLC client (see Figure 30). The server hosted a copy of the EnvivioDash3 video content hosted at [152]. The duration of each simulation was 193 seconds, a period of time equal to the playing time of the video content.

Note: Quality of Experience is a subject metric typically measured using panels of human assessors which are expensive to convene and this need for human assessors makes it impossible to measure QoE in real-time for streamed video content. It has been demonstrated that the number of stalling events that occur during the streaming of a video is closely related to the users QoE. In this work the number of stalling events is used as a metric for measuring QoE, a high number of stalling events indicates a low QoE while a lower number of stalling events indicates a higher QoE.

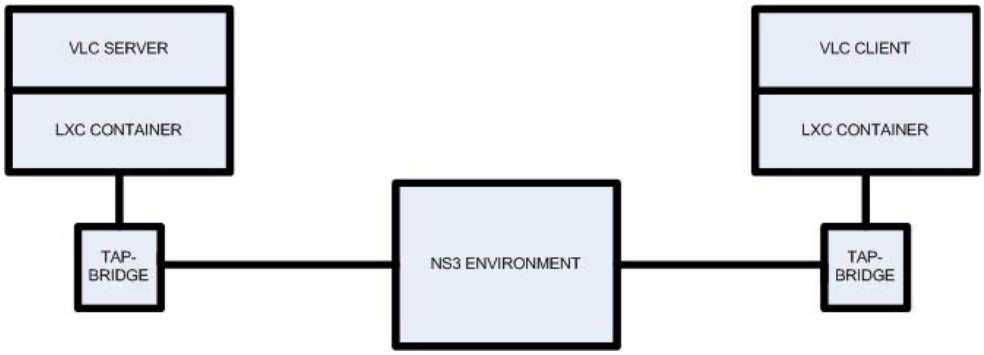

Figure 30 Overview of the test environment

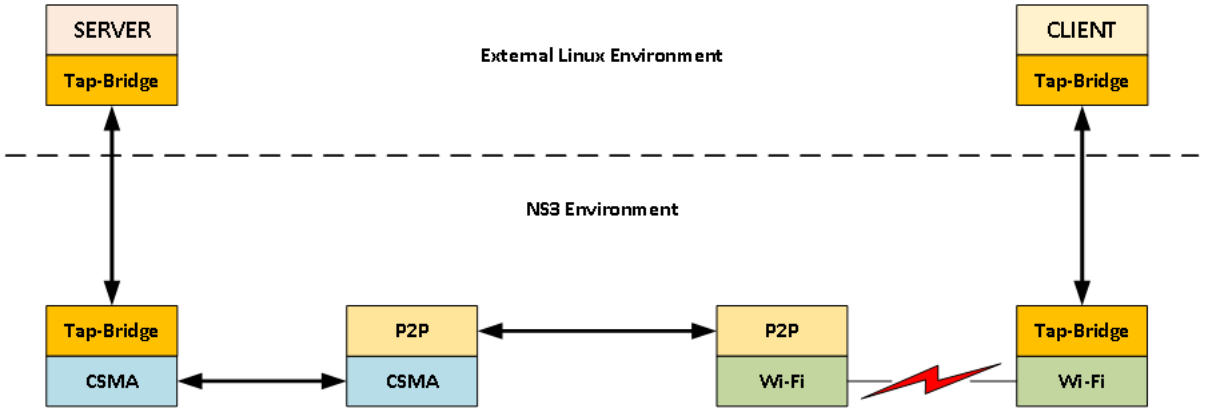

Figure 31 NS-3 Internal Network Structure

The following four scenarios were investigated using the test environment described previously:

1. User in a stationary position (0 metres per second)
2. Mobile user travelling at 1.4 metres per second, the average pedestrian speed in an urban environment
3. Mobile user travelling at 5 metres per second, the average speed of public transport in urban centres
4. Mobile user travelling at 10 metres per second, the legal speed limit in Dublin City Centre

For each of the 4 scenarios under investigation the simulations were conducted in two parts and the results compared. In Part 1 of each simulation the mobile host operated without MDF and in Part 2 of the simulation MDF's behaviours were implemented. During Part 1 of a simulation the mobile user would always attempt to connect to alternative networks when

and if such a network was detected. In Part 2 of each simulation the mobile user would attempt to connect to an alternative network only if the SONS module (Chapter 4) deemed that a useful connection could be established.

For the purposes of the MDF simulations the following assumptions were made:

- The user equipment (UE) is assumed to be a smartphone
- Initially the UE has an established connection to the cellular network
- The mobile node would establish a connection to a Wi-Fi AP as soon as it was in range
- Connection attempts to both Wi-Fi and cellular networks were never refused
- Connection delays were constant, Wi-Fi had an 8 second delay, LTE had a 7 second delay
- No data traffic was dropped and a different constant bit rate (CBR) was employed for both Wi-Fi and cellular links, Wi-Fi 11 Mbps, LTE 5 Mbps
- Wireless interfaces were deactivated when connections were dropped
- It is assumed that a mobile node travels through the widest part of an APs coverage area

The simulated wireless HetNet environment consisted of 3 Wi-Fi APs and an LTE eNodeB as shown in Figure 16 below. The coverage area of each of the APs was 80 metres in diameter and the coverage areas were non-overlapping. The simulation environment parameters are presented in Table 16.

Table 16 NS-3 Simulation Environment Parameters

NS-3 Simulation Parameters	
Parameter	**Value**
eNodeB Configuration	1 eNodeB, single cell, UL Throughput 5 Mbps, DL Throughput 5 Mbps
Wi-Fi AP Configuration	3 APs, 1 data source per AP, DL Through-put 11 Mbps
UE Configuration	1 UE, speed 0 – 10 metres per second, 3 data sinks (1 per AP), 1 LTE interface
All NS3 helpers e.g. wifiHelper, etc.	Defaults used

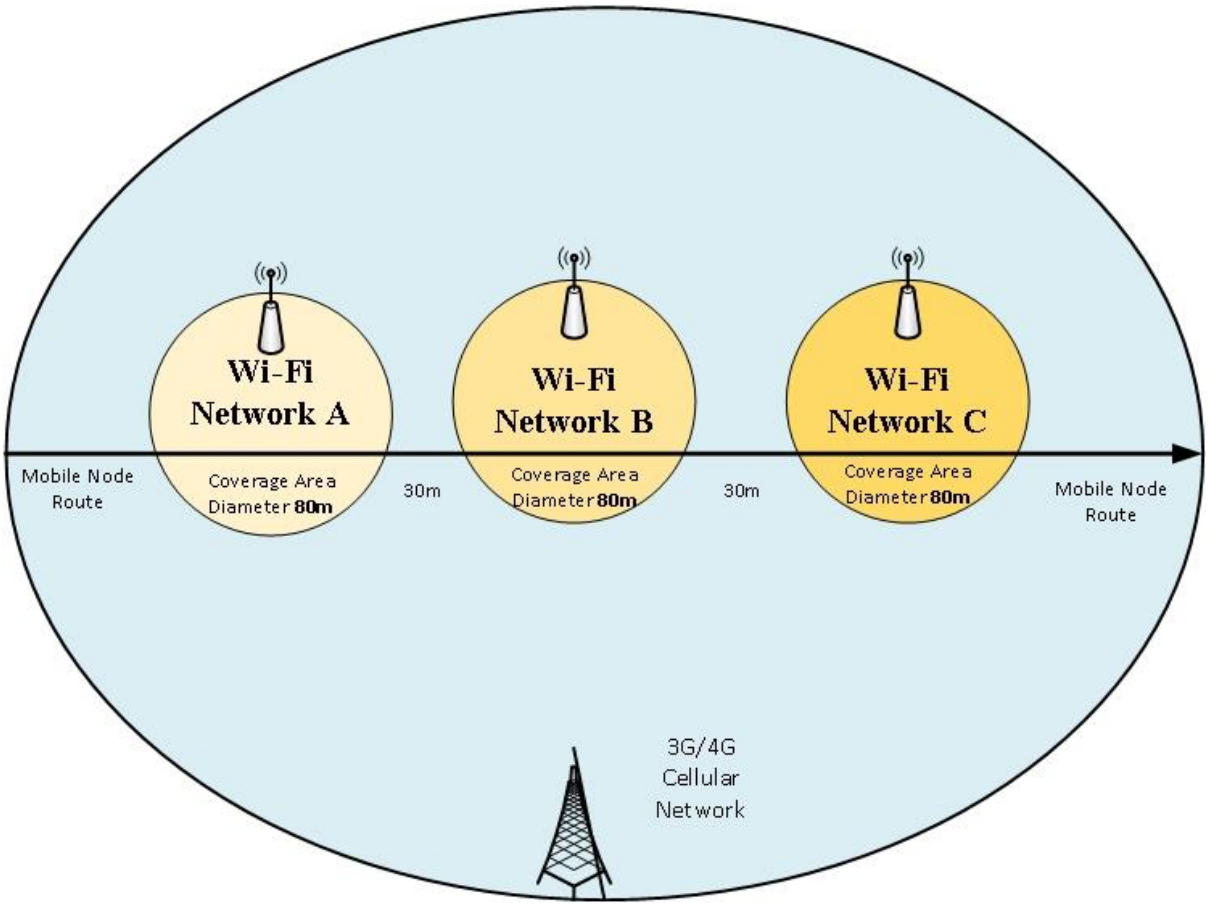

Figure 32 MDF simulation environment

Scenario 1 Part 1: In Scenario 1 Part 1 the mobile node was in a stationary position, approximately 10 metres to the left of the centre of Wi-Fi AP A's coverage area. The coverage areas of the Wi-Fi APs were 80 metres in diameter and initially, the mobile node established a connection to the cellular network's eNodeB. The cellular connection was maintained for the duration of an 8 second connection delay while a new connection was established to the Wi-Fi AP. In this simulated scenario there was a single handover from cellular to Wi-Fi without any loss of connectivity due to the "make before break" nature of the handover. When the connection to the AP was completed the cellular connection was dropped. As the node remained stationary for the duration of the simulation and remained within the coverage area of the AP there were no further handovers. The simulation had a duration of 193 seconds equivalent to the duration of the video content used [152]. MDF was not implemented.

Scenario 1 Part 2: In Scenario 1 Part 2 the mobile node remained in a stationary position approximately 10 metres to the left of the centre of Wi-Fi AP A's coverage area. The coverage area of the AP was 80 metres in diameter. Initially, the mobile node established a connection to the cellular network's eNodeB and the cellular connection was maintained during the 8 second connection delay while a new connection was established to AP A. Again in this simulation there was a single handover from cellular to Wi-Fi without any loss of connectivity due to the "make before break" nature of the handover. When the connection to the AP was completed the LTE connection was dropped. As the node remained stationary for the duration of the simulation and remained within the coverage area of the AP there were no further handovers. The simulation had a duration of 193 seconds equivalent to the duration of the video content used [152]. MDF was implemented during the simulation.

Scenario 2 Part 1: In Scenario 2 Part 1 the mobile node travelled a straight line path through the coverage areas of the 3 Wi-Fi APs at a constant speed of 1.4 metres per second, the average pedestrian speed in an urban environment. On start-up and before the mobile node began its journey from a position 50 metres to the left of the centre of AP A's coverage area the mobile node established a connection to the LTE network. The mobile node connected to the Wi-Fi APs whenever it came in range of one and once a connection was made to an AP the cellular connection was dropped. All handovers from cellular to Wi-Fi were "make before break" with no loss of connectivity but each connection to an AP experienced an 8 second connection delay during which the mobile node maintained its LTE connection. In

contrast, all handovers from Wi-Fi to cellular were “break before make” during which the mobile node experienced a 7 second total loss of connectivity for each handover completed. The simulation finished with the end of the video content, a period of 193 seconds during which MDF was not implemented.

Scenario 2 Part 2: In Part 2 of Scenario 2 the mobile node again travelled a straight line path through the coverage areas of the 3 Wi-Fi APs at a speed of 1.4 metres per second. When the simulation started the mobile node established a connection to the LTE network before it began its journey from a position 50 metres to the left of the centre of AP A’s coverage area. All handovers from cellular to Wi-Fi were “make before break” with no loss of connectivity but a connection delay of 8 seconds was experienced by the node and once the connection to the AP was made the LTE connection was dropped. As soon as the mobile node reached the edge of an APs coverage area it lost connectivity to the AP and all handovers from Wi-Fi to cellular were “break before make”. Each of these handovers incurred a 7 second total loss of connectivity for each one that took place. MDF was implemented in this simulation and the simulation finished with the end of the video content, a duration of 193 seconds.

Scenario 3 Part 1: When the simulation began the mobile node established a connection to the cellular network and then proceeded to travel a straight line path through the coverage areas of the 3 Wi-Fi APs at a speed of 5 metres per second. The node started its journey from a position 50 metres to the left of the centre AP A’s coverage area. Whenever the mobile node came within range of an AP it attempted to establish a connection to the AP. All handovers from cellular to Wi-Fi were “make before break” during which the node maintained its LTE connection while establishing a connection to the AP which resulted in no loss of connectivity. The “make before break” handovers had a duration of 8 seconds, equivalent to the Wi-Fi connection delay presented in Chapter 4 Section 4.2.3. If the node was successful in connecting to the AP, the LTE connection was dropped and when the node reached the right-hand edge of the APs coverage area the Wi-Fi connection was lost. All handovers from Wi-Fi to cellular were “break before make” and incurred a 7 second loss of connectivity for each handover that occurred. The duration of the simulation was 193 seconds and MDF was not implemented during the simulation.

Scenario 3 Part 2: The mobile node travelled a straight line path through the coverage areas of the 3 Wi-Fi APs at a speed of 5 metres per second. It started its journey from a position

50 metres to the left of the centre of AP A's coverage area having first established a connection to the cellular network. When the mobile node reached the left-hand edge of an AP's coverage area it attempted to establish a connection to the AP. All handovers from cellular to Wi-Fi were "make before break" during which the node maintained its LTE connection while attempting to establish a connection to the AP. The "make before break" handovers took 8 seconds to complete during which there was no loss of connectivity. As soon as the mobile node reached the right-hand side of the AP's coverage area the connection to the AP was lost. The node then began the process of re-connecting with the cellular network, however, all handovers from Wi-Fi to cellular were "break before make" and incurred a 7 second loss of connectivity for each hand over that occurred. The duration of the simulation was 193 seconds and MDF was implemented by the mobile node.

Scenario 4 Part 1: Having established a connection to the cellular network the mobile node travelled a straight line path through the coverage areas of the 3 Wi-Fi APs at a speed of 10 metres per second. As was the case in the previous simulations it started its journey from a position 50 metres to the left of the centre of AP A's coverage area. When the mobile node reached the left-hand side of an AP's coverage area it attempted to establish a connection to the AP. All handovers from cellular to Wi-Fi were "make before break" during which the mobile node maintained its LTE connection while connecting to the AP, this strategy resulted in no loss of connectivity. When the mobile node reached the right-hand side of the AP's coverage area it lost the connection to the AP at which time it started the process of re-connecting to the cellular network. A total loss of connectivity for a period of 7 seconds was experienced by the mobile node during the re-establishment of its LTE connection. The duration of the simulation was 193 seconds and MDF was not implemented.

Scenario 4 Part 2: On initialisation the mobile node established a connection to the cellular network before traveling a straight line path through the coverage areas of the 3 Wi-Fi APs at a speed of 10 metres per second. The node started its journey from a position 50 metres to the left of the centre of AP A's coverage area. As soon as the node reached the left-hand side of an AP's coverage area it began the process of establishing a connection to the AP. All handovers from cellular to Wi-Fi were "make before break" with a duration of 8 seconds during which the node maintained its LTE connection while attempting to connect to the AP with no loss of connectivity. If the node established a connection to the Wi-Fi AP, the LTE connection was dropped. When the node arrived at the right-hand edge of the AP's coverage

area the Wi-Fi connection was dropped and the node attempted to re-establish the cellular connection. However, all handovers from Wi-Fi to cellular were “break before make” and incurred a 7 second loss of connectivity for each hand over that occurred. The duration of the simulation was 193 seconds and MDF was implemented.

Table 17 MDF Scenario Summary Table

Scenario ID	Node Speed Over Ground	MDF Implemented	AP Coverage Area Diameter	Initial Conn Type	Node Start Position	Mobile Node Travel Pattern	Number of Wi-Fi APs	Number of eNodeBs	Simulation Duration
Scenario 1 Part 1	0 metres per sec	NO	80 metres	cellular	10 m to left of the centre of AP A's coverage area	Straight line left to right	3	1	193 seconds
Scenario 1 Part 2	0 metres per sec	YES	80 metres	cellular	10 m to left of the centre of AP A's coverage area	Straight line left to right	3	1	193 seconds
Scenario 2 Part 1	1.4 metres per sec	NO	80 metres	cellular	50 m to left of the centre of AP A's coverage area	Straight line left to right	3	1	193 seconds
Scenario 2 Part 2	1.4 metres per sec	YES	80 metres	cellular	50 m to left of the centre of AP A's coverage area	Straight line left to right	3	1	193 seconds
Scenario 3 Part 1	5 metres per sec	NO	80 metres	cellular	50 m to left of the centre of AP A's coverage area	Straight line left to right	3	1	193 seconds
Scenario 3 Part 2	5 metres per sec	YES	80 metres	cellular	50 m to left of the centre of AP A's coverage area	Straight line left to right	3	1	193 seconds
Scenario 4 Part 1	10 metres per sec	NO	80 metres	cellular	50 m to left of the centre of AP A's coverage area	Straight line left to right	3	1	193 seconds
Scenario 4 Part 2	10 metres per sec	YES	80 metres	cellular	50 m to left of the centre of AP A's coverage area	Straight line left to right	3	1	193 seconds

5.4 Results

In the following section graphs representing the playout buffer levels for each simulation are presented. For example, Figure 33 presents the playout buffer level for a simulation in which MDF was not implemented and it can be seen that the buffer level varies frequently and rapidly. When MDF was not implemented the DASH enabled client reverted to its default behaviour of downloading segments having the highest bitrate possible based on the client's estimation of the link capacity. The reason why the effective media throughput, as represented by the playout buffer levels, does not improve when increasing the segment size is that the available bandwidth in the simulation fluctuates over time. When larger segments are used the client is not able to adjust as quickly and flexibly as would be possible with shorter segments and therefore the buffer levels deteriorate for longer segment lengths. Figure 34 presents the playout buffer for a simulation in which MDF was implemented and it can be seen that in these simulations the buffer levels were more consistent. This was a result of the client not seeking to download the segments having the highest bitrates but downloaded smaller segments.

Some of the fluctuations in buffer levels cannot be explained by the behaviour of the DASH enabled client, for example, in Figure 36 we see two drops in buffer levels not associated with handover activity. It is not clear as to the cause of these events but we speculate that they arise form processing bottle necks in the simulations caused by the use of external Linux containers mapped onto the NS3 simulated network via tap-bridge devices.

5.4.1 Scenario 1 Part 1 Mobile Node in Stationary Position MDF Not Implemented

Figure 33 presents the playout buffer level in seconds over the course of the simulation for a stationary node which has not implemented MDF (Scenario 1 Part 1). When MDF is not implemented the DASH enabled media player reverts to its default behaviour of requesting the video segments based on available link capacity. In this scenario the DASH enabled player requests video segments having the highest bit rate without regard to the device screen size. Downloading larger files than required takes longer and during the download process the buffer levels drop as the existing buffer contents are consumed.

In heterogeneous networks users may experience delays in establishing connections to wireless networks when they change their point of attachment from one network to another. They may also face periods of time during which they are completely disconnected and rely on the contents of the playout buffer to maintain their video playback. The red horizontal line in Figure 33 represents a desired buffer level of 8 seconds, if the buffer level falls below this 8 second mark the mobile node may be unable to bridge gaps in connectivity due to handovers from Wi-Fi to cellular networks. This can result in depletion of the playout buffer and the occurrence of stalling events which have a negative impact on end-user Quality of Experience (QoE). It is apparent from Figure 33 that on multiple occasions during the simulation run the buffer level fell below the 8 second mark. Two stalling events were also observed at the 45 second and 189 second points of the simulation when the buffer level fell to zero.

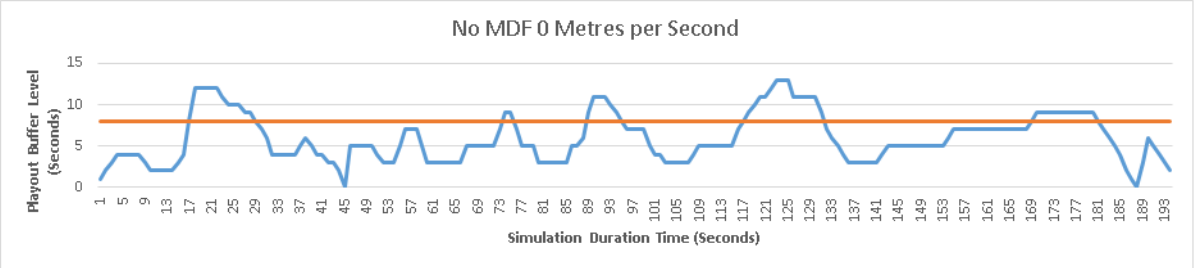

Figure 33 Stationary user MDF not implemented

5.4.2 Scenario 1 Part 2 Mobile Node in Stationary Position with MDF Implemented

Figure 34 presents the playout buffer levels in seconds for a stationary node which has implemented MDF, the segment sizes requested for downloading were based on a smartphone screen resolution and a segment duration of 2 seconds. Segments were not requested based on available bandwidth and severe fluctuations in buffer levels was avoided. Apart from a period of 9 seconds at the start of the video the playout buffer level remained above the 8 second mark (indicated by the red horizontal line) for the duration of the simulation run and no stalling events were observed. Maintaining a playout buffer level of at least 8 seconds is important for stationary nodes since there is no guarantee that they will remain in the same location for the duration of the video. Changing position has the potential to cause a handover from the Wi-Fi AP to the cellular network which will result in a break in connectivity due to the delay in establishing a cellular data connection. If the buffer level is less than 8

seconds when this event occurs there is a risk that the playout buffer will become depleted before the data connection is re-established and stalling events will occur.

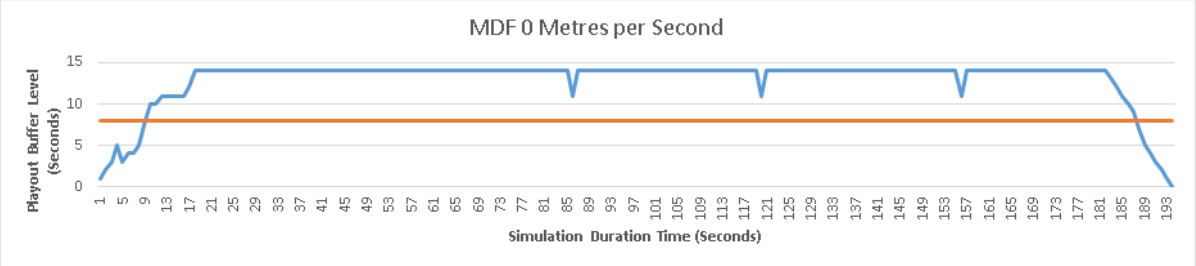

Figure 34 Stationary user MDF implemented

5.4.3 Scenario 2 Part 1 Mobile Node Moving at 1.4 metres per second MDF Not Implemented

The playout buffer levels in seconds over the course of the simulation for a mobile node travelling at 1.4 metres per second while not implementing MDF are presented in Figure 48. The node requested segments based on available bandwidth and not screen size leading to larger than necessary segments being downloaded. The graph shows wide fluctuations in buffer levels similar to those in Scenario 1 Part 1. It is clear from the graph that for the majority of the simulation the buffer level was below 8 seconds. In addition, there were 4 stalling events at 117 seconds, 68 seconds, 86 seconds and 117 seconds where the playout buffer level fell to zero.

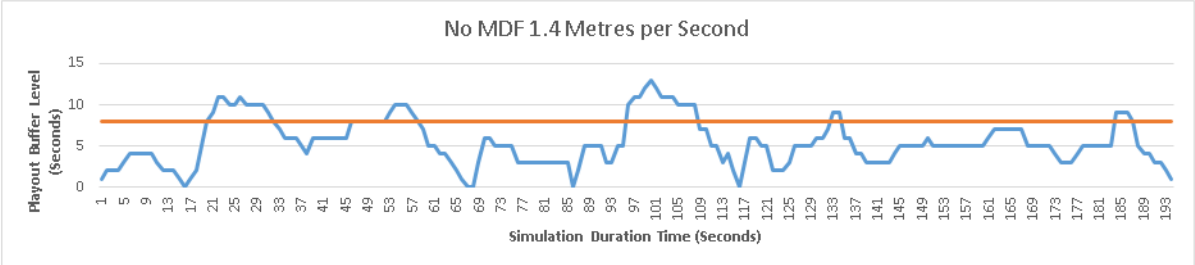

Figure 35 Mobile user travelling at 1.4 mps MDF not implemented

5.4.4 Scenario 2 Part 2 Mobile Node Moving at 1.4 metres per second MDF Implemented

Figure 36 shows the playout buffer levels for a mobile node travelling at 1.4 metres per second which has implemented MDF. Segment requests were based on a smartphone screen size and a duration of 2 seconds. Buffer levels were stabilised with no stalling events being observed and for the majority of the simulation the buffer level was above the 8 second mark. However, the buffer level fell below 8 seconds on two separate occasions centred around the 68 second and 142 second points in the simulation which match handovers between Wi-Fi and cellular networks.

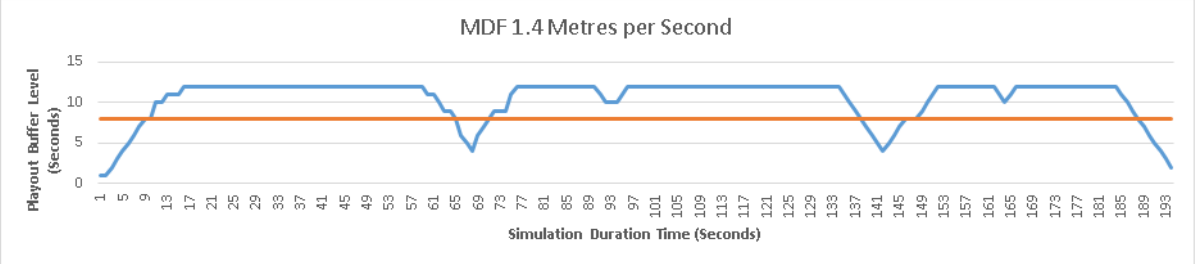

Figure 36 Mobile user travelling at 1.4 mps MDF implemented

5.4.5 Scenario 3 Part 1 Mobile Node Moving at 5 metres per second MDF Not Implemented

The playout buffer levels for a mobile node which has not implemented MDF travelling at 5 metres per second are presented in Figure 37. The segment selection strategy was the DASH enabled media player default and segments were selected based on available bandwidth. Five stalling events were observed at the 22 second, 70 second, 79 second and 125 second points during the simulation run. In addition, for the majority of the simulations duration the buffer level was below 8 seconds.

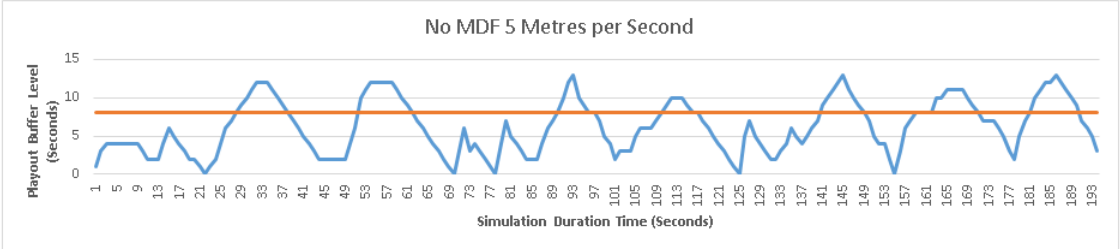

Figure 37 Mobile user travelling at 5 mps MDF not implemented

5.4.5 Scenario 3 Part 1 Mobile Node Moving at 5 metres per second MDF Implemented

Figure 38 presents the playout buffer levels for a mobile node which had implemented MDF travelling at 5 metres per second. The segment selection strategy was to select segments having a duration of 2 seconds and a resolution that reflected a smartphone screen size. No stalling events were observed during the simulation and the buffer level remained above 8 seconds for approximately 96% of the simulation duration time. The buffer level fell below the 8 second mark on two occasions centred around the 24 second and 66 second points in the simulation which matches handovers between Wi-Fi and cellular networks.

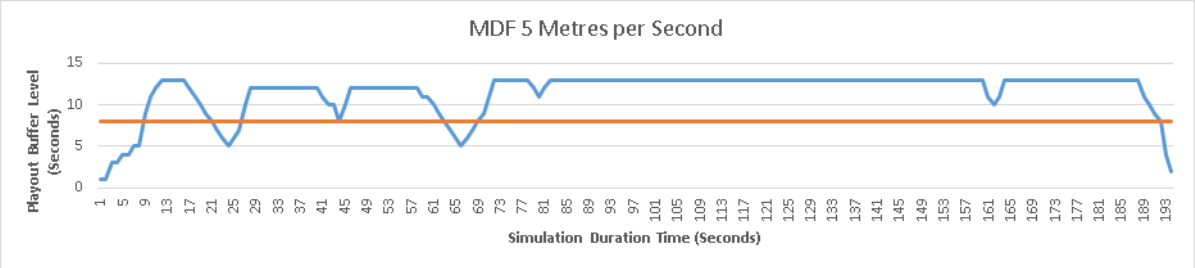

Figure 38 Mobile user travelling at 5 mps MDF implemented

5.4.6 Scenario 4 Part 1 Mobile Node Moving at 10 metres per second MDF Not Implemented

The playout buffer levels for a mobile node which had not implemented MDF and was travelling at 10 metres per second are presented in Figure 39. The default segment selection strategy of selecting segments based on available bandwidth was employed. For the majority of the simulation run time the buffer level was below the 8 second mark. Three stalling events were also observed, occurring at the 54 second, 61 second and 74 second points in the simulation.

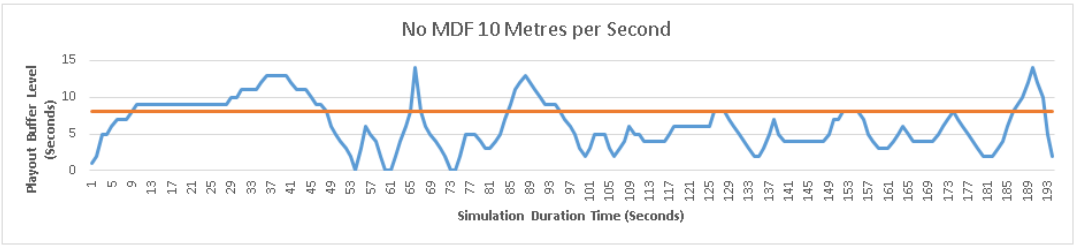

Figure 39 Mobile user travelling at 10 mps MDF not implemented

5.4.6 Scenario 4 Part 1 Mobile Node Moving at 10 metres per second MDF Implemented

Figure 40 presents the playout buffer levels for a mobile node travelling at 10 metres per second which had implemented MDF. The segment selection strategy was to select segments of 2 seconds duration with a resolution appropriate for a smartphones screen size. The buffer level remained above the 8 second mark for the majority of the simulation apart from a brief period of time during the initial stage of the streaming operation. No stalling events were observed during the simulation.

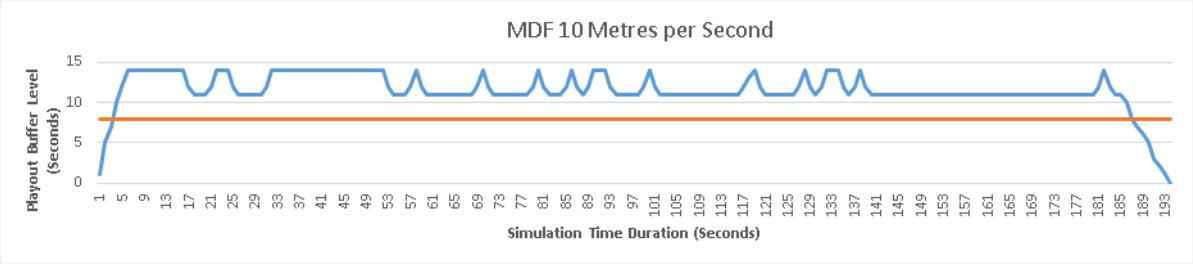

Figure 40 Mobile user travelling at 10 mps MDF implemented

5.5 Periods of Connectivity as a Percentage of Simulation Duration

The percentage of the simulation time during which the mobile node was connected to either a cellular or Wi-Fi network or was in a completely disconnected state are shown in Table 18. Also shown is the percentage of the simulation during which the playout buffer was less than 8 seconds and the MDF implementation status for each simulation.

Table 18 MDF Status & Percentage of Sim Time Buffer Level < Than 8 seconds

Node Speed	**MDF Status**	**Wi-Fi % of Time**	**LTE % of Time**	**% of Time Discon-nected**	**% of Time Buffer < 8 secs**
0 mps	OFF	91.2%	8.8%	0%	69.4%
0 mps	ON	91.2%	8.8%	0%	5.2%
1.4 mps	OFF	61.65%	30.6%	7.25%	76.7%
1.4 mps	ON	61.65%	30.6%	7.25%	12.4%
5 mps	OFF	12.5%	76.6%	10.9%	67.8%
5 mps	ON	12.5%	76.6%	10.9%	5.7%
10 mps	OFF	0%	92.75%	7.25%	70.5%
10 mps	ON	0%	100%	0%	0%

5.6 Data Transfers

Table 19 presents the total amount of video content received by the client in Megabytes (MB) with a breakdown of the amount of video content downloaded over Wi-Fi and cellular connections. Also shown is the reduction in the amount of data downloaded over the cellular connection when MDF is implemented.

Table 19 Data Downloads and MDF Status

Node Speed	MDF Status	Total Download (MB)	Wi-Fi Down-load (MB)	LTE Down-load (MB)	Reduction in LTE Data MDF (MB)	Reduction in LTE Data MDF (%)
0 mps	OFF	123	112.2	10.8	NA	NA
0 mps	ON	27	24.6	2.4	8.4	77.8%
1.4 mps	OFF	96	64.3	31.7	NA	NA
1.4 mps	ON	30	20.1	9.9	21.8	68.8%
5 mps	OFF	96	13.44	82.56	NA	NA
5 mps	ON	26	3.64	22.36	60.2	73%
10 mps	OFF	58	0	58	NA	NA
10 mps	ON	19	0	19	39	67.2%

5.7 Analysis

5.7.1 MDF not implemented by mobile node

Part 1 of each simulation begins with the mobile user having an established connection to the cellular network prior to attempting to establish a connection to alternative Wi-Fi networks. In every simulation in which MDF was not implemented by the client the playout buffer level fell below the 8 second mark (Figures 33, 35, 37, 39) on multiple occasions. When the node was stationary the buffer level was below 8 seconds 69.4% of the time, at a speed of 1.4 metres per second the buffer level was less than 8 seconds 76.7% of the time, at 5 metres per second the buffer level was less than 8 seconds 67.8% of the time and at a speed of 10 metres per second the buffer level was less than 8 seconds 70.5% of the time. In

addition, stalling events occurred when the buffer level fell to zero and the playout buffer was depleted. The client downloaded segments based on available bandwidth and not the device screen size resulting in excessive amounts of data being requested over cellular and Wi-Fi links.

5.7.2 MDF implemented by mobile node

In Part 2 of each of the simulations the mobile node implemented MDF and the stationary node implementing MDF saw the greatest benefits from switching from an existing cellular connection to an alternative Wi-Fi network when one was available. The initial handover from cellular to Wi-Fi was made using a "make before break" strategy which resulted in no loss of connectivity (Figure 33). The stationary user was able to achieve a 77.8% reduction in the amount of data downloaded over a cellular network (Table 19). The percentage of simulation time during which the buffer level was below 8 seconds was reduced from 69.4% of the time to 5.2% of the time.

A mobile user travelling at 1.4 metres per second with MDF implemented had a playout buffer level of less than 8 seconds 12.4% of the time. In comparison, the same user who did not implement MDF had a buffer level of less than 8 seconds 76.7% of the time. No stalling events were experienced and a reduction of 68.8% in the amount of data down loaded over the cellular connection was achieved.

At a speed of 5 metres per second a mobile user employing MDF experienced a buffer level of less than 8 seconds 5.7% of the time. In contrast to this a mobile user travelling at 5 mps without implementing MDF had a buffer level of less than 8 seconds 67.8% of the time. Additionally, no stalling events were observed and there was a 73% reduction in the amount of data downloaded over the cellular connection.

When the mobile user was travelling at 10 metres per second with MDF implemented the playout buffer levels never fell below the 8 second mark and no stalling events were observed. Although all data was downloaded over a cellular connection a reduction of 67.2% in the total was achieved in comparison to the total amount of data downloaded when MDF was not implemented. This reduction in total data downloaded was achieved by reducing the video segment resolution to match the device screen size as closely as possible. The use of a smartphone is assumed.

5.8 Conclusions

This section introduces an innovative MPEG-DASH-based Framework (MDF) for improving end-user video streaming experience in heterogeneous multi-network wireless environments. These improvements include improved end-user QoE through a reduction in the number of stalling events that occur, the capacity to 'bridge' gaps in connectivity due to hard handovers from Wi-Fi to cellular networks and protecting user data-caps through reducing the overall amount of data downloaded over a cellular connection. The results of the simulations clearly demonstrate that implementation of MDF is beneficial to the user. MDF aims to maintain the client's playout buffer levels above 8 seconds, the average delay experienced when attempting to connect to a Wi-Fi AP. This approach reduces the number of stalling events due to loss of connectivity during handovers from Wi-Fi to cellular networks by ensuring that the buffer contains enough content to maintain the video play back until a new connection can be established.

Maintaining playout buffer levels greater than 8 seconds is only required when connected to Wi-Fi networks in order to be prepared for handovers. While connected to a Wi-Fi AP the client can maximise their playback time by pre-loading video segments to the buffer without effecting their mobile provider imposed data-cap. When connected to a cellular network and conditions such as speed over ground make connecting to a Wi-Fi network non-viable the DASH enabled client reverts to its default behaviour when requesting video segments. It does not seek to maintain a buffer level greater than 8 seconds since no hard handovers will take place in this situation. However, MDF does attempt to match the devices screen size as closely as possible to available video segment resolutions to avoid downloading segments with resolutions too large for the device's screen. This approach reduces the amount of data downloaded over the cellular link. MDF fulfils Thesis Objectives 2 and 3 introduced in Chapter 1, Section 1.6 and reproduced below:

2 Enable a mobile user to maximise Quality of Experience while streaming SD video

4 Reduce the overall amount of data transferred over cellular connections

MDF helps maintain QoE by reducing the number of stalling events during a SD video streaming session in a heterogeneous wireless environment. It reduces the amount of data transferred over a cellular connection by streaming content over Wi-Fi where possible and by matching video segment resolution as closely as possible to device screen size.

CHAPTER 6 ADAPTIVE INTERFACE SELECTION (AIS)

In the previous chapter, the MDF module for managing SD video streams was introduced. This chapter presents AIS which is initialised by SONS when a HD video stream is detected. AIS is described in detail and its operation verified through a series of simulations. The results of the simulations are presented and discussed.

6.1 Motivation

Modern video streaming employs adaptive bitrate (ABR) algorithms to deal with the fluctuations in available bandwidth experienced by mobile users in wireless environments. MPEG-DASH enabled clients adapt the bitrate of requested video segments to match as closely as possible the available bandwidth of the wireless link in use. If link conditions are poor or the link is congested, the ABR algorithm will select the lowest available bitrate supported by the video stream. This approach works well for standard video content since there is no restriction on using low bitrates.

However, for users wishing to view High Definition (HD) content, there is a lower bound to the bitrate below which they cannot go if the content is to be considered HD. ABR algorithms, forced to maintain HD bitrates over wireless links with constrained bandwidth, experience significant numbers of stalling events. High numbers of stalling events, particularly towards the end of a video, have a serious negative impact on end-user QoE.

This section introduces Adaptive Interface Selection (AIS) to address this issue. AIS leverages the multi-homed nature of modern mobile devices and the segmented nature of video content prepared for DASH-enabled clients to overcome the bandwidth constraints that hobble ABR algorithms when streaming HD content. When necessary AIS uses multiple wireless interfaces in parallel to download content to maintain the playout buffer level to prevent stalling events. By minimizing the number of stalling events and streaming at HD compatible bitrates end-user QoE is maintained. AIS also seeks to download those video segments that are of the lowest HD compatible resolution. This strategy aids in maintaining buffer levels and downloading as little content as possible over mobile networks helps protect the end-user's mobile data-cap. If AIS is employing both the cellular interface and the Wi-Fi interface in parallel it will attempt to down load the video segment need next over the cellular interface and pre-load a second segment for later user over the Wi-Fi interface.

Other solutions that seek to protect the user's data-cap have also been developed including [153]. Koch et al. propose prefetching complete videos while the user is connected to alternative networks such as Wi-Fi or wired broadband. This strategy does protect the user's data-cap but is not very flexible and unlike AIS does not permit protecting the data allowance during spontaneous streaming of video content. It also gives rise to serious estimation and prediction challenges that are difficult to overcome. There have also been QoE optimisation schemes aimed at streaming ultra-HD video over 5G networks such as [154]. Unlike AIS these schemes are not concerned with reducing the amount of data downloaded over the cellular network and completely ignore the fact the mobile device is multi-homed.

6.2 Introduction

Successful delivery of High Definition (HD) video content to mobile users in heterogeneous, multi-network wireless environment is challenging. Due to its high bitrate HD video requires reliable, high capacity communication links. However, communications in wireless networks are characterized by unreliable communication links and rapidly changing levels of available bandwidth. End-user satisfaction with a video streaming service is closely tied to user quality of experience (QoE).

The end-user's QoE is influenced by many factors including the initial delay in playing the selected video, the number and duration of any interruptions to the video playback (stalling events), the location of any stalling events within the video itself (start, middle or end of video), and drastic shifts in the quality of the video [155], [156]. The amount of bandwidth available and fluctuations in the bandwidth have a direct, powerful effect on the QoE of the viewer [157]. In particular, consumers of HD video will experience a dramatic reduction in QoE if the content being viewed suddenly switches from HD to a low bitrate representation due to a reduction in available bandwidth [158].

Dynamic Adaptive Streaming over HTTP (DASH) is an example of an adaptive streaming strategy that has been introduced to address the challenges posed by large fluctuations in available bandwidth. DASH operates by segmenting the content to be delivered into chunks of various durations e.g. 2 seconds, 4 seconds, 6 seconds, 10 seconds and encoding each chunk at various bit rates. This process results in multiple copies of the content being stored on the HTTP server.

Figure 41 DASH enabled video streaming over single link

DASH enabled clients deal with fluctuations in the available bandwidth by dynamically adapting the bitrate of the content being requested to match as closely as possible the available bandwidth of the wireless link. An example of the DASH-based adaptive delivery process is illustrated in Figure 41. If the amount of available bandwidth increases the bitrate of the requested segments increases and if there is a reduction in the available bandwidth there is a reduction in the bitrate of the video segments being requested. However, in the case of HD video content there is a lower bound to the video bitrate below which the DASH enabled client cannot go if the content is to be viewed as HD video. The HD lower bound for video is typically stated as 720p (1280p x 720p) requiring a minimum link capacity of 2.5 Mbps.

In this context, the challenge is to maintain QoE for the end-user employing a DASH-enabled media player when the lowest acceptable bitrate for HD video is greater than the currently available bandwidth of a link over any single network. One possible solution is to employ an approach that uses multiple wireless links simultaneously in order to support high bitrate content delivery such as that of HD video. This section introduces Adaptive Interface Selection (AIS), a client-side interface controller and content requesting system that seeks to maintain QoE for mobile consumers of HD video content in bandwidth constrained wireless environments.

AIS leverages both the multi-homed nature of mobile devices and the segmented nature of video content created for delivery using DASH enabled servers and clients. It activates and de-activates the wireless interfaces in response to changes in network conditions downloading video segments over multiple interfaces where appropriate to maintain playout buffer levels to reduce stalling events and switches in bitrate.

The proposed system protects the end-user's data-cap on mobile networks by downloading content over Wi-Fi where feasible and restricting segment bitrates, reduces energy consumption by restricting unnecessary Wi-Fi scanning operations and maintains QoE by reducing stalling events and bitrate switches. AIS is invoked by the SONS Resolution Discovery Module described in Chapter 4 when the RDM determines that the MPD file for a HD video stream has been downloaded to the MPD-Cache directory (Chapter 4, Section 4.5).

6.3 Adaptive Interface Selection (AIS) Overview

Modern smart-phones and other portable devices such as tablets are equipped with multiple wireless interfaces by default, these typically consist of a Wi-Fi interface and an interface for connecting to mobile networks. Users of cellular data plans are constrained by monthly download limits known as data-caps and breaching these data-caps can result in excessive charges per MB of data over the limit and possible throttling of the connection.

Under normal usage conditions the various interfaces of a device are used in isolation, as illustrated in Figure 42. However, using multiple interfaces simultaneously would enable the user to increase the overall bandwidth share available to them and support higher bitrate streaming such as that of HD video.

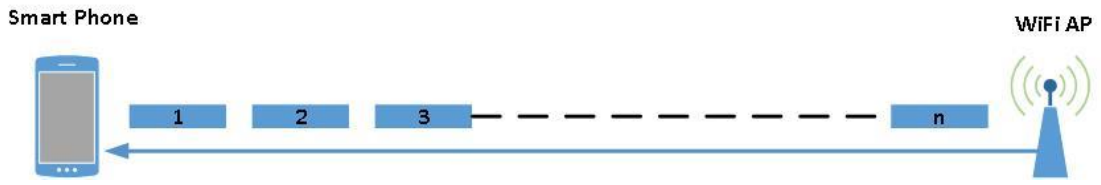

Figure 42 MPEG-DASH segment transfer over a single Wi-Fi link

Instead of constantly adjusting the bitrate of the content to match the available bandwidth, we increase the available bandwidth to match the bitrate required for the target HD content, which is restricted to the video segments having the lowest HD compatible resolution (Figure 44). This is achieved by activating both interfaces simultaneously and load balancing the segment transfers over both links (Figure 43, Figure 44) when necessary. Load balancing is a well-known technique for the simultaneous transfer of data across multiple resources such as physical or wireless interfaces.

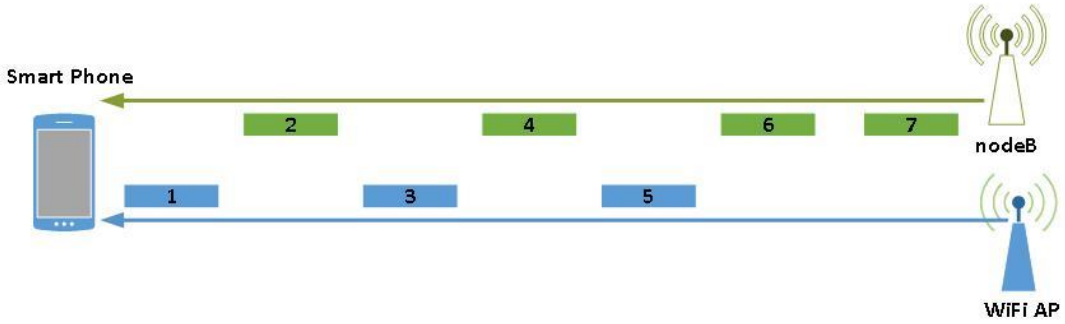

Figure 43 MPEG-DASH segment transfer over multiple links

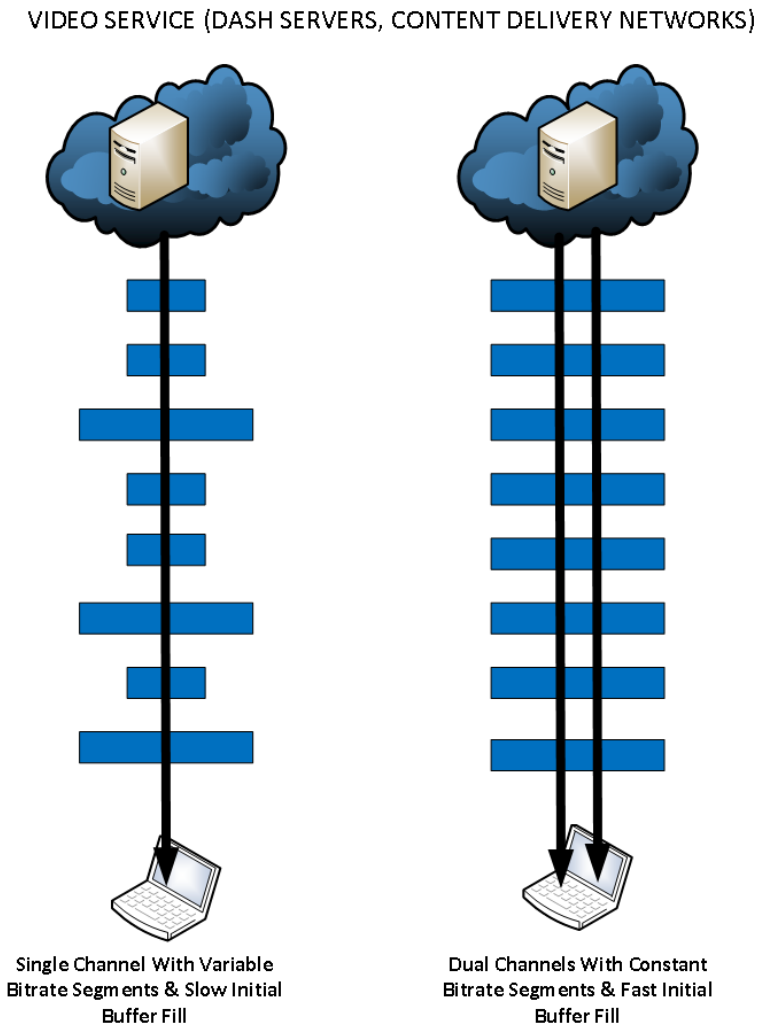

Figure 44 Single Channel with Variable Bitrate Segments vs Dual Channel with Fixed Bitrate Segments

The aim of load balancing is to maximise throughput and to reduce overloading any single link or interface. Load balancing can also increase reliability and availability through redundancy by using multiple components instead of a single component. For mobile devices load balancing comes at an increased cost in terms of energy consumption and should be used cautiously.

The segmented nature of the content on a DASH server in conjunction with the multi-homed capabilities of modern mobile device provide the basis of a solution to the problem of how to stream fixed bitrate video content over bandwidth constrained links. Typically, mobile users of smart devices will connect to Wi-Fi where available in order to reduce the amount of data they download over their cellular connection and then disconnect their cellular data connection. They do so to take advantage of often faster connection speeds, reduce their energy consumption and protect their data-cap. Normally a DASH enabled client must determine the throughput of the wireless link currently in use in order to be able to request segments having a bitrate that can be accommodated on the link. Monitoring the level of the playout buffer as demonstrated by [129], [139], [159] provides a simple means of inferring the capacity of the link without having to calculate the actual throughput.

If the level of the playout buffer falls too low, then the video playback begins to experience stalling events which are detrimental to the QoE of the viewer. Stalling events indicate that the link capacity is insufficient to support the current segment bitrate. If the client device is connected to a Wi-Fi AP, the AIS system re-activates the cellular link and begins downloading the next segment in the sequence as indicated by the MPD (manifest) over the cellular network. The playout buffer levels are increased rapidly by having alternating segments downloaded over both the Wi-Fi link and the cellular link until the playout buffer level has stabilized. When the playout buffer has regained a level at which the risk of stalling events occurring has been reduced to near zero, the cellular link is de-activated and the buffer continues to be monitored. A new temporary higher trigger level is implemented (e.g. activating the secondary interface when the buffer level is 10 seconds rather than 2 seconds), so that the standby interface can be activated sooner if necessary when the buffer level drops too low again.

6.4 AIS System Architecture and Operation

The Adaptive Interface Selection (AIS) system architecture consists of the following modules (Figure 45):

- SONS Interface (SI)
- Buffer Monitor (Bmon)
- Segment Selection Module (SSM)
- Interface Controller (IC)
- Prefetch

The SONS Interface (SI) polls the SONS module (Chapter 4) for the latest utility score to determine whether or not conditions are such that a useful connection could be established to a Wi-Fi AP, in addition, the SI also requests the "freshness" timer value from SONS. When SONS calculates the utility score a "freshness" timer is set to a value of 5 seconds and is decremented by 1 for every second that elapses, and when the timer reaches 0 the utility score is recalculated. The freshness timer value indicates the period of time in seconds until the next utility score is calculated. AIS takes the freshness counter value and uses it to synchronise its utility score requests with the utility value calculations in order to ensure that it has the most recent utility score. To reduce the initial delay, if the freshness timer value is greater than 3 AIS uses the current utility score and sets its update time to match the freshness timer. If the freshness timer value is less than 3 AIS sets its update timer to match the freshness value and waits until the utility score is updated.

The retrieved utility score is passed to the Interface Controller (IC). If a utility score of 1 is returned by SONS then conditions are suitable for attempting to connect to Wi-Fi, and a returned value of zero indicates that conditions are not suitable for connecting to Wi-Fi APs. On receipt of a SONS utility score of 1 the Interface Controller (IC) checks the status of the Wi-Fi interface, if the interface has already been activated the IC ignores it otherwise the IC activates the interface.

Buffer Monitor (Bmon) is responsible for monitoring the level of the DASH enabled client's playout buffer. It requests actions from either the Segment Selection Module (SSM) or the Interface Controller (IC) based on the observed buffer level. On system initialization the Bmon alerts the Interface Controller (IC) to activate the cellular data connection if it is not

in an active state. The Bmon then checks the playout buffer level and if the buffer is empty the Bmon sends a request to the SSM to begin downloading the required video content. When the playout buffer has been populated and play back begins the Bmon requests that the SI requests the current utility score and freshness value from SONS. SONS invoke the user defined network detection and selection algorithm if it calculates a utility score of 1. If a suitable Wi-Fi network is found and a connection is established the next video segments in the sequence are downloaded over the Wi-Fi link and the cellular connection is suspended.

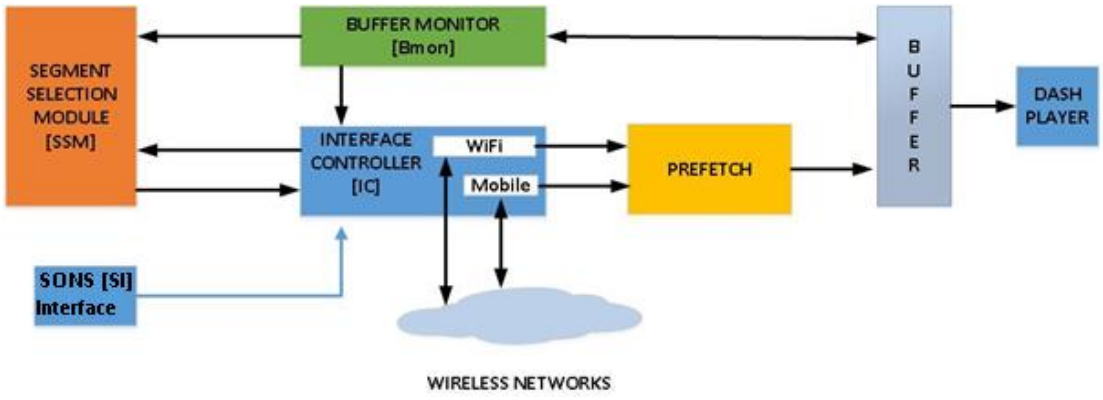

Figure 45 AIS Block diagram of system components

The Bmon continues to monitor the playout buffer and if the buffer level drops below 30 seconds the Bmon alerts the IC to enable the cellular connection in parallel with the Wi-Fi interface. Segments are then downloaded over both connections until the buffer level has reached satisfactory levels once again. In the event that the IC detects that the Wi-Fi interface has lost connectivity with the AP the IC automatically deactivates the Wi-Fi interface and continues downloading video segments over the cellular interface. When the Wi-Fi interface is de-activated the Bmon sets the target playout buffer level to 2 seconds in order to reduce the amount of data downloaded over the cellular link in anticipation of re-establishing a connection to a Wi-Fi network in the near future. Since there will be no hard handovers from Wi-Fi to cellular there is also no need to maintain a buffer level sufficient to enable the loss of connectivity due to a hard handover to be bridged.

The Segment Selection Module (SSM) is responsible for selecting the video segments to download and it uses a segment list generated during the initial stage of operations as part of the process. The Interface Controller interacts with the SSM and keeps it informed as to

the currently active interfaces. This information is used by the SSM to decide in which order the video segments will be downloaded. At the start of operations, the appropriate manifest file (mpd file) and header files are identified and a list of suitable video segments is generated. The list consists of segments having a resolution of 720p, the minimum HD resolution employed by multiple streaming services such as YouTube [160], Netflix [161], etc. If segments having a resolution of 720p are not available segments having the lowest HD resolution greater than 720p are selected. The initial video segments that meet the HD video requirements are identified and these video segments, the mpd (manifest) file and the header file are marked for downloading to Prefetch using the cellular network.

If the Wi-Fi interface is identified as the primary interface with the cellular data connection deactivated, the SSM requests segments from the front of the list. However, when the cellular interface is identified as the primary interface due to congestion on the Wi-Fi link the SSM requests segments from both the front and the back of the list simultaneously.

Segments from the front of the list are downloaded over the cellular connection (current highest capacity link) as they are required immediately. Segments from the back of the list are downloaded over the congested Wi-Fi link as they will not be required until later and there is more time available to retrieve them. The Interface Controller (IC) is responsible for activating and de-activating the wireless interfaces in response to messages from the SI and the Bmon. The IC must determine, based on input from the SI, whether or not a useful connection to an Access Point (AP) might be established since multiple short lived connections have a negative impact on both data transfer rates and energy consumption [162].

6.5 Test and Simulation Environment

6.5.1 AIS Testing Strategy

The testing of the AIS concepts was conducted in two stages as follows:

- Stage 1 examined the feasibility of using multiple communication links to maintain the playout buffer level and reduce stalling events using a physical testbed
- Stage 2 examined segment selection and interface activation performance

Note: During Stage 1 of testing no attempts were made to manage segment selection and downloading, the DASH enabled clients default segment retrieval strategies were accepted for use.

6.5.2 Stage 1 Testing

The physical test device employed for Stage 1 was a Lenovo ThinkPad laptop equipped with an Intel i7 processor, 16GB of RAM and an SSD drive running Debian 8 GNU/Linux.

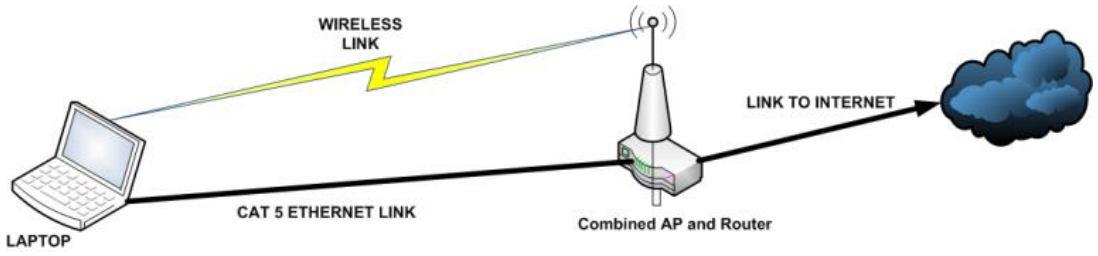

Figure 46 AIS Stage 1 Test Environment

The video content used in Stage 1 testing was the DASH Industry Forum [163] Spring 4Ktest Stream which was viewed with the browser based DASH IF Reference Client 2.4.1 [164]. The content was streamed from online the server over a 30 Mbps link supplied by a cable network operator connected via an on-premises Cisco EPC2425 wireless router. The distance between the test device and the wireless router was approximately 15 metres. The physical test bed is illustrated in Figure 46.

During testing a 15 metre CAT 5 Ethernet link was used to simulate the cellular network data connection and an onboard 802.11g adapter provided the Wi-Fi connectivity. The link speed for the emulated cellular connection was based on a speed of 6 Mbps (Chapter 4, Section 5.9.2, mean + 2 STD), this was the average cellular connection speed experienced by users in an Irish context. Wi-Fi connectivity speeds varied between 4.4 Mbps and 1.8 Mbps, reflecting the lower range of speeds experienced in real world usage of publically available Wi-Fi (Chapter 4, Section 5.9.2, mean – 2 STD). Link speeds were manipulated to produce the desired test speeds using the WonderShaper bandwidth throttling utility [165] available on the Linux platform.

In all test scenarios the test device remained stationary and the following assumptions were made:

- All links, Wi-Fi or emulated cellular, would be instantly available when required
- Cellular link speed was a constant 6 Mbps
- Wi-Fi link speeds varied between 4.4 Mbps and 1.8 Mbps
- Once a particular Wi-Fi link speed was selected it remained constant for the specified time periods

The proposed use of multiple interfaces to download video segments was emulated through the use of Linux interface bonding software [166]. The ifenslave utility was employed for this purpose and its round-robin load balancing mode selected. In this mode the higher capacity link, if it exists, is restricted to the speed of the lower capacity link.

Testing was conducted in two phases and the results compared. In Phase 1 the client device streamed a video formatted for MPEG DASH from the online media server without modifying its behaviour using a single wireless interface. During Phase 2 the client implemented the AIS interface selection strategy and employed one or both wireless interfaces as necessary depending on the playout buffer state.

6.5.3 Stage 1 Test Scenarios

The following scenarios were investigated and the results recorded for comparison.

Stage 1 Test Scenario 1: A single cellular connection was emulated using a 15 metre CAT 5 Ethernet cable, the speed of the link was restricted to a maximum rate of 6 Mbps using the WonderShaper bandwidth throttling utility. The DASH IF Spring 4K test Stream was streamed and viewed using the DASH IF Reference Client 2.4.1 in a Firefox browser. AIS was not implemented.

Stage 1 Test Scenario 2: The DASH IF Spring 4K test Stream was streamed over a Wi-Fi connection, the wireless link was throttled to 4.4 Mbps (mean – 2 STD), a link speed informed by 5.9.2. The content was viewed using the DASH IF Reference Client 2.4.1 in a Firefox browser. AIS was not implemented.

Stage 1 Test Scenario 3: The DASH IF Spring 4K test Stream was streamed over a Wi-Fi connection, the wireless link was throttled to 2.8 Mbps which reflected the frequent reductions in Wi-Fi speed experienced during real world streaming (Chapter 4, Section 4.9.2). The video content was viewed using the DASH IF Reference Client 2.4.1 in a Firefox browser. AIS was not implemented.

Stage 1 Test Scenario 4: The DASH IF Spring 4K test Stream was streamed over a Wi-Fi connection, the wireless link was further throttled to 1.8 Mbps to reflect the frequent reductions in Wi-Fi speed experienced during real world streaming (Chapter 4, Section 4.9.2). The video content was viewed using the DASH IF Reference Client 2.4.1 in a Firefox browser. AIS was not implemented.

Stage 1 Test Scenario 5: In this test both the emulated cellular network connection and the throttled Wi-Fi connection were employed. The emulated cellular connection was throttled at 6 Mbps and the Wi-Fi connection was throttled at 2.8 Mbps. The value of 2.8 Mbps was selected as a reasonable representative value for degraded Wi-Fi connectivity based on practical observations (Chapter 4, Section 4.9.2). The streaming of the content was initiated over the emulated cellular connection and after a period of 20 seconds the restricted Wi-Fi link was activated and the emulated cellular connection was shut-down. No further interventions were carried out.

Stage 1 Test Scenario 6: Both the emulated cellular connection and the restricted Wi-Fi link were used in this test. Streaming of the video content was initiated over the emulated cellular link and after a period of 20 seconds the Wi-Fi interface was activated and the emulated cellular connection was deactivated. In this test AIS was implemented and whenever the Wi-Fi link was unable to maintain the video stream on its own the emulated cellular interface was reactivated.

6.5.4 Stage 1 Test Results

Table 20 presents the results of the various test scenarios described above. In Test 1 an emulated cellular link with a restricted maximum bandwidth of 6 Mbps was used to stream content from the DASH Industry Forum website. Based on the available bandwidth of the link the DASH-enabled client selected segments having a bit rate of 4.873 Mbps, the average observed level of the playout buffer was greater than 16 seconds. No stalling events were observed during playback and the actual duration of the video content matched the stated duration of 2 min 45 sec.

In Test 2 a Wi-Fi connection, restricted to 4.4 Mbps was used to stream from the DASH IF website. The segment bitrate selected by the DASH-enabled client in this case was 2.859 Mbps, the reduction in segment bitrate being due to the reduction in link capacity. The average observed level of the playout buffer was 7 seconds, and no stalling events occurred during video playback. Again the actual duration of the video playback matched the stated duration of 2 minutes 45 seconds.

The purpose of Test 1 and Test 2 was to determine the number of stalling events during the video stream under optimal conditions i.e. maximum link capacity specific to the test, stationary device and non-shared network capacity.

Table 20 Stage 1 Stalling Events and Bitrates

Test Num	Scenario Description	Link Speed [Mbps]	Avg. Quality	No. of Stalling Events	Resulting Delay in Seconds	Reference Duration [Minutes]	Observed Duration	Segment Bitrate [Mbps]
1	Simulated Cellular Connection	6	1080p@5Mbps	**0**	0	2:45	2:45	4873
2	Wi-Fi	4.4	720p@3Mbps	**0**	0	2:45	2:45	2859
3	Wi-Fi (Degraded)	2.8	720p@3Mbps	**35**	31	2:45	3:16 (+31 sec)	2859
4	Wi-Fi (Degraded)	1.8	720p@3Mbps	**37**	135	2:45	5:00 (+ 135 sec)	2859
5	**LTE/Wi-Fi No Intervention**	6 2.8	1080p@5Mbps 720p@3Mbps	**34**	**32**	2:45	**3:17 (+ 34 sec)**	4873 2859
6	**LTE/Wi-Fi Intervention**	6 2.8	1620p@3Mbps 720p@3Mbps	**1**	**5**	2:45	**2:50 (+ 5 sec)**	4873 2859

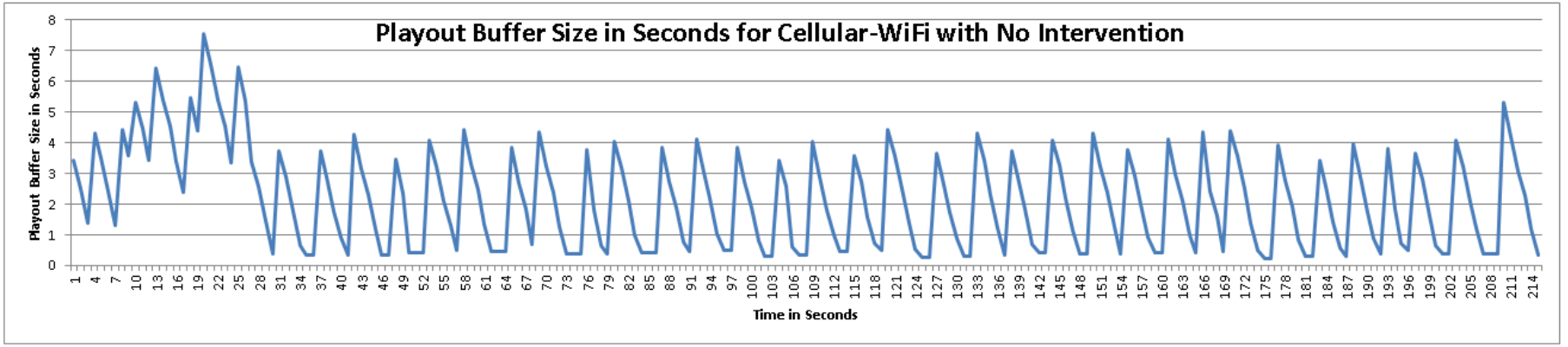

Figure 47 Stage 1 Playout Buffer Level in Seconds for Cellular-Wi-Fi Test with No Intervention

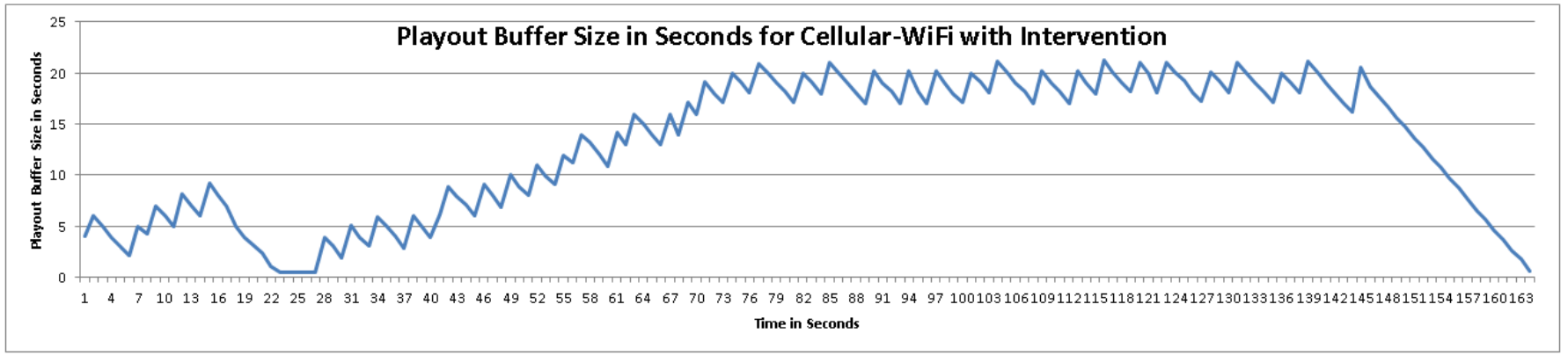

Figure 48 Stage 1 Playout Buffer Size in Seconds for Cellular-Wi-Fi Test with Intervention

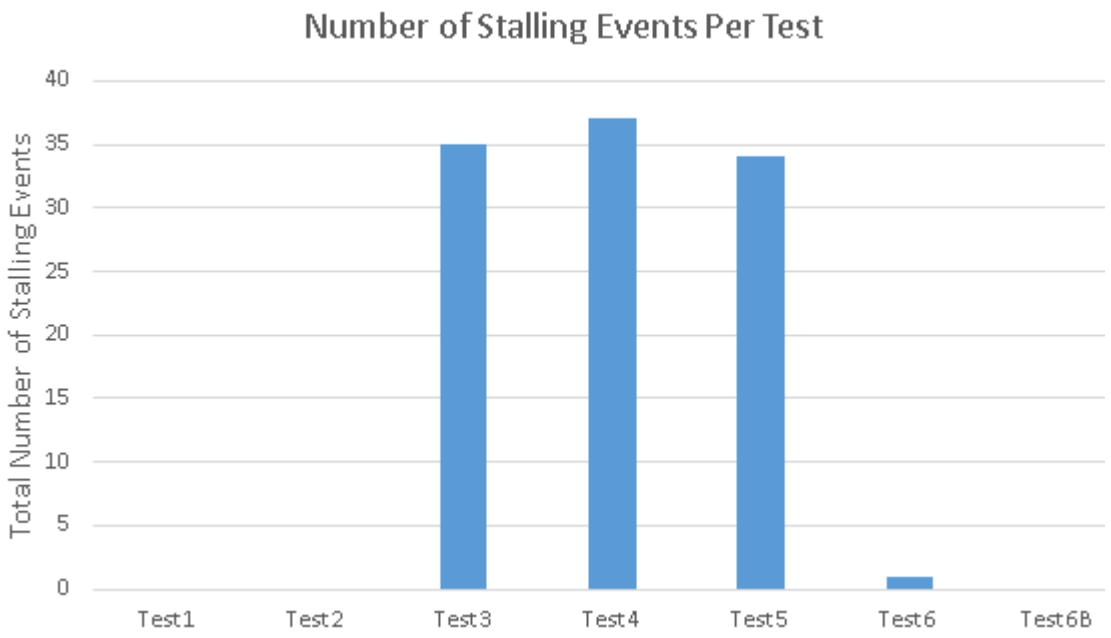

Figure 49 Number of Stalling Events per Test

Test 3 and Test 4 were concerned with examining the number of stalling events that might occur if the Wi-Fi connection was operating at sub-optimal speeds. In Test 3 the Wi-Fi link was restricted to a speed of 2.8 Mbps and in Test 4 the Wi-Fi link speed was further reduced to 1.8 Mbps. As can be seen from Table 20 under these conditions a large number of stalling events occurred during which the video stream was almost unwatchable. The primary difference between the two sets of stalling events is that each individual stalling events observed during Test 3 were of far shorter duration than individual stalling events observed during Test 4 due to the higher wireless link speed of Test 3.

Test 5 examined the case in which both the emulated cellular connection and the restricted Wi-Fi connection were employed. The emulated cellular link speed was maintained at 6 Mbps and the Wi-Fi link was restricted to 2.8 Mbps. In this test the initial connection to the media server was via the emulated cellular link. When the connection was established and the video content had begun streaming the Wi-Fi link was activated and the emulated cellular link was disabled. No other interventions were carried out and as expected the video stream began experiencing stalling events shortly after switching to the Wi-Fi link. As was the case with Test 3 and Test 4 the observed number of stalling events was high.

Test 6 examined the case in which both the emulated cellular connection and the restricted Wi-Fi connection were employed. The emulated cellular link speed was maintained at 6 Mbps and the Wi-Fi link was restricted to 2.8 Mbps. However, in this instance AIS was implemented and interventions were carried out as required. The initial streaming of the video content was established over the emulated cellular connection and when the video streaming commenced the Wi-Fi link was initialized and the emulated cellular link disabled. However, in this test when the level of the playout buffer approached 2 seconds the emulated cellular interface was re-activated. As can be seen from Table 20 there was a significant reduction in the number of stalling events that occurred. Stalling events, although greatly reduced in number, did occur and had a far longer duration than had been observed in previous tests. The stalling event in Test 6, which occurred at approximately 25 second into the simulation, lasted for 5 seconds and no other stalling events were observed for the remainder of the test. It was not immediately clear what caused the single stalling event in Test 6 and we investigated the matter by repeating the test (designated Test 6B) with an increased target buffer level of 10 seconds. No stalling events of any duration occurred during Test 6B and with no new evidence coming to light as to the cause of the stalling event in Test 6 it remains undetermined**.** The occurrence of a single stalling event does not impact on the effectiveness of AIS since one of the aims of AIS is to reduce the number of stalling events experienced by the user and clearly the number of stalling events has been reduced.

6.5.5 Stage 1 Analysis

The results of Test 1 and Test 2, presented in Table 20 and illustrated in Figure 49, show that under optimal conditions (single user, bandwidth not shared) no stalling events occur and no delays were experienced during video playback.

Results from Test 3 and Test 4, presented in Table 20 and detailed in Figure 49, show that in the test environment a single user employing only Wi-Fi with reduced capacity will experience stalling events. In addition to multiple stalling events, the time required to stream the video content will be significantly increased due to retransmission of packets and greater amount of time is required to fill the playout buffer. The increase in the amount of time taken to stream the entire video also results in an increase in energy consumption.

In Test 5 both the emulated cellular link (6 Mbps) and the Wi-Fi connection (2.8 Mbps) were used. The video stream was initialized using the emulated cellular link and when the video

stream had been established the test device activated the Wi-Fi interface and de-activated the emulated cellular interface. The emulated cellular interface then remained inactive for the remainder of the test. A high number of stalling events were observed during this test and the amount of time taken to complete streaming increased by 32 seconds from 2 minutes 45 seconds to 3 minutes 17 seconds.

For Test 6 the emulated cellular connection was used to initialise the video streaming session, and once the streaming session was established the Wi-Fi interface was activated and the emulated cellular interface was deactivated. In contrast to Test 5 AIS was implemented and the emulated cellular interface was reactivated when the deliberately reduced capacity of the Wi-Fi link proved inadequate to maintain an uninterrupted video stream. A single stalling event occurred during the streaming of the video, this stalling event lasted for 5 seconds and was of significantly longer duration than previously observed stalling events.

Figure 47 depicts the level of the playout buffer in seconds for the DASH-enabled client over the duration of the video streaming session in Test 5. Playout buffer level readings were taken every second in addition to a count of the stalling events. During the first 26 seconds of the test in which the emulated cellular connection was established the playout buffer size was in excess of 1 second and no stalling events occurred. When the Wi-Fi interface was activated and the emulated cellular interface deactivated the playout buffer began to deplete faster that it was being replenished resulting in the buffer size continually falling below 1 second. During this period of the test multiple stalling events were observed and the number and frequency of the stalling events was such as to make the video stream all but unwatchable.

Figure 48 depicts the playout buffer levels in seconds for Test 6 with the buffer size being recorded once a second for the duration of the test. During the initial phase of the test the emulated cellular connection was used exclusively and the buffer level remained above 1 second with no stalling events occurring. Following activation of the Wi-Fi interface and the deactivation of the emulated cellular connection the buffer level fell below 1 second. The cellular interface was reactivated to support the Wi-Fi interface during which time a single stalling event of approximately four or five seconds occurred. The remainder of the streaming session completed without any further stalling events.

6.5.6 Testing Stage 2

The Stage 2 test environment was deployed on a Lenovo ThinkPad laptop equipped with an Intel i7 processor and 16GB of RAM running the Debian 9 Linux [167] operating system. Two Oracle VM Virtualbox [168] Debian virtual machines (VMs) were installed. One VM contained a modified version of the DASH Industry Federation JavaScript Reference Player [169]. The second VM had an Apache web server installed to host the media files.

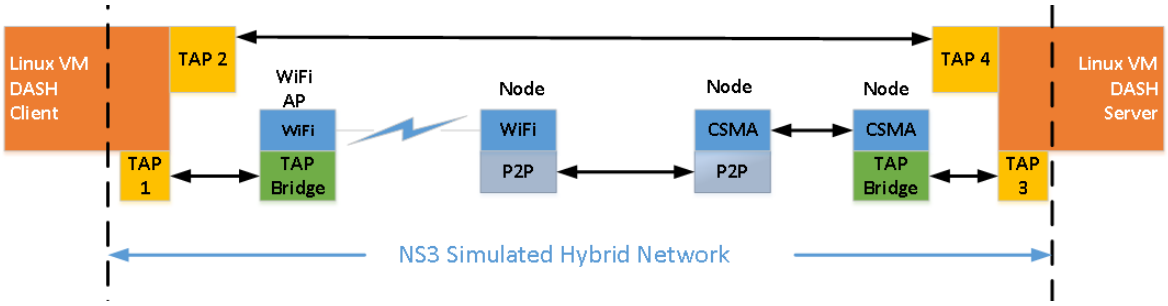

Figure 50 Stage 2 Combined VM and NS-3 Testing Environment

Figure 50 depicts the components of the Linux VM/NS-3 simulation environment. Two Linux VMs, the left hand VM containing the DASH enabled client and the right hand side VM containing the DASH content server, connect to the NS-3 simulated network through TAP devices implemented in software. These TAP software devices enable external entities to establish a connection to an NS-3 environment and send data traffic across the NS-3 virtual links. The TAP 2 interface on the DASH client VM is connected to the TAP 4 interface on the server VM by a CBR link that emulates a cellular connection. On the client VM the TAP 1 interface provides a connection to a TAP interface attached to a node on the NS-3 network that emulates a Wi-Fi AP. The internal NS-3 network that TAP1 and TAP 3 connect to is a mixture of wireless and fixed line links.

The DASH enabled player, which is written in JavaScript and runs inside a browser, requested content from the Apache server running on the second VM. A mixed wireless/fixed Network Simulator 3 simulated network provided connectivity between the DASH player and the server (Figure 50).

Media Source Extensions (MSE) [170] as implemented by the browser are responsible for rendering the video content. Media Source Extensions is a W3C specification, it enables JavaScript to send byte streams to media codecs within a web browser that supports HTML5

video and audio. MSE also supports the implementation of client-side pre-fetching and buffering entirely in JavaScript. The video streaming session begins when the player loads the manifest file, known as the Media Presentation Description (MPD), which among other information such as segment location, describes the video segment characteristics. This description includes video length, segment length and encoded bitrate.

6.5.7 Comparative Solutions

The proposed AIS system was evaluated against the three ABR algorithms included by default in the DASH IF Reference player, these ABR algorithms are outlined in the following section.

The default ABR strategies are BOLA-E [171], DYNAMIC [172] and THROUGHPUT [173]. The performance of AIS was evaluated against the default ABR strategies in terms of the total number of stalling events and average buffer levels during playback. NS-3 links were loaded during testing with a constant bitrate traffic to model network congestion for mobile users. The video used for testing was EnvivioDASH3 from [37] which contained multiple HD bitrate segments with a total duration of 194 seconds. The video depicts a variety of characters in elaborate makeup and costumes. Segment bitrates ranged from a minimum 2 Mbps (3220x180) to a maximum 5.3 Mbps (1920x1080). Each of the ABR algorithms was evaluated using this video content.

The BOLA-E strategy selects the bitrate for requested video segments based on the current playout buffer level, with higher bitrate segments requested for higher buffer levels [155]. The THROUGHPUT strategy chooses the bitrate for requested video segments based on the recent throughput history [172]. The DYNAMIC strategy switches smoothly between BOLA-E and THROUGHPUT in real time, in order to exploit the strengths of both [173].

6.5.8 Stage 2 Test Scenarios

The four ABR strategies, BOLA-E, THROUGHPUT, DYNAMIC and AIS were each tested in the following three simulation scenarios.

Stage 2 Test Scenario 1 examined streaming HD content from the Linux VM Apache server over the NS-3 network to the Linux VM DASH enabled client with no additional load placed on the NS-3 network. This simulation scenario was run a total of four times with one of the four ABR strategies being employed each time.

Stage 2 Test Scenario 2 examined streaming HD content from the Linux VM Apache server over the NS-3 network to the Linux VM DASH enabled client with a Constant Bitrate (CBR) load of approximately 30 Mb per second imposed on the NS-3 network. This simulation scenario was run a total of four times with one of the four ABR strategies being examined each time.

Stage 2 Test Scenario 3 considered streaming HD content from the Linux VM Apache server over the NS-3 network to the Linux VM DASH enabled client with a CBR load of 40Mbps imposed on the NS-3 network. This simulation scenario was run a total of four times with one of the four ABR strategies being examined each time.

6.5.9 Stage 2 Video Content Selection

The video content selected for streaming in each of the three scenarios was EnvivioDASH3 [37], the first video segment of this stream meeting the HD criteria of 1280x720 pixels had a bit rate of 2.85 Mbps [173]. This bitrate was selected as the default bitrate for all tests. ABR algorithms were not permitted to reduce the bitrate of the stream in response to changes in link conditions. This was to reflect the assumption that the user wished to view the content in HD only.

6.5.10 Stage 2 Results

Test results for the Stage 2 simulations are presented in Table 21 The table shows the ABR algorithm used, the average buffer level in seconds for each test, network load, percentage of traffic on both Wi-Fi and cellular links, the bitrate of the streamed content and the total number of stalling events.

For the first set of tests no additional load was imposed on the network and all ABR algorithms produced very similar results. A load was introduced onto the network gradually and the results recorded and no significant changes in output were observed until the additional CBR load on the network approached 30 Mbps.

When the imposed network load approached 30 Mbps the number of stalling events experienced by the default ABR algorithms in the reference player increased rapidly (Figure 51). The DYNAMIC ABR experienced 24 stalling events during 194 seconds of video playback, BOLA_E experienced 28 as did the THROUGHPUT ABR algorithm. The proposed solution, AIS experienced no stalling events but did load balance across two interfaces with 63% of the video segments being streamed over the primary interface (cellular) and 37% over the secondary Wi-Fi interface.

Table 21 Simulation results for All AIS Scenarios

ABR Algorithm	Avg Buffer Level in Seconds	Content Bitrate (Kbps)	Network Load	% WiFi	% Cellular	Total Number of Stalls
Scenario 1 - No Additional Load on NS3 Network						
DYNAMIC	17.9	2850	0	100	0	0
BOLA-E	16.6	2850	0	100	0	0
THROUGHPUT	17.9	2850	0	100	0	0
AIS	18.4	2850	0	100	0	0
Scenario 2 - 30 Mbps Additional Load on NS3 Network						
DYNAMIC	1.95	2850	30 Mbps CBR	100	0	24
BOLA-E	1.85	2850	30 Mbps CBR	100	0	28
THROUGHPUT	1.94	2850	30 Mbps CBR	100	0	28
AIS	14.15	2850	30 Mbps CBR	37	63	0
Scenario 3 – 40 Mbps Additional Load on NS3 Network						
DYNAMIC	1.39	2850	40 Mbps CBR	100	0	135
BOLA-E	NA	2850	40 Mbps CBR	100	0	unwatchable
THROUGHPUT	NA	2850	40 Mbps CBR	100	0	unwatchable
AIS	9.73	2850	40 Mbps CBR	34.7	65.3	2

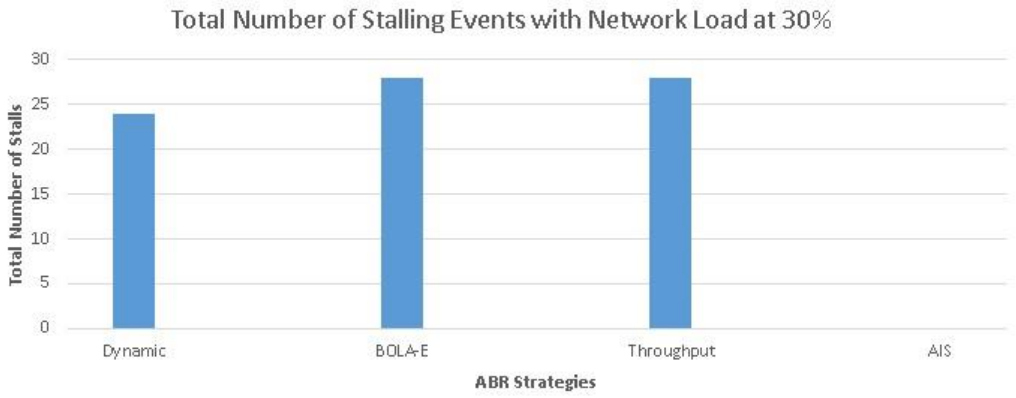

Figure 51 Number of stalling events for each ABR at 30% CBR load on network

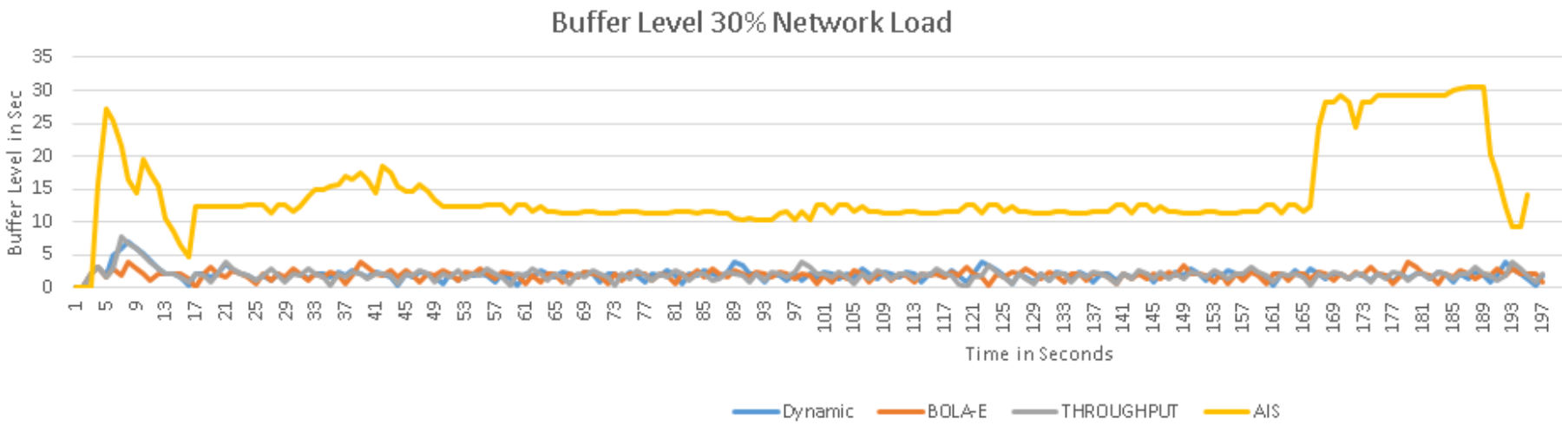

Figure 52 Average buffer levels in seconds events with 30% CBR load on network

The average playout buffer levels when a 30 Mbps CBR load is placed on the network are shown in Figure 52. The three default ABR buffer levels dropped from almost 18 seconds for no additional network load to just under 2 seconds during the 30 Mbps loading. The average playout buffer level for AIS has dropped from just over 18 seconds with no network load to 10 seconds with the network load imposed. Figure 52 also shows a distinctive rise in the AIS playout buffer levels towards the end of the video streaming session. Recall that on initialisation AIS creates a list of video segments to be downloaded during the streaming session. In the event that the Wi-Fi link is not sufficient to maintain buffer levels the cellular link is used to download segments from the front of the list as they will be required immediately and the slower Wi-Fi link is used to download segments from the back of the list. This strategy of downloading the video segments from the front and end of the segment list in parallel results in high playout buffer levels towards the end of the stream. The impact of this strategy is that no stalling events occur at the end of the session during which they would have the greatest negative impact on QoE.

Figure 53 depicts the number of stalling events experienced when a 40 Mbps CBR load is placed on the NS-3 network with only two ABRs represented: DYNAMIC and AIS. At this load level both the BOLA-E and the THROUGHPUT ABRs had so many stalling events that they were unwatchable. Of the three default ABRs examined the DYNAMIC ABR performed best but did experienced 135 stalling events during video playback.

Figure 53 Number of stalling events with 40% CBR load on network

AIS performed in much the same way as it did with a 30 Mbps load on the network. It experienced two stalling events during the test and received 65.3% of video segments over the primary cellular link and 34.7% over the secondary Wi-Fi link. Figure 54 depicts the average buffer levels for both DYNAMIC and AIS, again we can see a rapid increase in the playout buffer levels towards the end of the session for AIS as previously described.

ABR schemes are designed to operate with whichever interface is in use be it a Wi-Fi or cellular interface. Performance of the ABR scheme suffers if the link in use suffers from network latency, ABR schemes do not have the functionality to control wireless interface s or switch between active interface at will. AIS, on the other hand is designed to use multiple interfaces if available and can reduce the impact of latency on one link by load balancing on another link with lower latency. In general, ABR schemes seek to maximise QoE by requesting the segments with the highest bitrate that it estimates the link in use can support, this approach worsens the impact of latency by taking longer to download segments. In an effort to reduce the amount of data downloaded over cellular networks AIS requests the segments having the minimum bitrate that satisfies the requirements for HD video thereby reducing load on the links.

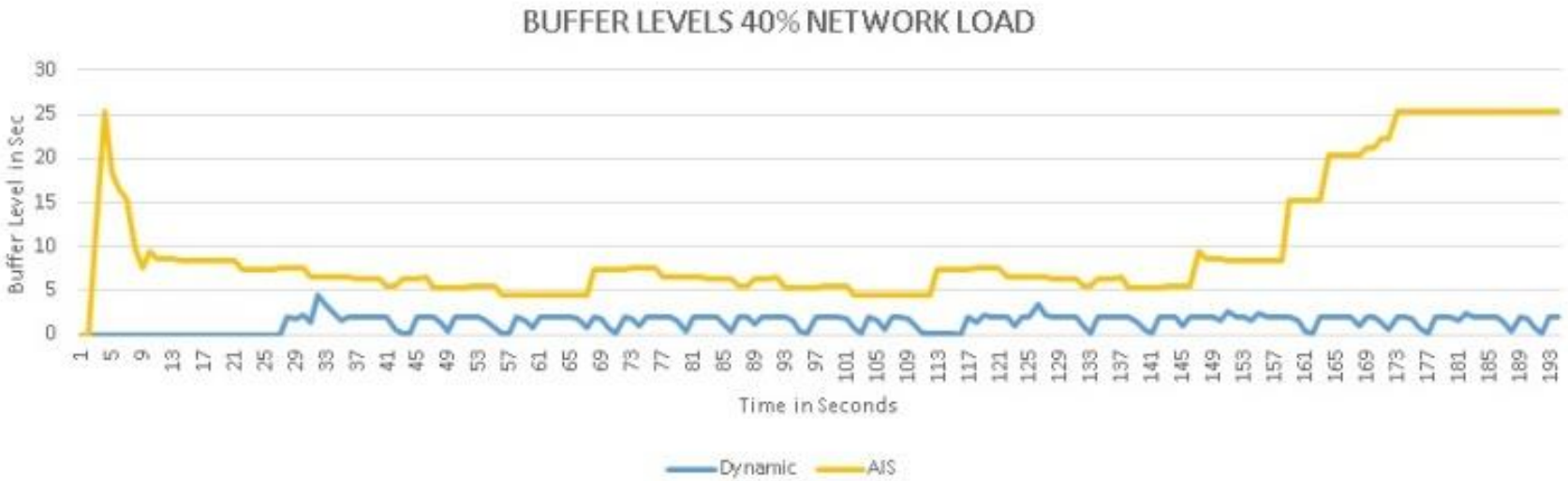

Figure 54 Average buffer levels in seconds with 40% CBR load on network

6.6 Conclusions

Many ABR algorithms have been developed to deal with the problem of fluctuating bandwidth in wireless networks in order to improve end-user QoE. This is achieved by varying the bitrate of requested video segments in response to changes in available bandwidth in order to avoid stalling events in the video playback. When the available bandwidth drops the ABR algorithm will request lower bitrate and therefore lower quality segments in response to the reduction in bandwidth. In the case of HD video, however, there is a lower bound to the bitrate below which the requested segments cannot fall if they are to be considered high definition.

Standard ABR strategies such BOLA-E, THROUGHPUT and DYNAMIC are successful at matching the size of the video segments that they request to the bandwidth available but fail at maintaining a HD video stream under fluctuating bandwidth conditions. This is due to the fact they have no control over the initialization and use of available wireless interfaces, they are restricted to using the current wireless interface in operation.

The proposed solution, AIS, makes use of the multi-homed nature of modern mobile devices and the segmented nature of video content formatted for use by DASH enabled clients to address the problem. AIS enables the streaming of HD video using multiple links by adapting the selection of the interfaces to match the required bitrate rather than adapt the bitrate to match the available bandwidth. AIS reduces stalling events, especially in the latter half of the streaming session where stalling events would have the greatest impact on QoE while also helping to protect the user's cellular network data-cap. The overall amount of data downloaded over the cellular interface is reduced by offloading data transfers to Wi-Fi whenever possible.

This proposed approach is not without its drawbacks: the use of multiple interfaces has a detrimental impact on energy consumption and in energy constrained devices this is not to be taken lightly. However, everything in mobile communications is a compromise, in this case the trade-off is between a reduction in stalling events during the streaming of video content in heterogeneous wireless environments with its consequent reduction in user QoE and an increase in energy consumption. The increase in energy consumption is due to the need for multiple wireless interfaces to be active simultaneously. Although not quantified in our tests, the additional time required to stream the test video content with its attendant need

for retransmission (Stage 1: Test Scenario 3, Test Scenario 4 and Test Scenario 5) would also result in increased energy consumption. The proposed strategy clearly demonstrates that it is possible to maintain video quality while reducing stalling events through the use of a solution which is easily deployable using existing components without the need for any alterations to equipment, servers or protocols.

AIS fulfils Thesis Objectives 3 and 4 introduced in Chapter 1, Section 1.6 and reproduced below:

3 Enable a mobile user to maximise Quality of Experience while streaming HD video
4 Reduce the overall amount of data transferred over cellular connections

AIS maintains QoE by reducing the number of stalling events, enabling the streaming of HD video over bandwidth constrained links and removing changes in bitrates between video segments. It also reduces the amount of data transferred over cellular links by reducing segment sizes to the lowest resolution that meets the requirements of HD video. Further reductions in the amount of data received at the mobile node are achieved by downloading as much data as possible over Wi-Fi.

7 CONCLUSIONS

7.1 Overview

Many urban dwellers rely solely on mobile devices such as smartphones to access data and services on the Internet. These devices are frequently equipped with multiple wireless interfaces that enable the user to take advantage of the opportunities presented by the heterogeneous, multi-network wireless environments in which they operate. Pay-as-you-go subscriptions with mobile service providers that have fixed data allowances are a very popular option amongst users and these fixed data allowances or data-caps are usually valid for the duration of the pay-as-you-go subscription which is typically 28 days. If the subscriber exceeds their data cap before the end of the 28-day subscription period, they face very high excess data charges or disconnection from data services. This is problematic for these users as much of the content they access is video which has the potential to rapidly consume a subscriber's data-cap. As network speeds and capacity grow through the deployment of systems such as 5G the situation will worsen for users. Data-caps will remain but the rate at which they are depleted will increase through downloading SD and HD video content at grater speeds. This will motivate users to seek ways to reduce the amount of data downloaded over cellular networks in a bid to protect their data allowance. Although users want to protect their data-caps they are also interested in maintain their QoE while viewing video content on their mobile devices, one important challenge is how to maintain or enhance the users QoE while protecting their data-cap.

7.2 Contributions

It is within this context that this thesis presents a system to address these issues by proposing three major, integrated mechanisms:

(1) Scan-Or-Not-to-Scan (SONS)

The novel Scan-Or-Not-to-Scan (SONS) framework decides, based on user remaining data-cap and device sensors when it is appropriate for the user to conduct network detection scans and execute network selection strategies and when it is not. The use of the SONS framework enables the user to conserve energy by shutting down unused interfaces and minimising the number of unnecessary scans for Wi-Fi APs. By reducing both

the number of connection events in which little or no data can be received over Wi-Fi and the number of unnecessary handovers SONS helps maintain the mobile user multimedia QoE at high levels. SONS employers the user's remaining data cap (RDC) as a significant input in its decision making process. The SONS utility score determines when the device's Wi-Fi interface should be activated and SONS Resolution Discovery Module decides how the video stream is to be treated. To the best of our knowledge SONS is the first scan or no scan decision making process that considers the user's data cap.

(2) MDF

The MPEG-DASH-based Framework (MDF), is a novel technology agnostic framework that seeks to optimise the performance of MPEG-DASH enabled clients streaming Standard Definition (SD) video in urban HetNet environments. MDF employs the SONS mechanism to determine when conditions are suitable for a user to attempt to switch their point of attachment from a cellular network to an alternative Wi-Fi networks. Attempting to connect to alternative networks only when a reasonable possibility of establishing a useful connection exists reduces disruption to connectivity and also improves QoE by reducing stalling events. MDF also matches the requested video segments with both connection type and device capabilities (e.g. screen size) to avoid downloading segments of a higher definition than can be utilised, this helps to protect the user's data-cap. The data-cap is also protected by downloading as little data over the cellular connection as is practicable.

(3) Adaptive-Interface-Selection (AIS)

Adaptive Interface Selection (AIS) addresses the issue of how to download fixed bitrate video segments over bandwidth constrained links. AIS leverages the multi-homed nature of modern mobile devices and the segmented nature of video content prepared for DASH-enabled clients to overcome the bandwidth constraints that hobble ABR algorithms when streaming HD content. When necessary AIS uses multiple wireless interfaces in parallel to download content to maintain the playout buffer level to prevent stalling events. By minimizing the number of stalling events and streaming video content at HD compatible bitrates end-user QoE is maintained. AIS also seeks to download those video segments that are of the lowest HD compatible resolution. This strategy aids in

maintaining playout buffer levels and downloading as little content as possible over mobile networks helps protect the end-user's mobile data-cap.

7.3 An Illustrative Example of an Integrated System

Figure 55 presents an example of a system in which SONS, MDF and AIS are integrated. It shows, at a high level the relationship between the system components, the DASH enabled client and the wireless interfaces and it provides an overview of the communication flows between them. The MPD-Cache directory is not show here for clarity.

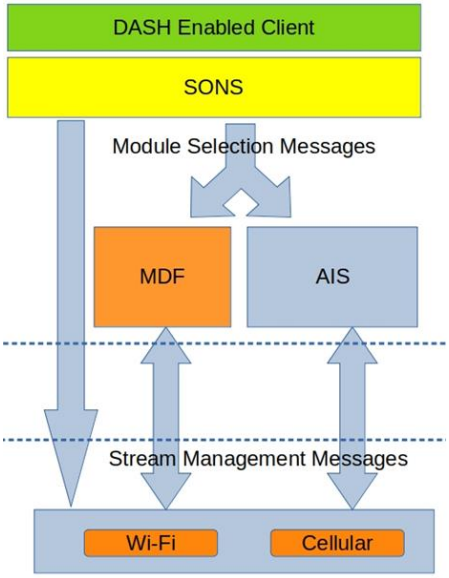

Figure 55 Integrated System

Figure 56 presents a typical travel HetNet environment through which an urban commuter might travel to work or college. Bob, our commuter wants to view video content on his smartphone during his commute to work and on his journey Bob will move through a HetNet comprised of mobile phone networks and individual, autonomous Wi-Fi networks. These independent networks are essentially islands of Wi-Fi connectivity within the cellular network coverage areas. The rate at which Bob travels varies greatly and depends on his mode of transport and the prevailing traffic conditions and occasionally Bob will stop to have a coffee or some food. Like a great many other young commuters Bob has a pay-as-you-go account with his service provider and the account has a fixed data allowance for each billing period. Bob wants to maximise his enjoyment of the video content that he streams to his device while protecting his account's data-cap. He protects his data allowance by

downloading as little data as possible over mobile networks to avoid depleting his allowance before the end of the billing period. Due to the constrains imposed by the limited capacity of his device's battery he also wishes to conserve energy whenever possible.

As he moves through the wireless HetNet Bob will have many potential opportunities to offload his data transfers from cellular to Wi-Fi. However, as a result of the natural changes in Bob's speed, direction of travel and the variations in signal strength and coverage of the wireless networks offloading data successfully is very challenging. The difficulties for users such as Bob are deciding on when to scan for available networks, deciding on the most appropriate level of video quality to stream in order to maximise QoE and deciding on when, if ever, to switch their point of attachment to an alternative network.

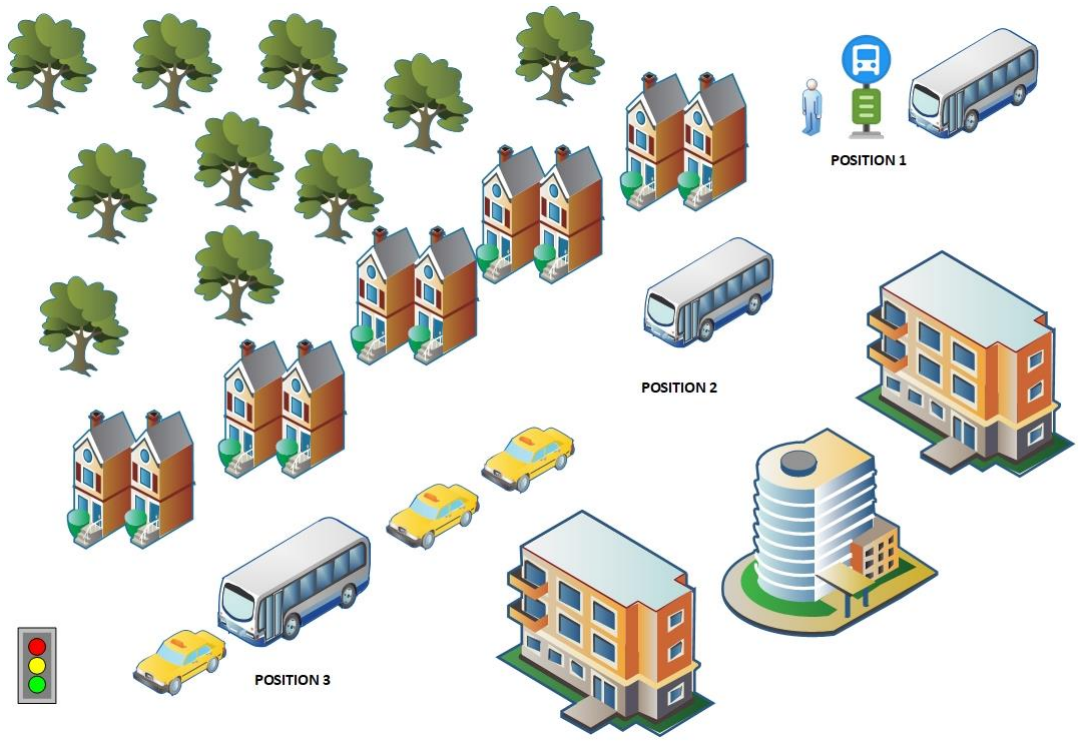

Figure 56 Bob's Commute

7.3.1 Illustrative Example

Figure 7-2 depicts various locations along Bob's route and they will be used as reference points during the description how the integrated system might operate as follows:

Position 1: Bob arrives at the bus stop and has to wait for the bus to arrive, to pass the time while he waits Bob decides to watch some videos. His device is currently connected to the cellular network of his mobile phone provider and several Wi-Fi networks are available at his location but initially the Wi-Fi interface is disabled to conserve energy. He selects a video stream for viewing and begins streaming the video using a DASH enabled client over his cellular connection.

It has been demonstrated in Chapter 4 that automatic scanning for and connecting to Wi-Fi networks by mobile devices can be detrimental to the user's energy consumption and data transfer rates. An automated scanning decision mechanism, SONS, is proposed. SONS runs in the background as a service on Bob's device. SONS uses multiple inputs such as the user's speed over the ground and remaining data-cap to decide when conditions are suitable for scanning for nearby Wi-Fi networks.

When SONS determines that conditions are such that a useful connection to a Wi-Fi network could be established by calculating a utility score equal to 1, it activates the Wi-Fi interface and invokes Bob's network detection and selection algorithm of choice. If one or more potential candidate Wi-Fi networks are detected Bob's preferred network selection algorithm chooses a Wi-Fi AP to connect to. When the connection has been established to a Wi-Fi Access Point (AP) the cellular data link is disconnected. The SONS MPD-Cache Monitor (MCM) checks the MPD-Cache directory for downloaded MPD or manifest files, when an MPD file is detected the MCM processes the file to determine if the target video is SD or HD. While the MPD file is beings processed the initial video segments are downloaded into the play-out buffer over which ever interface is active.

If the processed MPD file shows that the target video is in SD the MDF module is initialised to handle the video stream management. On the other hand, if the requested video stream is in HD the MCM module passes control of the stream to the AIS component of the system. In this scenario Bob starts watching his chosen video which is a Standard Definition (SD) stream and after a short period of time his bus arrives.

Position 2: Bob pauses the video and boards the bus to continue his journey. The SONS module detects that Bob is travelling at a speed greater that the speed at which a useful Wi-Fi connection can be maintained and re-establishes a data connection over the cellular network to stream the video. The Wi-Fi interface is de-activated, video segments are downloaded over the cellular interface and SONS enters 'Bus/Train Mode'.

Many urban public transport systems have on-board Wi-Fi networks that provide Internet access to passengers and the SONS 'Bus/Train Mode' addresses this scenario. When in Bus/Train Mode SONS re-activates the UE's Wi-Fi interface after a period of time equal to 60 seconds and scans for available Wi-Fi APs. A duration of 60 seconds was selected as the vehicle is assumed to have passed beyond the coverage area of the AP to which Bob had been connected in this time. This strategy will avoid the problem of establishing a connection to the original AP while seated on the bus only to lose the connection when the bus departs from the stop. When the scan has been completed SONS stores the details of any detected APs and shuts down the Wi-Fi interface in order to avoid automatic scanning and to conserve energy. After a period of 90 seconds SONS reactivates the Wi-Fi interface and scans for available APs, on completion of the scan SONS compares the SSIDs of any detected APs with the SSIDs of any APs recorded during the previous Bus/Train Mode scanning operation. If an SSID is detected in both scan results it is considered to be an on-board AP and an attempt is made to establish a connection to it. When a connection to the on-board Wi-Fi network is established the cellular connection is closed. All subsequent video segments are streamed over the Wi-Fi interface while the connection is stable.

Point 3: The bus enters rush hour traffic and slows down, Bob decides to begin watching a newly released feature film in HD. When the videos MPD file is downloaded SONS checks the file and determines that HD video content is to be downloaded and switches management of the stream from the MDF module to the AIS module.

AIS manages the new video stream and downloads the segments with the lowest bitrate that qualify as HD. The proposed system monitors the play-out buffer level as a proxy for throughput on the wireless link [129], [139]. AIS can download the video content over the cellular connection only, the Wi-Fi interface only or it can download segments over both interfaces simultaneously depending on the available networks. The primary aim of the AIS module is to maintain a HD video stream for the user while protecting the user's data cap as much as possible. The amount of data downloaded over the cellular connection is minimised

by only selecting the segments with the lowest bitrate that still meet the requirements for HD video. Using Wi-Fi when available to either replace the cellular connection or supplement it helps to further protect the data cap. This illustrative example shows how the various components might work together to optimise Bob's QoE by reducing stalling events and sudden switches in video quality, protect Bob's data-cap by offloading data transfers to Wi-Fi whenever possible and selecting the smallest segments that meet his requirements and conserves energy by shutting down wireless interfaces that are not in use.

7.4 Thesis Objectives

Of the thesis objectives proposed in Chapter 1, Section 1.6, all except Objective 5 have been fully achieved. While full completion of Objective 5 has been elusive the initial calculations regarding the Wi-Fi interface have been completed in Chapter 5, Section 5.9

Fulfilment of the 5 thesis objectives proposed in Chapter 1, Section 1.6 are considered below. Recall that this thesis set out a number of specific objectives as follows:

1. To develop a decision making process that takes into consideration a user's remaining data-cap
2. Enable a mobile user to maximise Quality of Experience while streaming SD video
3. Enable a mobile user to maximise Quality of Experience while streaming HD video
4. Reduce the overall amount of data transferred over cellular connections
5. Reduce energy consumption by deactivating wireless interfaces when possible

1. The novel SONS utility function described in Chapter 4 employs the users remaining data-cap (RDC) as an input into its calculation of a utility score for use in triggering user defined network detection and selection algorithms. Development of the SONS utility function achieves Objective 1.

2. MDF (Chapter 5) reduces the number of stalling thus maximising QoE in the context of streaming SD video content and meets Objective 2.

3.AIS (Chapter 76 achieves Objective 3 by never permitting the DASH enabled client to access video segments that do not meet HD video requirements with regard to resolution

i.e. segments must be at least 720p. Minimises stalling events by managing playout buffer levels.

4. AIS (Chapter 6) and MDF (Chapter 5) help protect the user's data-cap by attempting to match requested segment bitrate to device screen size and by minimising where possible data downloads over cellular links thereby meeting Objective 4.

5. Both AIS and MDF help reduce energy consumption by shutting down wireless interfaces when not in use thereby meeting Objective 5.

7.5 Future Work

The work presented in this thesis is primarily focused on proposing the SONS, MDF and AIS concepts as viable solutions in aiding urban commuters to maintain QoE and to protect their data-caps while streaming DASH enabled video content. This work is presented in the context of commuters moving through urban heterogeneous, multi-network wireless environments. However, SONS, MDF and AIS are of a heuristic design based primarily on modelling and simulation in software of wireless technologies. To move further towards the goal of maintaining QoE while protecting the subscriber's data-cap real world implementation and testing of the components is required in the near future. This would enable perceptual testing to take place with the results of the tests being compared with the results of the simulations conducted during the course of this work. Additionally, the impact of wide spread adoption by urban commuters of these components on other HetNet users should also be examined.

An interesting research direction would be the use of Machine Learning to predict wireless conditions for commuters. Commuters in urban areas travel the same routes, at the same times and at approximately the same speeds every working day. Consistency in behaviour provides an opportunity to gather data on wireless conditions along the commuter's route at regular positions throughout the year. This would allow a model to be built that could determine when to connect to a particular network and when to avoid doing so. The use of an online database would enable other commuters to build their own models without the need to conduct extensive scanning operations themselves.

BIBLIOGRAPHY

[1] Cisco Systems, “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2018-2023”, Whitepaper, [Online], Available https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html, [Accessed March 7th 2020]

[2] OpenSignal, “Germany's rural 4G users still spend one-fourth of their time on 3G and 2G networks”, Article, [Online], Available https://www.opensignal.com/blog/2019/06/13/germanys-rural-4g-users-still-spend-one-fourth-of-their-time-on-3g-and-2g-networks [Accessed July 2019]

[3] Opensignal, “The State of Mobile Video”, Report, [Online], Available https://www.opensignal.com/reports/2019/11/state-of-mobile-video-2019 [Accessed December 2019]

[4] Opensignal, “Smartphone Users in Melbourne and Sydney experience 4G Download Speeds 15 Mbps Slower During Busy Periods”, Article, [Online], Available https://www.opensignal.com/blog/2019/07/05/smartphone-users-in-melbourne-and-sydney-experience-4g-download-speeds-15-mbps-slower-during-busy [Accessed November 2020]

[5] Pew Research Center, “Mobile Technology and Home Broadband 2019”, Report, [Online], Available https://www.pewresearch.org/internet/2019/06/13/mobile-technology-and-home-broadband-2019/ [Accessed June 2019]

[6] OpenSignal, “Global State of Mobile Networks (February 2017)”, Report, [Online], Available https://opensignal.com/reports/2017/02/global-state-of-the-mobile-network [Accessed February 2018]

[7] M. Isaac, S. Frenkel, “Facebok Is ‘Just Trying to keep the lights on’ as Traffic Soars in Pandemic”, Article, [Online], Available https://www.nytimes.com/2020/03/24/technology/virus-facebook-usage-traffic.html [Accessed April 2020]

[8] A. Nordrum, “Popular Internet of Things Forecast of 50 Billion Devices by 2020 is Outdated”, Article, [Online], Available http://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecast-of-50-billion-devices-by-2020-is-outdated [Accessed September 2016]

[9] S. Segan, “In Las Vegas, AT&T Pulls Back the Curtain (Slightly) on Its 5G Strategy”, Article, [Online], Available https://www.pcmag.com/news/369475/in-las-vegas-at-t-pulls-back-the-curtain-slightly-on-its [Accessed July 2019]

[10] R. Amadeo, “The Latest Barrier to 5G Speeds? The Summer”, Article, [Online], Available https://arstechnica.com/gadgets/2019/07/the-latest-barrier-to-5g-speeds-the-summer/ [Accessed July 2019]

[11] E. Dahlman, S. Parkvall, J. Skold, P. Beming, “3G Evolution 2nd Edition (HSPA and LTE for Mobile Broadband)”, Chapter 1: Background of 3G Evolution, Academic Press, August 2008, ISBN: 9780123745385

[12] M. Guowang, Z. Jens, et al, ‘Fundamentals of Mobile Data Networks’, Cambridge University Press, ISBN 1107143217, 2016

[13] IEEE, “IEEE 802.11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications" (PDF). (2012 revision). IEEE-SA, 5 April 2012

[14] IEEE, “IEEE 802.11b-1999 - IEEE Standard for Information Technology - Telecommunications and information exchange between systems - Local and Metropolitan networks - Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Higher Speed Physical Layer (PHY) Extension in the 2.4

GHz band “[Online] Available https://standards.ieee.org/standard/802_11b-1999.html [Accessed August 2020]

[15] IEEE, “IEEE 802.11a-1999 - IEEE Standard for Telecommunications and Information Exchange Between Systems - LAN/MAN Specific Requirements - Part 11: Wireless Medium Access Control (MAC) and physical layer (PHY) specifications: High Speed Physical Layer in the 5 GHz band “[Online] Available https://standards.ieee.org/standard/802_11a-1999.html [Accessed August 2020]
[16] IEEE, “IEEE 802.11g-2003 - IEEE Standard for Information technology-- Local and metropolitan area networks-- Specific requirements-- Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Further Higher Data Rate Extension in the 2.4 GHz Band “[Online] Available https://standards.ieee.org/standard/802_11g-2003.html [Accessed August 2020]
[17] IEEE, “802.11n-2009 - IEEE Standard for Information technology-- Local and metropolitan area networks-- Specific requirements-- Part 11: Wireless LAN Medium Access Control (MAC)and Physical Layer (PHY) Specifications Amendment 5: Enhancements for Higher Throughput “[Online] Available https://standards.ieee.org/standard/802_11n-2009.html [Accessed August 2020]
[18] IEEE, “802.11ac-2013 - IEEE Standard for Information technology--Telecommunications and information exchange between systems—Local and metropolitan area networks--Specific requirements--Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications--Amendment 4: Enhancements for Very High Throughput for Operation in Bands below 6 GHz. “[Online] Available https://standards.ieee.org/standard/802_11ac-2013.html [Accessed August 2020]
[19] IEEE, “IEEE P802.11ax - IEEE Draft Standard for Information Technology -- Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks -- Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment Enhancements for High Efficiency WLAN” [Online] Available https://standards.ieee.org/project/802_11ax.html [Accessed August 2020]
[20] IEEE, “P802.11be - Standard for Information technology--Telecommunications and information exchange between systems Local and metropolitan area networks--Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment: Enhancements for Extremely High Throughput (EHT)” [Online] Available https://standards.ieee.org/project/802_11be.html [Accessed August 2020]
[21] IEEE 802.21 Working Group, [Online], Available http://www.ieee802.org/21/ [Accessed June 2017]
[22] IEEE Standard for Information technology—Telecommunications and information exchange between systems Local and metropolitan area networks—Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," in IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012) , vol., no., pp.1-3534, 14 Dec. 2016, doi: 10.1109/IEEESTD.2016.7786995.
[23] IEEE,”Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Sponsored by the LAN/MAN Standards Committee”[Online] Available https://www.academia.edu/6564159/Part_11_Wireless_LAN_Medium_Access_Control_MAC_and_Physical_Layer_PHY_Specifications_Sponsored_by_the_LAN_MAN_Standards_Committee [Accessed June 2018]
[24] ITU, “E.800: Definitions of terms related to quality of service”, [Online], Available https://www.itu.int/rec/T-REC-E.800-200809-I [Accessed September 2016]

[25] K. Brunnstr¨om, S. A. Beker, K. De Moor, A. Dooms, S. Egger, et al. “Qualinet White Paper on Definitions of Quality of Experience”. Qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeting. 2013
[26] ITU, “P.10: Vocabulary for performance and quality of service”, [Online], Available https://www.itu.int/rec/T-REC-P.10 [Accessed October 2016]
[27] M. Fiedler, T. Hossfeld, Tobias, P. Tran-Gia, “A generic quantitative relationship between Quality of Experience and Quality of Service”, IEEE Network, 24(2):36—41, March 2010
[28] IETF, “A Transport Protocol for Real-Time Applications”, RFC3550, July 2003, [Online] Available https://tools.ietf.org/html/rfc3550 [Accessed July 2016]
[29] IETF, “Real Time Streaming Protocol (RTSP)”, RFC2326, April 1998, [Online] Available https://www.ietf.org/rfc/rfc2326.txt
[30] Telestream, “Adaptive Bitrate Encoding”, [Online], Available https://www.telestream.net/telestream-cloud/adaptive-bitrate-encoding.htm [Accessed March 2019]
[31] Bitmovin, “Adaptive Streaming”, [Online], Available https://bitmovin.com/adaptive-streaming/ [Accessed March 2019]
[32] AT&T Video Optimizer, “Adaptive Bit Rate Video Streaming”, [Online], Available https://developer.att.com/video-optimizer/docs/best-practices/adaptive-bitrate-video-streaming [Accessed December 2018]
[33] Metzger F. et al. (2018) Context Monitoring for Improved System Performance and QoE. In: Ganchev I., van der Mei R., van den Berg H. (eds) Autonomous Control for a Reliable Internet of Services. Lecture Notes in Computer Science, vol 10768. Springer, Cham. https://doi.org/10.1007/978-3-319-90415-3_2
[34] AWS, “Latency in Live Streaming” [Online] Available https://aws.amazon.com/media/tech/video-latency-in-live-streaming/ [Accessed August 2020]
[35] ISO, “ISO/IEC 23009-1:2014 Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats” [Online] Available https://www.iso.org/standard/65274.html [Accessed June 2020]
[36] J. Rodriguez, “Fundamentals of 5G Mobile Networks”, June 2015, John Wiley & Sons Inc., ISBN: 978-1-118-86752-5
[37] Qualcomm, “Rising to Meet the 1000x Mobile Data Challenge”, Whitepaper 2012, [Online], Available https://www.qualcomm.com/media/documents/files/rising-to-meet-the-1000x-mobile-data-challenge.pdf [Accessed May 2016]
[38] 4G Americas, “Understanding 3GPP Release 12: Standards for HSPA+ and LTE Enhancements”, February 2015, [Online], Available http://www.5gamericas.org/files/6614/2359/0457/4G_Americas_-_3GPP_Release_12_Executive_Summary_-_February_2015.pdf [Accessed March 2016]
[39] Ericsson, “Ericsson Mobility Report”, June 2016, [Online], Available https://www.ericsson.com/res/docs/2016/ericsson-mobility-report-2016.pdf [Accessed December 2016]
[40] ITU-T, “The Tactile Internet”, Technology Watch Report August 2014, [Online], Available https://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000230001PDFE.pdf [Accessed April 2015]
[41] 5G.co.uk, “Network Slicing” [Online] Available https://5g.co.uk/guides/what-is-network-slicing/ [Accessed August 2020]
[42] Huawei, “Service Based Architecture for 5G Core Networks” [Online] Available https://www.3g4g.co.uk/5G/5Gtech_6004_2017_11_Service-Based-Architecture-for-5G-Core-Networks_HR_Huawei.pdf [Accessed August 2020]

[43] A. Mishra, M. Shin, W. Arbaugh, “An EmpiricalAnalysis of the IEEE 802.11 MAC Layer Handoff Process”, SIGCOMM Comput. Commun. Rev. 33, 2, pp 93-102, Apr. 2003
[44] C. Pei, Z. Wang, Y. Zhao, Z. Wang, Y. Meng, D. Pei, Y. Peng, W. Tang, X. Qu, “Why it Takes so Long to Connect to a WiFi Access Point?”, Cornell University Library, May 2017, [Online], Available https://arxiv.org/abs/1701.02528v3 [Accessed June 2017]
[45] G. Castignani, A.E.A Moret, N. Montavont, “A study of the discovery process in 802.11 networks”, ACM Sigmobile - Mobile computing and communications review, 2011, 15 (1), pp.25-36
[46] S. Shin, A. Singh, H. Schulzrinne, “Reducing MAC Layer Handoff Latency in IEEE 802.11 Wireless LANs”, International Conference on Mobile Computing and Networking, Proceedings of the Second International Workshop on Mobility Management and Wireless Access Protocols, 2004
[47] V. Mhatre, K. Papagiannaki, “Using Smart Triggers for Improved User Performance in 802.11 Wireless Networks”, MobiSys06, Uppsala, Sweden, 2006
[48] Y. Liao, L. Gao, “Practical Schemes for Smooth MAC Layer Handoff in 802.11 Wireless Networks”, Proceedings of the International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM’06), IEEE, Niagara-Falls, Buffalo-NY, 26-29 June 2006
[49] H. Velayos, G. Karlsson, “Techniques to reduce the IEEE 802.11b handoff time”, IEEE International Conference on Communications, Paris, France, 20-24 June 2004
[50] I. Ramani, S. Savage, “SyncScan: practical fast handoff for 802.11 infrastructure networks”, Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), Miami, FL, USA, 2005
[51] A. Gami Niteshkumar, K., Buddhdev Prasann, D. Kaushal, V. Agarwal, “An Advance Approach to Reduce Handover Delay by Reducing Scanning Delay using Media Independent Handover in 802.11”, IEEE International Conference on Computing Sciences, Omaha, NB, USA, 2012
[52] Z. Chang, O. Alanen, E.H. Ong, J. Kneckt, “Enhanced Channel Scanning Schemes for Next Generation WLAN System”, First IEEE International Conference on Communications in China: Wireless Networking and Applications (WNA), Beijing, China, August 15-18, 2012
[53] M. Q. Khan, S. H. Andresen, "An Intelligent Scan Mechanism for 802.11 Networks by Using Media Independent Information Server (MIIS)," 2011 IEEE Workshops of International Conference on Advanced Information Networking and Applications, Singapore, 2011, pp. 221-225, doi: 10.1109/WAINA.2011.26.
[54] F. Kaleem, A. Mehbodniya, K. K. Yen,F. Adachi, "Application of fuzzy TOPSIS for weighting the system attributes in overlay networks," 2012 14th Asia-Pacific Network Operations and Management Symposium (APNOMS), Seoul, 2012, pp. 1-6, doi: 10.1109/APNOMS.2012.6356098.
[55] S. Lee, K. Sriram, K. Kim, et al., “Vertical Handoff Decision Algorithm Providing Optimized Performance in Heterogeneous Wireless Networks”, proceedings IEEE GLOBECOM 2007
[56] T. Bi, R. Trestian, et al., “Reputation-based Network Selection Solution for Improved Video Delivery Quality in Heterogeneous Wireless Network Environments”, International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), London, 2013
[57] 3GPP, “3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; Access Network Discovery and Selection Function (ANDSF) Manage-

ment Object (MO) (Release 14)", accessed December 2016, available at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1077
[58] R. Trestian, O. Ormond, G.-M. Muntean, "Power-Friendly Access Network Selection Strategy for Heterogeneous Wireless Multimedia Networks", IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Shanghai, China, Apr. 2010, pp. 1-5
[59] Ali E. Abbas, "Foundations of Multi-Attribute Utility", University Printing House, Cambridge CB2 8BS, UK, First Published 2018, ISBN: 978-1-107-15090-4 Hardback
[60] O. Ormond, J. Murphy, G. Muntean, "Utility-based Intelligent Network Selection in Beyond 3G Systems," 2006 IEEE International Conference on Communications, Istanbul, 2006, pp. 1831-1836, doi: 10.1109/ICC.2006.254986
[61] Y. Yu, B. Yong, C. Lan, "Utility-Dependent Network Selection using MADM in Heterogeneous Wireless Networks," 2007 IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, Athens, 2007, pp. 1-5, doi: 10.1109/PIMRC.2007.4394328
[62] L. Liang, H. Wang, P. Zhang, "Net Utility-Based Network Selection Scheme in CDMA Cellular/WLAN Integrated Networks," 2007 IEEE Wireless Communications and Networking Conference, Kowloon, 2007, pp. 3313-3317, doi: 10.1109/WCNC.2007.610
[63] P. Kosmides, A. Rouskas, M. Anagnostou, "Network selection in heterogeneous wireless environments," 2011 18th International Conference on Telecommunications, Ayia Napa, 2011, pp. 250-255, doi: 10.1109/CTS.2011.5898929
[64] R. A. Powers, "Batteries for Low Power Electronics", Proceedings of the IEEE, Volume 83, Issue 4, pp 687–693, April 1995
[65] A. Carroll, G. Heiser, "An Analysis of Power Consumption in a Smartphone", Proceedings of the Usenix Annual Technical Conference (USENIXTAC10), 2010
[66] G.P. Perrucci, F.H.P. Fitzek, J. Widmer, "Survey on Energy Consumption Entities on the Smartphone Platform", IEEE Vehicular Technology Conference (VTC Spring), 2011
[67] M. Tawalbeh, A. Eardley, L. Tawalbeh, "Studying the Energy Consumption in Mobile Devices", Procedia Computer Science, Volume 94, 2016, Pages 183-189, [Online], Available http://www.sciencedirect.com/science/article/pii/S1877050916317756 [Accessed January 2017]
[68] R. Kravets, P. Krishnan, "Power Management Techniques for Mobile Communications", Proceedings of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM'98), Dallas, Texas, USA, October 25-30, 1998
[69] LAN MAN Standards Committee of the IEEE Computer Society, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications", ANSI/IEEE Std802.11, 1999 Edition (R2003), Reaffirmed 12 June 2003
[70] S. Chakraborty, "Reducing Energy Consumption of Network Interfaces in Handhelds", [Online], Available www.cs.purdue.edu/homes/li/cs690Z/Outline/srijan.ptt, February 2002 [Accessed May 2018]
[71] E. Shih, P. Bahl, M.J. Sinclair, "Wake on Wireless: An Event Driven Energy Saving Strategy for Battery Operated Devices", The Eight Annual Conference on Mobile Computing and Networking (MOBICOM'02), Atlanta, Georgia, USA, September 23-26, 2002
[72] R. Krashinsky, H. Balakrishnan, "Minimizing Energy for Wireless Web Access with Bounded Slowdown", The Eight Annual Conference on Mobile Computing and Networking (MOBICOM'02), Atlanta, Georgia, USA, September 23-26, 2002
[73] G. Anastasi, M. Conti, E. Gregori, A. Passarella, "Saving Energy in Wi-Fi Hotspots through 802.11 PSM: an Analytical Model", Proceedings of the Workshop on Modeling and

Optimization in Mobile, Ad Hoc and Wireless Networks (WiOPT2004), University of Cambridge, UK, March 24-26 2004
[74] J. Flinn, M. Satyanarayanan, “Energy-aware adaptation for mobile applications”, 17th ACM Symposium on Operating Systems Principles (SO SP ’99), Published as Operating Systems Review, 34(5):48-63, December 1999
[75] L.M. Feeney, M. Nilsson, “Investigating the Energy Consumption of a Wireless Network Interface in an Ad Hoc Networking Environment”, IEEE Conference on Computer Communications (Infocom’01), April 2001
[76] G. Kalic, I. Bojic, M. Kusek, “Energy consumption in android phones when using wireless communication technologies”, Proceedings of the 35th International Convention MIPRO, 2012 , pp. 754 – 759
[77] R. Trestian, O. Ormond, G.-M. Muntean, “On the impact of wireless network traffic location and access technology on mobile device energy consumption”, Local Computer Networks (LCN), 2012 IEEE 37th Conference on , 2012 , pp. 200–203
[78] H. M. K. G. Bandara, H. A. Caldera, "Towards optimising Wi-Fi energy consumption in mobile phones: A data driven approach," 2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, 2015, pp. 226-235, doi: 10.1109/ICTER.2015.7377693.
[79] P. S. Deogun, S. Ranjan, P. Rathod, A. Karandikar, N. Akhtar, "Energy efficient IEEE 802.11 WLAN discovery for heterogeneous 3GPP LTE network," 2015 IEEE Wireless Communications and Networking Conference (WCNC), New Orleans, LA, 2015, pp. 1392-1397, doi: 10.1109/WCNC.2015.7127672.
[80] N. Brouwers, M. Zuniga, K. Langendoen, "Incremental Wi-Fi scanning for energy-efficient localization," 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom), Budapest, 2014, pp. 156-162, doi: 10.1109/PerCom.2014.6813956.
[81] T. H. Lim, S. H. Rhee, "An Adaptive Power Management Scheme for WLANs using Reinforcement Learning," 2019 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea (South), 2019, pp. 412-415, doi: 10.1109/ICTC46691.2019.8939921.
[82] E. Rattagan, "Wi-Fi usage monitoring and power management policy for smartphone background applications," 2016 Management and Innovation Technology International Conference (MITicon), Bang-San, 2016, pp. MIT-171-MIT-175, doi: 10.1109/MITICON.2016.8025223.
[83] S. Dimatteo, P. Hui, B. Han, V.O.K. Li, “Cellular Traffic Offloading through WiFi Networks”, Eight IEEE International Conference on Mobile Ad-Hoc and Sensor Systems, 2011
[84] N. Ristanovic, J.Y. Le Boudec, A. Chaintreau, V. Erramilli, “Energy Efficient Offloading of 3G Networks”, in Proc. IEEE 8th International Conference on Mobile Ad-Hoc and Sensor Systems (MASS), Valencia, Spain, October 2011, pp. 202-211
[85] E. Bulut, B.K. Szymanski, “WiFi Access Point Deployment for Efficient Mobile Data Offloading”, in Proceedings of the ACM International Workshop on Practical Issues and Applications in Next Generation Wireless Networks (PINGEN 2012), Istanbul, Turkey, August 26, 2012
[86] K. Lee, J. Lee, Y. Yi, I. Rhee, S. Chong, “Mobile Data Offloading: How Much Can WiFi Deliver?”, IEEE/ACM Transactions on Networking, Volume 21, No. 2, April 2013
[87] A. Balasubramanian, R. Mahajan, A. Venkataramani, “Augmenting Mobile 3G using WiFi”, 8th Annual International Conference on Mobile Systems, Applications, and Services, San Francisco, CA, USA, June 15-18, 2010

[88] I. Trestian, S. Ranjan, A. Kuzmanovic, “Taming the Mobile Data Deluge with Drop Zones”, IEEE/ACM Transactions on Networking, Volume 20, No. 4, pp.1010-1023, August 2012
[88] M. Stiemerling, S. Kiesel, “Cooperative P2P Video Streaming for Mobile Peers”, in Proc. IEEE 19th International Conference on Computer Communications and Networks (ICCCN 2010), Zurich, Switzerland, August 2–5, 2010
[89] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, Y. Koucheryavy, “Cellular traffic offloading onto network-assisted device-to-device connections”, IEEE Communications Magazine, Volume 52, No. 4, pp. 20-31, April 2014
[90] S.I. No. 182/1997 - Road Traffic (Traffic and Parking) Regulations, 1997, available http://www.irishstatutebook.ie/eli/1997/si/182/made/en/print, accessed December 2016
[91] Dr. Vukan, R. Vuchic, Urban Public Transportation: Systems and Technology, ISBN 0139394966, Prentice-Hall College Division, NJ 1981
[92] UK Commission for Integrated Transport, Fact Sheet No. 6, 2005
[93] Dublin Bus, Dublin Bus Network Review (Report), February 2007 [Online] Available https://www.dublinbus.ie/PageFiles/2430/2007EnglishReport.pdf [Accessed June 2015]
[94] R. Coyne, “Opening Statement to Joint Oireachtas Committee on Transport, Tourism and Sport, 5th October 2016”, [Online], Available http://data.oireachtas.ie/ie/oireachtas/committee/dail/32/joint_committee_on_transport_tourism_and_sport/submissions/2016/2016-10-05_opening-statement-mr-ray-coyne-bus-atha-cliath-dublin-bus_en.pdf [Accessed Dec 2016]
[95] John J. Fruin, “Pedestrian Planning and Design, 3rd Edition”, Chapter 3-Traffic and space characteristics of pedestrians, ISBN 99909 4745 9789990914740, Elevator World Inc, Mobile, Al, USA, 1987
[96] Seth B. Young, “Evaluation of Pedestrian Walking Speeds in Airport Terminals”, Transportation Research Record 1674, Transportation Research Board of the National Academies, Volume 1674/1999, pp 20-26, 1999
[97] R.L. Knoblauch, M.T Pietrucha, M. Nitzburg, “Field Studies of Pedestrian Walking Speed and Start-up Time”, Transportation Research Board, Transportation Research Record No. 1538, Pedestrian and Bicycle Research, 1996
[98] U.S. Department of Transportation, Federal Highway Administration, Manual on Uniform Traffic Control Devices for streets and highways, Section 4E.06 Pedestrian Intervals and Signal Phases, 2009 edition, [Online], Available https://mutcd.fhwa.dot.gov/pdfs/2009/mutcd2009edition.pdf [Accessed May 2016]
[99] NYC Department of City Planning, Transportation Division, New York City Pedestrian Level of Service Study Phase 1, [Online] Available https://www1.nyc.gov/assets/planning/download/pdf/plans/transportation/td_fullpedlosb.pdf [Accessed June 2019]
[100] S. Chandra, A.K. Bharti, “Speed Distribution Curves for Pedrestrians during Walking and Crossing”, 2nd Conference of Transportation Research Group of India (2nd CTRG), 12th to 15th December 2013 Agra in North India, Utter Pradesh, India
[101] R.G. Golledge, “Path Selection and Route Preference in Human Navigation: A Progress Report”, Proceedings of the European Conference on Spatial Information Theory (COSIT), pp 207-222, Austria, 1995
[102] R. Conroy Dalton, “The Secret is to Follow Your Nose – Route Path Selection and Angularity”, Proceedings of the 3rd International Space Syntax Symposium, Atlanta, 2001
[103] B. Hillier, Moving Diagonally, Some Results and Conjectures, University College London, London, UK, 1997
[104] D. Helbing, P. Molnar, I.J. Farkas, “Self-organising pedestrian movement”, Environment and Planning B: Planning and Design 2001, Volume 28, pp 361-383, 2001

[105] D. Tang, M. Baker, “Analysis of a Metropolitan-Area Wireless Network”, Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking, pp 13-23, Seattle, Washington, United States, 1999
[106] T. Henderson, D. Kotz, I. Abyzov, “The Changing Usage of a Mature Campus-wide Wireless Network”, Proceedings of the 10th Annual International Conference on Mobile Computing and Networking (MobiCom 2004), SESSION: Experimental Testbeds and Data, pp 187-201, Philadelphia, PA, USA, 2004
[107] D. Kotz, K. Essien, “Analysis of a Campus-wide Wireless Network”, Proceedings of the Eight Annual International Conference on Mobile Computing and Networking (MOBICOM’02), pp 107 – 118, September 2002
[108] M. McNett, G.M. Voelker, “Access and Mobility of Wireless Mobile Device Users”, Mobile Computing and Communications Review, Volume 9, Number 2, pp 40-55, April 2005
[109] D. Schwab, R. Bunt, “Characterising the Use of a Campus Wireless Network”, IEEE INFOCOM 2004, Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, Volume 2, pp 862-870, March 2004
[110] M. Balazinska, P. Castro, “Characterising Mobility and Network Usage in a Corporate Wireless Local Area Network”, Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, pp 303-316, San Francisco, California, 2003
[111] C.C. Tossel, P. Kortum, A. Rahmati, C. Shepard, L. Zhong, “Characterizing Web Use on Smartphones”, SIGCHI Conference on Human Factors in Computing Systems, May 5-10, 2012, Austin, Texas, USA
[112] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, G. Cheng, “Characterizing User Behavior in Mobile Internet”, IEEE Transactions on Emerging Topics in Computing, Volume 3, No. 1, March 2015
[113] F. Kuipers, R. Kooij, D. De Vleeschawer, K. Brunnstrom, “Techniques for Measuring Quality of Experience”, in Proc. 8th International Conference, WWIC 2010, Lulea, Sweden, June 1-3, 2010
[114] ITU-T, “P.800 : Methods for subjective determination of transmission quality”, [Online], Available https://www.itu.int/rec/T-REC-P.800-199608-I/en [Accessed December 2016]
[115] Q. Huynh-Thu, M. Ghanbari, “Scope of Validity of PSNR in image/video quality assessment”, Electronics Letters, Volume 44, No. 13, June 19, 2008
[116] M. Pinson, S. Wolf, “A New Standardised Method for Objectively Measuring Video Quality”, IEEE Transactions on Broadcasting, Volume 50, No. 3, September 2004
[117] S. Khirman, P. Henriksen, “Relationship between Quality of Service and Quality of Experience for Public Internet Service”, Proceedings of Passive and Active Measurement (PAM2002), Fort Collins, Colorado, USA, 2002
[118] R.K.P. Mok, E.W.W Chan, R.K.C Chang, “Measuring the Quality of Experience of HTTP Video Streaming”, IFIP/IEEE International Symposium on Integrated Network Management, Dublin, Ireland, May 23 – 27, 2011
[119] H.J. Kim, G.G. Yun, H-S. Kim, K.S. Cho, S.G. Choi, “QoE Assessment Model for Video Streaming Service using QoS Parameters in Wired-Wireless Network”, 14th International Conference on Advanced Communication Technology (ICACT), Phoenix Park, PyeongChang, South Korea, 2012
[120] S.Y. Yoon, S. Lee, Y. Kim, P. Lee, C.Y. Oh, I. Youn, E. Monroy, Z. Hasany, J. Choi, “Mobile data service QoE analytics and optimisation”, IEEE International Conference on Communication Workshop (ICCW), London, UK, June 8-12, 2015

[121] A. Sideris, E. Markakis, N. Zotos, E. Pallis, C. Skianis, “MPEG-DASH users QoE: The segment duration effect”, IEEE 7th International Workshop on Quality of Multimedia Experience (QoMEX), Costa Navarino, Messinia, Greece, May 26-29, 2015
[122] N. Khambari, B. Ghita, L. Sun, "QoE-driven video enhancements in wireless networks through predictive packet drops," 2017 IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Rome, 2017, pp. 355-361, doi: 10.1109/WiMOB.2017.8115811.
[123] M. Seufert, V. Burger, F. Kaup, "Evaluating the Impact of WiFi Offloading on Mobile Users of HTTP Adaptive Video Streaming," 2016 IEEE Globecom Workshops (GC Wkshps), Washington, DC, 2016, pp. 1-6, doi: 10.1109/GLOCOMW.2016.7848897.
[124] B. Koch, A. Lins, R. Rizk, R. Steinmetz, D. Hausheer, "vFetch: Video prefetching using pseudo subscriptions and user channel affinity in YouTube," 2017 13th International Conference on Network and Service Management (CNSM), Tokyo, 2017, pp. 1-6, doi: 10.23919/CNSM.2017.8256011.
[125] S. K. Park, A. Bhattacharya, M. Dasari, S. R. Das, "Understanding User Perceived Video Quality Using Multipath TCP Over Wireless Network," 2018 IEEE 39th Sarnoff Symposium, Newark, NJ, USA, 2018, pp. 1-6, doi: 10.1109/SARNOF.2018.8720402.
[126] ISO, ISO/IEC 23009-1:2014, Information technology — Dynamic adaptive streaming over HTTP (DASH) — Part 1: Media presentation description and segment formats, [Online], Available https://www.iso.org/standard/65274.html [Accessed November 2018]
[127] K. Spiteri, R. Urgaonkar, R. K., Sitaraman, “BOLA: Near-Optimal Bitrate Adaptation for Online Videos”, arXiv:1601.06748v2, April 2016
[128] T. Huang, R. Johari, N. McKeown, M. Trunnell, M. Watson, “A Buffer-Based Approach to Rate Adaptation: Evidence from a Large Video Streaming Service”, Proc. of the ACM conference on SIGCOMM, Pages 187-189, Chicago, Illinois, August 2014
[129] W. Huang, Y. Zhou, X. Xie, D. Wu, M. Chen, E. Ngai, "Buffer State is Enough: Simplifying the Design of QoE-Aware HTTP Adaptive Video Streaming," IEEE Trans. on Broadcasting, vol. 64, no. 2, pp. 590-601, June 2018.
[130] J. Jiang, V. Sekar, H. Zhang, “Improving fairness, efficiency and stability in http-based adaptive video streaming with FESTIVE”, Proc. of the 8th International Conference on Emerging Network Experiments and Technologies, ACM, 97-108, 2012
[131] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, D. Oran, “Probe and Adapt: Rate Adaptation for HTTP Video Streaming at Scale”, IEEE Journal on Selected Areas in Communications 32, 4(2014), 719-733
[132] L. De Cicco, V. Caldaralo, V. Palmisano, S. Masolo, “Elastic: a client-side controller for dynamic adaptive streaming over http (dash)”, Pocket Video Workshop (PV), 20th International, IEEE, 1-8, 2013
[133] A. Beben, P. Wisniewski, J. M. Batalla, P. Krawiec, “ABMA+: lightweight and efficient algorithm for HTTP streaming”, Proceedings of the 7th International Conference on Multimedia Systems, ACM, 2, 2016
[134] K. Spiteri, R. Sitaraman, D. Sparacio, “From Theory to Practice: Improving Bitrate Adaption in the DASH Reference Player”, MMSys’18, Proc. of the 9th ACM Multimedia Systems Conference, Pages 123-127, Amsterdam, Netherlands, June 2018
[135] DASH Reference Client 2.9.3, [Online] Available https://reference.dashif.org/dash.js/v2.9.3/samples/dash-if-reference-player/index.html, [Accessed April 2019]
[136] A. Bentaleb, A. C. Gegen, R. Zimmermann, “Game Theory Based Bitrate Adaptation for DASH.js Reference Player”, 2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) ,San Diego, CA. USA, 23-27 July 2018

[137] C. Timmerer, M. Maiero, B. Rainer, S. Petscharnig, “Quality of Experience of Adaptive HTTP Streaming in Real-World Environments”, Vol.10, No.3, May 2015, IEEE COMSOC MMTC E-Letter
[138] Cisco, “Cisco Annual Internet Report (2018–2023)”, White Paper, available at https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html
[139] W. Huang, Y. Zhou, X. Xie, D. Wu, M. Chen, E. Ngai, "Buffer State is Enough: Simplifying the Design of QoE-Aware HTTP Adaptive Video Streaming," IEEE Trans. on Broadcasting, vol. 64, no. 2, pp. 590-601, June 2018.
[140] Xiaokang Sang, Qian Wu, Hewu Li, "iScan: Efficient WiFi Scan for mobile device based on client and network behavior learning," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 1109-1114, doi: 10.1109/ISCC.2017.8024674.
[141] J. Han, J. Kim, C. Joo, S. Bahk, "SplitScan: Sharing Wi-Fi Scan Information through Bluetooth Low Energy," 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA, 2019, pp. 1-5, doi: 10.1109/VTCFall.2019.8891362.
[142] WiGLE Wardriving application, [Online] Available https://wigle.net/ [Accessed 30 September 2016]
[143] NS-3 Network Simulator, available at: http://www.nsnam.org/ [Accessed10 August 2017]
[144] Three Ireland Ltd., Available https://www.three.ie/explore/about-three/ [Accessed June 2020]
[145] OOKLA Speedtest for Android, available at: http://www.speedtest.net/mobile/android/ [Accessed 30 September 2016]
[146] OmNet++, [Online] https://omnetpp.org/ [Available August 2020]
[147] NetSim v12.2, [Online] https://www.tetcos.com/software-download.html [Available August 2020]
[148] W. Zhang, H. He, S. Ye, Z. Wang, Q. Zheng, “Enhancing QoE for Mobile Users by Environment-Aware HTTP Adaptive Streaming”, Sensors. 2018; 18(11):3645
[149] M. Seufert, F. Wamser, P. Casas, R. Irmer, P. Tran-Gia, R. Schatz, “YouTube QoE on Mobile Devices: Subjective Analysis of Classical vs. Adaptive Video Streaming “, International Wireless Communications and Mobile Computing Conference, Dubrovnik, Croatia, Aug. 2015
[150] Linux Containers, [Online], Available https://linuxcontainers.org/ [Accessed December 2015]
[151] VideoLAN, [Online], Available https://www.videolan.org/vlc/ [Accessed June 2015]
[152] EnvivioDASH3, [Online], Available https://dash.akamaized.net/envivio/EnvivioDash3/ [Accessed December 2015]
[153] B. Koch, A. Lins, R. Rizk, R. Steinmetz, D. Hausheer, "vFetch: Video prefetching using pseudo subscriptions and user channel affinity in YouTube," 2017 13th International Conference on Network and Service Management (CNSM), Tokyo, 2017, pp. 1-6, doi: 10.23919/CNSM.2017.8256011.
[154] J. Nightingale, P. Salva-Garcia, J. M. A. Calero, Q. Wang, "5G-QoE: QoE Modelling for Ultra-HD Video Streaming in 5G Networks," in IEEE Transactions on Broadcasting, vol. 64, no. 2, pp. 621-634, June 2018, doi: 10.1109/TBC.2018.2816786
[155] A. Ali, A. Al Ajami, J. Alotaibi, "Subjective and Objective Evaluation of the Effect of Packet Loss and Delay on Video Streaming Quality", International Journal of Computer and Information Technology, no. 2, March 2016, ISSN 2279-0764

[156] A. Moldovan, I. Ghergulescu, C. H. Muntean, "VQAMap: A Novel Mechanism for Mapping Objective Video Quality Metrics to Subjective MOS Scale," IEEE Trans on Broadcasting, vol. 62, no. 3, pp. 610-627, Sept. 2016
[157] A. Moldovan, I. Ghergulescu, S. Weibelzahl, C. H. Muntean, "User-centered EEG-based multimedia quality assessment," IEEE Int. Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), London, 2013, pp. 1-8.
[158] A. Asan, W. Robitza, I.-H. Mkwawa, L. Sun, E. Ifeachor, A. Raake, "Impact of Video Resolution Changes on QoE For Adaptive Video Streaming", IEEE International Conference on Multimedia and Expo ICME, Hong Kong, 2017
[159] T. Huang, R. Johari, N. McKeown, M. Trunnell, M. Watson, "A Buffer-Based Approach to Rate Adaptation: Evidence from a Large Video Streaming Service", Proc. of the ACM conference on SIGCOMM, Pages 187-189, Chicago, Illinois, August 2014
[160] Youtube, [Online] Available https://www.youtube.com [Accessed August 2020]
[161] Netflix, [Online] Available https://www.netflix.com [Accessed August 2020]
[162] T. Casey, G.-M. Muntean, "Scan-Or-Not-To-Scan – Balancing Network Selection Accuracy and Energy Consumption", Proc. of the 11th Int. Wireless Communication and Mobile Computing Conference (IWCMC), Dubrovnik, Croatia, August 24-28 2015
[163] DASH IF Spring 4Ktest Stream, [Online], Available http://dash.edgesuite.net/akamai/streamroot/050714/Spring_4Ktest.mpd [Accessed February 2017]
[164] DASH IF Reference Client 2.4.1, [Online], Available http://dashif.org/reference/players/javascript/v2.4.1/samples/dash-if- reference-player/index.html, [Accessed February 2017]
[165] WonderShaper, [Online], Available https://github.com/magnific0/wondershaper [Accessed February 2017]
[166] Interface Bonding, [Online], Available https://wiki.linuxfoundation.org/networking/bonding [Accessed February 2017]
[167] Debian, [Online], Available https://www.debian.org/ [Accessed November 2018]
[168] Oracle VM Virtualbox, [Online] Available https://www.oracle.com/ virtualization/virtualbox/ [Accessed Nov. 2018]
[169] DASH Reference Client 2.9.3, [Online] Available https://reference.dashif.org/dash.js/v2.9.3/samples/dash-if-reference-player/index.html, [Accessed April 2019]
[170] Media Source Extensions (MSE), [Online] Available https://www.w3.org/ TR/media-source/, [Accessed Feb. 2019]
[171] K. Spiteri, R. Urgaonkar, R. K., Sitaraman, "BOLA: Near-Optimal Bitrate Adaptation for Online Videos", arXiv:1601.06748v2, April 2016
[172] K. Spiteri, R. Sitaraman, D. Sparacio, "From Theory to Practice: Improving Bitrate Adaption in the DASH Reference Player", MMSys'18, Proc. of the 9th ACM Multimedia Systems Conference, Pages 123-127, Amsterdam, Netherlands, June 2018
[173] HD Video Encoding Settings, [Online] Available https://support.video.ibm.com/hc/en-us/articles/207852117-Internet-connection-and-recommended-encoding-settings [Accessed Jan. 2019]