Differentiated QoS Connectivity for Society-Critical Use Cases in the 5G Era

Obiodu, Emeka

Awarding institution:
King's College London

Differentiated QoS Connectivity for Society-Critical Use Cases in the 5G Era

Chukwuemeka Obiodu

Supervisors: Prof. Hamid Aghvami & Prof. Nishanth Sastry

Department of Engineering, Faculty of Natural and Mathematical Sciences
King's College London

This dissertation is submitted for the degree of
Doctor of Philosophy

April 2023

Abstract

There is an understandable expectation across society that 5G will support Society-Critical System (SCS) use cases such as connected vehicles and industrial robots. This is based on design and network capacity improvements in 5G that deliver improved reliability. But will these improvements guarantee consistent quality of service (QoS) to SCS use cases? The evidence from mature 4G and early 5G networks suggest otherwise. Using latency as a proxy for performance, our field experiments on 4G networks in London on four New Year Days (plus additional January dates) from 2016/17 - 2019/20, show that a user has only a 58% chance to get a latency below 50ms. Likewise, our measurements on early 5G networks in London (UK) and Seoul (South Korea) in 2021 show that a user has only a 50% chance to get a latency below 50ms. If mature 4G and early 5G networks are unable to provide consistent performance for SCS, then it follows that mature 5G networks will likely perform similarly.

Working with colleagues during my PhD studies, I explored if and how a Differentiated QoS (D-QoS) approach can improve QoS guarantees. I focused on using prioritisation, redundant multi-connectivity or exclusive infrastructure (i.e. Private Networks) to provide QoS assurance for selected services. Given the long and mostly unsuccessful history of D-QoS (e.g. IntServ, DiffServ, 4G QCI, etc), I critically examined the merits of a more targeted, and officially-supported D-QoS approach. I then did a deep dive to evaluate the reality of a redundant multi-connectivity D-QoS example. I did this experimentally, showing via a system implementation and drive-through measurements on over 800 kilometers of roads in South East England, that the use of redundant connections in Connected Cars, preferably user-managed on the demand side, can improve performance by up to 28 percentage points.

In this thesis, I bring together these insights to investigate the proposal that an officially-supported D-QoS approach is a suitable means of providing QoS assurances for SCS use cases in the 5G era. I interrogate this thesis extensively using the insights and outputs from the work done during my PhD to confirm that the thesis is supported by the evidence from empirical evaluation. I then practicalise the thesis for implementation by technical, commercial and policy stakeholders by proposing the D-QoS Guidance Framework (DGF), conceptualising a 999 app and using the DGF to evaluate how an officially supported 999 app can be made to work globally.

To the loving memory of the two women who gave everything so that I can be anything.

…

Mum: *Ezinne Ijeoma Christiana Obiodu (nee Emereuwa). Ezinne m, this is for all the hollandais you couldn't buy while paying my school fees.*

Wife: *Dr Ulumma Vivian Obiodu (nee Chukwu). Nkem, I made you a promise on your death bed in Year 1 of a PhD that you encouraged me to get on with. Yes, I didn't quit. I finished it.*

…

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this report are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This report is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 100,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Chukwuemeka Obiodu
April 2023

Acknowledgements

The seed of this PhD journey started in November 1996 in my first year as an undergraduate in FUTO. Since then, so many people have inspired, encouraged, cajoled, guided, mentored or supported me to finally get the PhD done. To each and every one of you, I express my heartfelt gratitude and appreciation.

I want to especially thank my PhD supervisors, Prof Nishanth Sastry and Prof Hamid Aghvami, for their support in my day to day PhD journey. Nishanth welcomed me to King's and adroitly guided me both academically and pastorally to see that I finished. Thank you for being a supervisor, a mentor and a friend. Hamid brought me to the finishing line, providing me with nuggets of wisdom and sagacious counsel. Having read about you in IEE journals in the late 1990s, it fills me with pride that I got the opportunity to be your student.

To my PhD colleagues (or more appropriately, supporting supervisors), I cannot thank you enough. Aravindh Raman, Abdullahi Kutiriko Abubakar, Pushkal Agarwal, Tooba Faisal and Sagar Joglekar, thank you for all the experiments, discussions, papers, laughs, lunches/dinners, parties etc. Dancing by 02:00am while observing mobile phone measurements in KCL lab has got to be one of the most memorably weird PhD escapades in history (no pun intended). Thank you also to my other collaborators - especially Simone Mangiante (Vodafone), Alex Geoman (Fujitsu) and Dongwook Kim (ETSI) - for all your support.

I am deeply grateful to my two employers (NVIDIA and GSMA) and work colleagues during the seven years of my part-time PhD journey. I thank you for all the support, insights, accommodations for my shortfall in other areas, and opportunities to use my PhD ideas in real-live scenarios that affect billions of people and multi-billion dollar businesses. Thank you also to my friends, school mates and colleagues, especially the members of the two socio-cultural groups (Ndiigbo Association Hertfordshire and XJAYS UK/Europe) who did a fantastic job in ensuring that I had a social life during the PhD.

My family is my everything and I remain very proud of my parents and my siblings for all their motivation and encouragement on this PhD journey. I thank my dad, Sir K, for all the intellectual stimulations, including during our weekly hour calls. To my siblings, Chinenye, Chidubem, Ngozi, Ikechukwu and Chidiebere, *ihe dịka unu akọkwala m.* That's 4/6 doctors in the house now. Powerful, isn't it?

Finally, I reserve my deepest gratitude and appreciation to my wife, Ijeoma, for everything she has done for me over the past four years. *Añulim* (my joy), thank you for lifting me up from my despairing widower state and restoring my confidence, belief and joy. To our little boys - Dinobichi and Netochi - the good news is that you now have two doctors in the house; and the really really good news is that daddy will no longer have an excuse not to want to play with you!

Chukwuemeka Obiodu
April 2023

Table of contents

List of figures **xiii**

List of tables **xv**

List of Abbreviations **xvi**

1 Introduction **1**

1.1 Introduction 1
1.2 Scope of Using D-QoS for SCS 5
1.3 Thesis Statement 5
1.4 Research Goals & Questions 7
1.5 Methodology 8
1.6 Contributions 9
1.7 List of Publications 10
1.8 Thesis Overview 11
1.9 Challenges 11

2 Background & Context **13**

2.1 Introduction 13
2.1.1 Key questions answered 13
2.1.2 Chapter layout 14
2.1.3 Publications linked to chapter 14
2.2 Society Critical Systems (SCS) 15
2.2.1 Introduction 15
2.2.2 Safety Critical Systems (SaCS) 15
2.2.3 Mission Critical Systems (MCS)/Business Critical Systems (BCS) . 18
2.2.4 Critical Information Systems (CIS) 19
2.2.5 Security Critical Systems (SeCS) 21
2.3 Understanding Mobile Network Performance & Reliability 23

2.3.1 Reliability & resilience in mobile connectivity systems 23
2.3.2 Speed, Throughput and Bandwidth 23
2.3.3 Latency, RTT, RTT Variance (Jitter) 24
2.4 5G-era Mobile Network Performance 24
2.4.1 Introduction 24
2.4.2 5G expected performance 25
2.4.3 5G performance drivers 27
2.4.4 5G actual performance 28
2.4.5 5G perception and 'Fake' 5G icons 29
2.4.6 4-stage framework for determining 5G icon status 30
2.4.7 Determining factors for network performance 33
2.4.8 Improving network performance for SCS 35
2.4.9 Measuring network performance 35
2.5 Differentiated Connectivity 37
2.5.1 Rationale and options 37
2.5.2 Net Neutrality 38
2.6 Redundancy Analysis for Connectivity 42
2.6.1 Redundancy for reliability & resilience 42
2.6.2 Redundancy for performance-constrained reliability 42
2.6.3 Redundancy setup for mobile services 42
2.6.4 Global feasibility of redundant mobile connectivity 43
2.7 Differentiated Connectivity for Connected Cars 45
2.7.1 Setting the context 45
2.7.2 Performance expectations for Connected Cars 45
2.7.3 State of market for Connected Cars 46
2.7.4 Overcoming challenges to mobile connectivity 47
2.7.5 Comparing V2X and smartphones 48
2.7.6 Improving reliability of V2X 48
2.8 Chapter Summary 49

3 Methodology 50
3.1 Introduction 50
3.1.1 Key questions answered 50
3.1.2 Chapter layout 51
3.1.3 Publications linked to chapter 51
3.2 Research Plan 52
3.2.1 Research Steps 52

3.2.2 Field Experiments 52
3.3 Experiment Setup and Details 53
3.3.1 Measurement Setup 53
3.3.2 Measurement Software 53
3.3.3 Measurement Hardware 54
3.3.4 Measurement parameters 56
3.3.5 Mobile Network Operators 56
3.3.6 Benchmarks for analysis 58
3.4 Chapter Summary 59

4 Understanding Mobile Network Performance 60
4.1 Introduction 60
4.1.1 Key questions answered 60
4.1.2 Chapter layout 61
4.1.3 Publications linked to chapter 61
4.2 Experiment 1: Performance in Pre-5G UK Networks 62
4.2.1 Aim 62
4.2.2 Methodology 62
4.2.3 Results 65
4.2.4 Discussion 76
4.2.5 Conclusion and next step 79
4.3 Experiment 2: Performance in an early UK 5G Network 80
4.3.1 Aim 80
4.3.2 Methodology 81
4.3.3 Results 82
4.3.4 Discussion 87
4.3.5 Conclusion and next step 88
4.4 Experiment 3: Performance in UK & South Korea 5G networks 90
4.4.1 Aim 90
4.4.2 Methodology 91
4.4.3 Results 93
4.4.4 Discussion 101
4.4.5 Conclusion and next step 102
4.5 Chapter Summary 103

5 Introducing Differentiated QoS Connectivity 104
5.1 Introduction . . . 104
5.1.1 Key questions answered . . . 105
5.1.2 Chapter layout . . . 105
5.1.3 Publications linked to chapter . . . 105
5.2 D-QoS for Reliability Assurance for SCS . . . 106
5.2.1 Introducing D-QoS mechanisms . . . 106
5.2.2 History of D-QoS mechanisms . . . 112
5.2.3 Rationale for D-QoS mechanisms . . . 112
5.3 Classifying D-QoS Mechanisms . . . 113
5.3.1 Purpose of D-QoS mechanism . . . 113
5.3.2 Infrastructure used for D-QoS . . . 114
5.3.3 Packet handling by D-QoS mechanism . . . 114
5.3.4 Applicable network layer for D-QoS mechanism . . . 115
5.3.5 Target market for D-QoS mechanism . . . 115
5.4 Predicting the success of a D-QoS Mechanism for SCS . . . 116
5.4.1 Lessons from the history of D-QoS mechanisms . . . 116
5.4.2 15 key questions on the success of a D-QoS mechanism for SCS . . . 117
5.5 Practical Considerations for the use of D-QoS for SCS in 5G era networks . . . 122
5.5.1 Taxonomy for identifying SCS use cases for D-QoS . . . 122
5.5.2 CLASP: framework for deciding 'official' SCS use cases . . . 127
5.6 Chapter Summary . . . 129

6 Evaluating D-QoS for Connected Cars in the Field 130
6.1 Introduction . . . 130
6.1.1 Key questions answered . . . 131
6.1.2 Chapter layout . . . 131
6.1.3 Publications linked to chapter . . . 131
6.2 Setting the Context . . . 132
6.2.1 Aim . . . 132
6.2.2 Motivation . . . 132
6.2.3 Background . . . 133
6.3 System Model . . . 134
6.3.1 Threshold for reliability parameters . . . 134
6.3.2 Number of multi-operator connectivity options needed . . . 135
6.4 Experiment Design, Parameters & Setup . . . 136
6.4.1 Experiment goals & design . . . 136

6.4.2 Experiment route and dates . . . 138
6.4.3 Hardware setup . . . 141
6.4.4 Software setup . . . 142
6.4.5 Benchmarks for analysis . . . 143
6.4.6 Comparison with drive-through tests . . . 143
6.5 Results . . . 143
6.5.1 Day 1 - Establishing the hypothetical benefit of DSM-MoC . . . 144
6.5.2 Day 2: Implementing DSM-MoC & validating its benefit . . . 147
6.5.3 Day 3/4: Determining impact of DSM-MoC on TCP/UDP . . . 149
6.6 Discussions, Implications & Recommendations . . . 154
6.6.1 DSM-MoC delivers superior performance . . . 154
6.6.2 SSM-MoC's network architecture creates challenges . . . 154
6.6.3 Operational complexities for DSM-MoC are surmountable . . . 156
6.6.4 Multi-radio connectivity as default for CCs . . . 156
6.6.5 CCs have computational power to manage DSM-MoC . . . 157
6.6.6 Cost constraints discourage DSM-MoC . . . 157
6.6.7 Policy guidance is imperative . . . 158
6.7 Key contribution . . . 159
6.8 Chapter Summary . . . 160

7 D-QoS Guidance Framework 161
7.1 Introduction . . . 161
7.1.1 Key questions answered . . . 162
7.1.2 Chapter layout . . . 162
7.1.3 Publications linked to Chapter . . . 162
7.2 Exploring the D-QoS Guidance Framework . . . 164
7.2.1 Scope . . . 164
7.2.2 Question 1: Is it an SCS use case? . . . 164
7.2.3 Question 2: Is mobile connectivity preferred? . . . 164
7.2.4 Question 3: Is usage in multiple locations? . . . 164
7.2.5 Question 4: Is usage only for specific applications? . . . 165
7.2.6 Question 5: Does usage meet criteria for official support? . . . 165
7.3 Implementability of DGF . . . 166
7.3.1 Policy and Regulatory Considerations . . . 166
7.3.2 Technical Considerations . . . 166
7.3.3 Commercial Considerations . . . 167
7.3.4 Time Considerations . . . 167

7.4 DGF Evaluation for a 999 CIS app 168
7.4.1 Introducing the 999 app 168
7.4.2 Question 1: 999 app is a CIS and an SCS 169
7.4.3 Question 2: 999 app needs mobile connectivity 170
7.4.4 Question 3: 999 app needs to work everywhere 170
7.4.5 Question 4: 999 app plays a specific role 171
7.4.6 Question 5: 999 app should be officially enforced 171
7.5 Chapter Summary 171

8 Summary, Limitations & Conclusion 173
8.1 PhD research in a nutshell 173
8.2 PhD report summary 174
8.2.1 Recap: Thesis, Research Goals & Research Questions 174
8.2.2 Summary & Takeaways 175
8.3 Limitations and future directions 177
8.3.1 Insufficient unique experimental results 177
8.3.2 Incomplete evaluation of introduced concepts 177
8.3.3 Gaps in understanding D-QoS 178
8.3.4 SCS connectivity requirements 178
8.3.5 Reusing data in analytical models 178
8.3.6 Increased cross disciplinary approach 178
8.4 Final Remarks 178

References 181

List of figures

1.1 What is a Society Critical System (SCS)? 2
1.2 Human congestion at Holborn Station London (01:00am; 01/01/2017) . . . 3

2.1 IEC 61508 framework 17
2.2 Government-owned emergency alert apps 20
2.3 5G NSA vs 5G SA 26
2.4 Spectral Outcomes for 5G capacity 27
2.5 Android Architecture Stack 31
2.6 GSMA/3GPP rules for 5G icon 32
2.7 Variation in network performance 34

3.1 Measurement Setup 54

4.1 Exp 1: Overall RTT readings for 3G and 4G devices 66
4.2 Exp 1: RTT Box for NP1 - NP4 67
4.3 Exp 1: RTT Box for 2017 - 2020 68
4.4 Exp 1: RTT CDF for NP1 - NP4 69
4.5 Exp 1: RTT CDF for 2017 - 2020 69
4.6 Exp 1: Median RTT (24 hours vs 6 hours) 70
4.7 Exp 1: Median RTT for 14 UK websites 71
4.8 Exp 1: Page Load Times 72
4.9 Exp 1: Number of base stations/sectors in use 73
4.10 Exp 1: Network Type on 3G & 4G devices 74
4.11 Exp 1: Downlink and Uplink Speeds 75
4.12 Exp 1: RTT in City Centre vs Train Station 75
4.13 Exp 1: Base Stations MNO vs MVNO 76
4.14 Exp 1: MNO vs MVNO comparison 77
4.15 Exp 2: RTT comparison 4G vs 5G 84
4.16 Exp 2: Downlink speed comparison 4G vs 5G 84

4.17 Exp 2: Uplink speed comparison 4G vs 5G 85
4.18 Exp 2: Base station diversity 86
4.19 Exp 2: Video playback comparison 4G vs 5G 86
4.20 Exp 3: Determining 5G Status with Netmonster Core 94
4.21 Exp 3: Overall RTT performance 95
4.22 Exp 3: RTT analysis 96
4.23 Exp 3: Downlink analysis 97
4.24 Exp 3: Uplink analysis 98
4.25 Exp 3: PLT analysis 99
4.26 Exp 3: Packet Loss analysis 100
4.27 Exp 3: Video playback time 101
4.28 Exp 3: Summary of 4G and 5G performance in UK and South Korea 102

5.1 Three types of D-QoS mechanisms based on the infrastructure used. 114
5.2 Predicting success for a D-QoS mechanism 117

6.1 Multi Operator Connectivity options for Connected Cars 133
6.2 Theoretical evaluation of Reliability and MTTF 137
6.3 Schematic setup for Connected Car experiments 138
6.4 Equipment for Connected Car experiments 139
6.5 Drive route for Connected Car experiments 140
6.6 Day 1 results: eight measurement parameters 145
6.7 Day 1 results: Availability 147
6.8 Day 1 results: predicted reliability vs Ofcom data 147
6.9 Day 2 results: page load for AWS-hosted page 148
6.10 Day 2 results: Latency comparisons 150
6.11 Day 3/4 results: Latency and RTT Variance 152
6.12 Day 4 results: DSM-MoC impact on TCP/UDP 153
6.13 Day 1 results: Comparison of SSM and DSM 155
6.14 Day 1 results: Best NP vs worst 2-network and 3-network DSM 155
6.15 Traceroute analysis for SSM-MoC 156
6.16 Location for managing redundancy 158

7.1 D-QoS Guidance Framework (DGF) 163
7.2 Advantages of 5G 165
7.3 Architectural comparison of a 999 app and the 999 call system 169

List of tables

1.1 SCS Services/Systems and D-QoS 6

2.1 IEC 61508: Safety Integrity Levels (SIL) benchmarks 17
2.2 Benchmark performance for 4G & 5G 26
2.3 Net Neutrality - Insights from other Utility-like industries 41
2.4 3GPP V2X categories and main KPIs 46

3.1 Websites used in measurement 55
3.2 Measurement parameters 57
3.3 Performance thresholds for measurement 58

4.1 Exp 1: Dates and rationale for mobile network measurements in London . . 64
4.2 Exp 3: Dates, locations and rationale for measurements in South Korea & UK 92

5.1 History of D-QoS Standards (1970s to date) 107
5.2 Differentiated service classes for mobile networks 124
5.3 CLASP Framework 128

6.1 Experiments for Multi Operator Connectivity in CCs 139
6.2 Performance thresholds for Connected Car experiments 143
6.3 Day 2 latency readings with MoC switching vs Oracle 150

7.1 CLASP Framework Analysis for 999 app 172

List of Abbreviations

2G Second Generation

3G Third Generation

3GPP Third Generation Partnership Project

4G Fourth Generation

5G Fifth Generation

5GC 5G Core

5GSI 5G Status Indicator

5QI 5G QoS Indicator

6G Sixth Generation

ABC Always Best Connected

ACB Access Class Baring

ACDC Access Control for general Data Connectivity

AI Artificial Intelligence

AJIT Accountable Just In Time

ALARP As Low As Reasonably Possible

API Application Programming Interface

ALP Allocation and Retention Policy

ARPU Average Revenue Per User

ATM Asynchronous Transfer Mode

AR/VR Augmented Reality / Virtual Reality

BCM Business Continuity Management

BCS Business Critical Systems

BS Base Station

CAP Common Alerting Protocol

CC Connected Car

CDF Cumulative Distribution Frequency

CDN Content Delivery Networks

CIP Critical Infrastructure Protection

CIS Critical Information Systems

CLASP Critical. Localised, Approved, Specific & Perishable

CN Core Network

CoS Class of Service

CP Content Providers

CPU Central Processing Unit

CT Connectivity Technology

cV2X Cellular Vehicle to Everything

D-QoS Differentiated Quality of Service

DC Dual Connectivity

DCHP Dynamic Host Configuration Protocol

DGF D-QoS Guidance Framework

DiffServ Differentiated Services

DSM Demand Side Managed

EDGE Enhanced Data Rates for GSM Evolution

EPC Enhanced packet Core

ERM Enterprise Risk Management

EU European Union

FCC Federal Communications Communication

FOTA Firmware Over The Air

GBR Guaranteed Bit Rate

GPRS General Packet Radio Service

GSA Global Mobile Suppliers Association

GSMA GSM Association

HSDPA High Speed Data Packet Access

HSE Health and Safety Executive

HSPA High Speed Packet Access

HTTP Hypertext Transfer Protocol

ICT Information & Communications technology

ICMP Internet Control Message Protocol

IEC International Electrotechnical Commission

IntServ Integrated Services

IoT Internet of Things

IP ToS Internet Protocol Type of Service

IPTV Internet Protocol TV

ISDN Integrated Services Digital Network

ISO International Standards Organisation

ISP Internet Service Provider

IT Information Technology

ITS Intelligent transport Systems

ITU-R International Telecommunications Union - Radiocommunication

KCL King’s College London

KPI Key Performance Indicators

LEO Low Earth Orbit

LTE Long Term Evolution

MAD Median Absolute Deviation

MB/GB Megabytes / Gigabytes

mbps Megabits Per Second

Mbps Megabytes Per Second

MCS Mission Critical Systems

MEC Multi-Access Edge Computing

MIMO Massive Input Massive Output

MNO Mobile Network Operator

MoC Multi-operator Connectivity

MPLS Multi Protocol Label Switching

ms Milliseconds

MSI Multiple Shared Infrastructure

MVNO Mobile Virtual Network Operator

NAT Network Address Translation

NP Network Provider (UK)

NR New Radio

NSA Non Standalone

OS Operating System

OT Operational Technology

PELR Packet Error Loss Rate

PES Programmable Electronic Systems (PESs)

PLMN Public Land Mobile Network

PLT Page Load Times

PoL Packet Loss

PoS Point of Sale

PWS Public Warning System

RIL Radio Interface Layer

RRC Radio Resource Control

RTT Round Trip Time

QCI Quality Control Indicator

QoS Quality of Service

QoE Quality of Experience

RAN Radio Access Network

RDS Radio Data Service

RQ Research Question

RSVP Resource Reservation Protocol

SA Standalone

SaCS Safety Critical Systems

SCS Society Critical Systems

SDK Software Development Kit

SDP Smart Data Pricing

SD-WAN Software Defined Wide Area Networks

SeCS Security Critical Systems

SEI Single Exclusive Infrastructure

SIB System Information Block

SIL Safety Integrity Level

SIM Subscriber Identity Module

SLA Service Level Agreement

SMS Short Message Service

SNP Network Provider (South Korea)

SSH Secure Shell

SSI Single Shared Infrastructure

SSM Supply Side Managed

TCP Transport Control Protocol

TFL Transport for London

TSN Time Sensitive Networking

UDP User Datagram Protocol

UE User Equipment

UMTS Universal Mobile Telecommunications System

URLLC Ultra Reliable and Low Latency Communication

V2I Vehicle to Infrastructure

V2N Vehicle to Network

V2P Vehicle to Pedestrian

V2X Vehicle to Everything

V2V Vehicle to Vehicle

ViLTE Video Over Long Term Evolution

VoLTE Voice Over Long Term Evolution

WME Wireless Multimedia Extension

XML eXtensible Markup Language

Chapter 1

Introduction

Ihe niile nwere mbido! (Igbo Proverb)

(Everything has a beginning!)

1.1 Introduction

The mobile industry has entered another generational change era with the commercial launch of 5G in late 2018. This promises to be a time when 5G networks will be at the centre of an ecosystem - together with parallel technological developments in computing, artificial intelligence, devices, etc - that will transform society. The GSMA recognizes the period between 2020 - 2030 as the '5G era', a time when an increasingly prevalent 5G infrastructure will co-exist with 4G to support existing and new services [1]. GSMA Intelligence projects that 5G will contribute $2.2 trillion to the global economy by 2034, transforming key industrial sectors such as manufacturing, transportation, utilities and financial services [2]. Ericsson examined the impact of 5G on ten industries and projected that 5G will enable $1.5 trillion in ICT revenues by 2030, about 39% of the global ICT revenues[3].

5G will build on the success stories of 2G/3G/4G while striving to break new grounds in supporting use cases that demand high reliability to avoid the significant negative consequences from their failure or malfunction[4]. These use cases can be safety critical[5], mission critical[6], business critical[7], critical information[8] or security critical[9]. In this work, I group all these together and refer to them as Society Critical Systems (SCS) (Figure 1.1).

I then ask two questions: first, is *'best effort'* mobile data connectivity sufficient to meet the demand for high reliability for SCS use cases? For example, between 2006 and 2016, the UK's Health and Safety Executive (HSE) public register for convictions shows that there were 2,217 health and safety cases, with total fines of £37,179,916, and wide-ranging penalties[10]. Businesses and policymakers would not want connectivity-related problems to aggravate these health and safety failings. Second, if connectivity is not to cause health and safety failings, should connectivity for SCS systems continue to be treated in the same way as for non-SCS systems? For example, can a Connected Car (CC) always rely on mobile connectivity for its safe operation and will society accept to handle the control traffic of CCs in the same way as the infotainment traffic for the passengers?

A Society Critical System (SCS) is any system whose failure or malfunction can lead to death or serious injury to people, loss or severe damage to property, significant tangible or intangible economic losses, or damage to the environment.

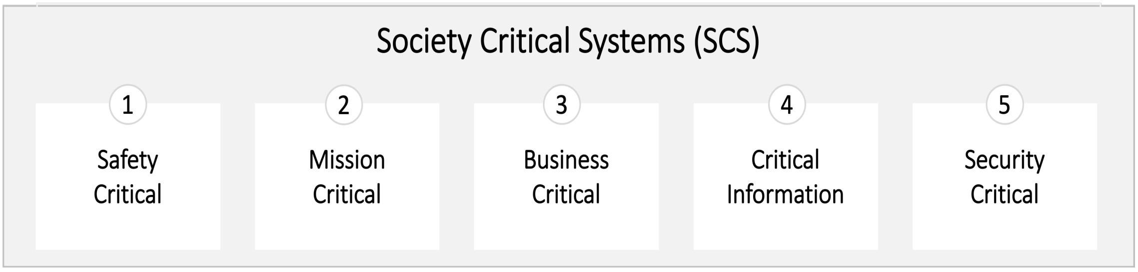


Fig. 1.1 Society Critical Systems (SCS) demand high reliability to avoid significant negative consequences from their failure or malfunction. Can 5G era mobile connectivity meet this required high reliability?

Serendipity triggered my motivation for this research. At Holborn Underground Station in London by 01:00am on 01 January 2017, there was human traffic congestion at the Station at a time when the police were struggling to marshal people away from busy transport nodes after the New Year fireworks celebration (Figure 1.2). This prompted me to ask: if travel information is already considered as critical to be prioritised on the nearly 50 year old Radio Data Service (RDS)[11], would it not be helpful, in the age of smartphones, to update the Transport for London (TFL) 'Journey Planner' app with live information and provide an assurance that the mobile network will always prioritise access to the app in the event of network congestion?

I recognize that some governments have created general emergency alert apps (e.g. '112 Suomi' in Finland, 'NINA' in Germany, 'Alertable' by several Canadian provinces and '311' by New York, USA) or alert apps for specific incidents (e.g. 'ShakeAlert' for earthquakes in the US). These apps purport to extend the Public Warning System (PWS) service beyond its original SMS format. But there is no expectation nor guarantee that these apps will

Fig. 1.2 Human congestion at Holborn Station London after the New Year Fireworks Celebration (00:50am 01/01/17). Thanks to mobile network congestion, most Internet services were down. So, despite having current transport information on the Transport for London (Tfl) Journey Planner webpage, most of the crowd could not access it. Could the mobile networks prioritise access to a Tfl app or service so that it works despite the network congestion? Would such a mechanism provide a better approach to disperse the crowd to other less congested train stations?

work during congestion. In contrast, a government-backed emergency app, with the same prioritisation and privileges as a 999 call, will offer a much better QoS guarantee. We detailed this in [8], the first time ever in the literature that a 999 app concept has been proposed.

My research took off from this observation on how to provide critical information services to society and was tasked with finding novel approaches to ensure that all other 5G era SCS systems are fit for purpose[2]. Some examples, such as driverless cars and remote-assisted surgery, are *prima facie* examples of safety critical and mission critical services. Some, such as Industry 4.0 services, are business critical as they underpin the transformation that will inform the next phase of industrialisation. Some, such as providing critical information with a 999 app, have yet to awaken society's consciousness. Indeed, as is often seen at major sporting events in the UK, it seems society is still comfortable with a police officer on horseback shouting instructions through loudspeakers to a crowd. Security critical systems are out of scope in this work.

However, designing 5G for SCS is happening at the same time as a boom in mobile data traffic. This is fuelled by exponential growth in video usage and exacerbated by the behavioural adaptations to the Covid-19 pandemic. For 2021, Ericsson estimated that the total global mobile data traffic will reach 65 exabytes/month with each smartphone user accounting for 11.4 GB/month; video traffic accounted for 69% of all mobile data traffic and that this will rise to 79% by 2027[12]. A classic industry response to the traffic boom is to add more mobile network capacity to increase the peak capacity of the network [13–15], albeit with economically disadvantageous overcapacity in public networks. For example, in 62 out of the 64 European mobile networks evaluated by Rewheel in 2017, average radio network capacity utilization in the 5% of their most loaded cells was less than 25% [16].

The prevailing industry expectation is that increased network capacity inevitably leads to improved network reliability and user quality of service (QoS) on public networks. This undoubtedly helps. But as 4G shows, extra capacity is not enough to eliminate all concerns about network reliability and user QoS [17–20]. Thanks to even more capacity and other design changes, 5G ought to be more reliable than previous generations of mobile networks [14]. For example, ITU-R specifies a minimum reliability requirement of $1\text{-}10^{-5}$ success probability of transmitting a layer 2/3 packet within a required maximum time for 5G [21]. To provide further reliability assurances, there is also a big effort to design for a new category of "Ultra Reliable and Low Latency Communication (URLLC)" 5G services [22]. Revisions in 3GPP Release 16/17/18 will pave the way for URLLC services[23].

In this work, I investigate if extra capacity and design improvements, for the overall network, can be relied on to always provide the required consistency in network reliability for 5G era SCS. This could be during *typical* scenario (e.g. normal day) or *atypical* scenarios (e.g. crowded events such as New Year fireworks or natural/man-made disasters such as during the 2022 Russian-Ukrainian conflict[1]). If not, my investigation shifts to identify how network management and operations can be adapted or tailored to deliver improved QoS to specific SCS use cases or the traffic from specific SCS use cases. I describe these efforts as the ***Differentiated QoS (D-QoS)*** approach.

Scanning through the tools and mechanisms at the disposal of telecoms industry stakeholders, I identify three primary D-QoS mechanisms to improve reliability. The first is by prioritising traffic from selected use cases. Examples of this includes Integrated Services (IntServ), Differentiated Services (DiffServ) and 4G Quality Control Indicator (QCI). The second is by using redundant multi-connectivity, enabling a specific use case to be served by multiple connectivity options. This approach is already very pervasive for consumer smartphones as part of a *price or quality arbitrage* mechanism with WiFi or multi-SIM devices[24]. However, our D-QoS focus is on *reliability arbitrage* for society-critical use cases such as CCs. The third is the use of exclusive infrastructure for selected users or use cases. Examples of this are Private 4G/5G networks and Leased Lines which provide

[1] https://blog.adaptivemobile.com/the-mobile-network-battlefield-in-ukraine-part-1?hsLang=en

exclusive physical infrastructure. By providing only *logical* exclusivity, 5G Network Slicing straddles options 1 and 3. We note in [25], that despite the wide availability of D-QoS since the 1970s, their commercial history has been relatively unsuccessful and the prevailing Net Neutrality regulatory paradigm of treating all data traffic equally do not encourage the use of D-QoS on public networks.

1.2 Scope of Using D-QoS for SCS

There is, as yet, little clarity on what constitutes an SCS, especially for services that rely on the mobile telecommunications network. While several electro-mechanical systems can be described as safety critical, this work goes further to create an all-encompassing class of services/systems which are critical for safety and the proper functioning of society.

Increasingly, there are commercial solutions that are seeking to deliver reliability assurance for SCS that require or rely on mobile connectivity. Examples include the ‘Emergency Alert’ in Android and iOS smartphones or the ‘Mark myself safe’ function on Facebook. However, a central argument in this work is that, despite the merits of these commercial solutions, they are insufficient to deliver the requisite enforceable assurances that is required for SCS.

In practice, not every SCS requires or relies on mobile connectivity and Table 1.1 describes the five main types of SCS and the types of services/systems which rely on mobile connectivity and therefore ought to be considered as deserving a D-QoS. A full exploration of the different types of SCS is presented in Chapter 2.

1.3 Thesis Statement

Given the reality about the inadequacy of existing public networks to provide reliability assurances for 5G era SCS, and the option of using D-QoS (despite its less than successful history and the current constraints with Net Neutrality), my thesis is that -

> A government-supported Differentiated QoS approach will provide improved QoS guarantees on public mobile networks for society critical systems in the 5G era.

In this report, I interrogate the thesis statement and distill it into investigate-able research goals and questions.

Table 1.1 SCS Services/Systems and D-QoS

SCS Type	**Description**	**Examples requiring or relying on mobile connectivity**
Safety Critical Systems	Systems whose failure or malfunction can lead to loss of life, serious injury to people or damage to property and the environment	Connected road vehicles (whether fully or partially autonomous), Smart Cities solutions (including cameras, signage/displays, sensors), robots & autonomous guided vehicles (in factories, mines, ports), Smart Healthcare (including Smart Ambulances, remote assisted consultation/diagnostics/surgery), drones/unmanned aerial vehicles.
Mission Critical Systems	Systems that are fundamentally essential to the success of a specific operation and their failure or malfunction will generally lead to an inability to complete the task	Any component, equipment or software that is critical to a mission.
Business Critical Systems	Systems whose failure or interruption can lead to tangible and intangible economic losses, customer dissatisfaction and inefficiencies in productivity for a business	Hardware and software solutions that deliver the Operational Technology, Connectivity technology and Information Technology solutions for a business (incl. IoT solutions and for Industry 4.0).
Critical Information Systems	Systems that deliver society's trusted and verified information/news sources to ensure that critical and important information is effectively communicated	Government-approved/sanctioned news on TV, Radio, newspapers, official government apps, Public Warning Systems.
Security Critical Systems	Systems that protect assets (including physical and digital assets) from threats, losses and attacks, regardless of whether they are intentional, accidental or malicious	Any hardware or software which is exposed to threats and attacks.

1.4 Research Goals & Questions

Based on the thesis statement, I identify three primary *Research Goals* and their constituent *Research Questions (RQ)* that will be answered in this report.

Research Goal 1: *Investigate the adequacy of public mobile networks to provide QoS guarantees for society critical services*

Over the past 30 years, improvements in the reliability of mobile networks has enabled numerous data services: real-time notifications (e.g. Uber, Whatsapp), dependable notifications (e.g. SMS reminders), IoT (e.g. vehicle tracking), mobile money etc. Despite the ubiquity of these services (for both consumers and enterprises), it is rare to see enforceable QoS guarantees from any mobile network provider, nor has there been any effort to extend the 999 voice call guarantees to any data service. These lead to two research questions:

> **RQ 1.1:** How reliable are 3G and 4G networks to deliver QoS guarantees to end users?
>
> **RQ 1.2:** Will extra capacity and design improvements in public 5G networks be sufficient to provide consistent and enforceable QoS guarantees for SCS?

Research Goal 2: *Explore the history of D-QoS approaches to understand if and how they can be used to provide QoS guarantees for SCS.*

Since the introduction of Leased Lines in the 1970s, there has been a long history of trying to use D-QoS to improve the quality, reliability and 'tailorability' of data networks[25]. On mobile networks, the 4G QoS Class Identifier (QCI) was the first mechanism to become available to tailor mobile data services for the mass market and 5G Network Slicing is poised to come next. Yet, despite the technical effort in developing these mechanisms, there is a lack of sustained commercial success for most of them. These lead to two research questions:

> **RQ 2.1:** Why and how can D-QoS mechanisms deliver better QoS guarantees for 5G era SCS?
>
> **RQ 2.2:** Given its relatively unsuccessful commercial history and current net neutrality paradigm, how should D-QoS be positioned for success?

Research Goal 3: *Evaluate the reality of D-QoS approaches to understand their relative advantages in providing QoS guarantees for SCS*

It is important to quantify the potential impact of any of the three D-QoS approaches to better understand how they can be used to improve QoS guarantees for 5G era SCS. So far, there has been more industry efforts on Network Slicing (a prioritisation approach) and the use of Private 5G networks (an exclusive infrastructure approach) but with little effort on using redundant multi-connectivity as a way to provide QoS guarantees to selected use cases. These lead us to three research questions:

> **RQ 3.1:** If D-QoS is a feasible approach, what are the options and implementation approaches to using D-QoS to provide QoS guarantees to SCS?
>
> **RQ 3.2:** What are the quantifiable benefits of using any of the D-QoS approaches on public mobile networks for selected use cases?
>
> **RQ 3.3:** How can key industry stakeholders (including policy makers) support the application of D-QoS?

1.5 Methodology

During my PhD studies, I have worked with colleagues and collaborators to validate the thesis statement and answer the research questions. Broadly, the approach was to understand the nature of performance in 3G/4G/5G networks, understand the applicability of D-QoS, and then conceptualise a framework for action. I break this down into five research steps :

1. **Field Measurements 1:** Conducted extensive, multi-year field measurements on mature 3G/4G networks in the UK to investigate the reality of user-perceived QoS. This is aimed at addressing RQ 1.1.

2. **Field Measurements 2:** Conducted extensive measurements on early 5G networks in the UK and South Korea to understand performance differences with mature 3G/4G. This is aimed at addressing RQ 1.2.

3. **Critical Review:** Underwent a historical review of D-QoS standards from 1970s to 2020s, to determine the criteria for success of D-QoS, including the role of official support. This is targeted at RQ 2.1, 2.2 and 3.1

4. **Field Measurement 3:** Implemented a multi-redundant connectivity system for a Connected Car and tested it along over 800 kilometers of roads in South East England. This provides a redundant multi-connectivity example for RQ 3.2.

5. **Framework Modeling:** Conceptualised a framework for action to guide commercial and policy stakeholders on when and how to use D-QoS. This is aimed at RQ 3.3.

1.6 Contributions

Based on the work that I have done with my colleagues/collaborators during this PhD, I make four contributions in this report.

- **QoS reality on public 4G/5G networks is inadequate for SCS:** Public 4G/5G networks are unable to deliver consistent and adequate connectivity QoS guarantees for SCS in the 5G era. Using latency of 50ms as a proxy for performance for SCS, our field experiments on 4G networks in London on four New Year Days, plus additional January dates from 2016/17 - 2019/20, show that a user has only a 58% chance to get a latency below 50ms. Likewise, our measurements on early 5G networks in London (UK) and Seoul (South Korea) in 2020 & 2021 show that a user has only a 50% chance to get a latency below 50ms. If mature 4G networks and early 5G networks are unable to provide consistent QoS that is suitable for SCS, then it follows that mature 5G networks will likely perform similarly.
- **D-QoS can improve QoS consistency and adequacy for SCS:** Officially-supported D-QoS mechanisms can support public 4G/5G networks to provide QoS guarantees for SCS. Based on our review of D-QoS mechanisms since the 1970s, we demonstrate that the use of D-QoS for purely commercial reasons have generally failed. However, for SCS that are integral to the smooth functioning of the digital society and economy, it is imperative that policy makers provide support for the use of D-QoS in specific and limited scenarios. We provide a taxonomy to identify classes of use cases that merit a D-QoS approach and introduce the CLASP (Critical, Localised, Approved, Specific & Perishable) framework to guide policymakers on how to make decisions.
- **Field evaluation of D-QoS mechanisms show their merit :** Our evaluation of a D-QoS mechanism shows that they deliver clear advantages with minimal disruption. In a system implementation and field measurements along over 800km of major/minor roads in South East England, we show that the use of redundant connections in CCs, preferably managed by the user on the demand side, can improve performance by up to 28 percentage points.
- **A D-QoS Guidance Framwork (DGF) for action:** I distill our work into a structured framework that can guide industry stakeholders on if, when, where and how to use a D-QoS mechansim to provide QoS guarantees for 5G era SCS.

1.7 List of Publications

During the seven years of my PhD, I was the lead author and main contributor for nine publications (published, accepted and under peer review) and one regulatory filing. In descending chronological order, these are:

1. (Under Review) Emeka Obiodu, Dongwook Kim, Abubakar Abdullahi, Aravindh Raman, Simone Mangiante, Nishanth Sastry and Hamid Aghvami. Characterizing 5G-enabled computing in the field: measurements & insights from South Korea & UK. In *Proceedings of the Internet Measurement Conference (IMC),* ACM, 2022.

2. Emeka Obiodu, Abubakar Abdullahi, Aravindh Raman, Puhskal Agrawal, Tooba Faisal, Sagar Joglekar and Nishanth Sastry. How Special is New Year Eve Traffic? Insights from Four Years 3G/4G/5G User Measurements. In *Proceedings of the 45th International Conference on Telecommunications and Signal Processing,* IEEE, 2022.

3. Emeka Obiodu, Aravindh Raman, Abubakar Abdullahi, Simone Mangiante, Nishanth Sastry, and Hamid Aghvami. DSM-MoC as baseline: Reliability assurance via redundant cellular connectivity in connected cars, *IEEE Transactions on Network and Service Management,* 2022.

4. Emeka Obiodu, Abubakar Abdullahi, Aravindh Raman, Nishanth Sastry, and Simone Mangiante. To share or not to share: reliability assurance via redundant cellular connectivity in connected cars. In *2021 IEEE/ACM 29th International Symposium on Quality of Service*. IEEE, 2021.

5. Emeka Obiodu, Abdullahi K Abubakar, and Nishanth Sastry. Is it 5G or not? Investigating doubts about the 5G icon and network performance. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6. IEEE, 2021.

6. Emeka Obiodu and Nishanth Sastry. From ATM to MPLS and QCI:The Evolution of Differentiated QoS Standards and Implications for 5G Network Slicing. *IEEE Communications Standards Magazine,* 4(2):14–21, 2020

7. Emeka Obiodu, Nishanth Sastry, and Aravindh Raman. Is it time for a 999-like (or 112/911) system for critical information services, In *NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium,* pp. 1–6, IEEE, 2020

8. Emeka Obiodu, Nishanth Sastry, and Aravindh Raman. CLASP: a 999-style priority lanes framework for 5G-era critical data services. In *2019 International Symposium ELMAR*, pages 101–104. IEEE, 2019.

9. Emeka Obiodu, Nishanth Sastry and Aravindh Raman. Towards a taxonomy of differentiated service classes in the 5G era. In *Proceedings of IEEE 5G World Forum, Santa Clara*, Jul 2018

10. (Regulatory filing) Emeka Obiodu, Aravindh Raman and Nishanth Sastry Contribution to BEREC's Consultation on Net Neutrality. *Consultation paper on the evaluation of the application of Regulation (EU) 2015/2120 and the BEREC Net Neutrality Guidelines*, April 2018 [2]

1.8 Thesis Overview

The rest of this report is structured as follows. In Chapter 2, the report provides a background to the principles, technologies, policies and methodologies that underpinned my work. Chapter 3 describes the methodology used in this PhD report, explaining the assumptions behind the thesis and the experiments conducted to investigate the thesis. Chapter 4 focuses on the field measurements conducted on the four UK mobile networks on typical and atypical days over four years, plus 5G measurements in UK and South Korea. In Chapter 5, we explore the history of D-QoS and identify the changes required to make them work for 5G era SCS. Chapter 6 focuses on the evaluation for the D-QoS example and provides the results of our road drive field measurements. In Chapter 7, we distill all of the findings in this PhD into a structured framework that can be used by stakeholders to make decisions regarding the use of D-QoS for SCS. Chapter 8 provides a summary of our work and the opportunities for further work in this space.

1.9 Challenges

There were several challenges encountered in the course of this PhD. Below, I highlight the key ones.

- **Covid-19 disruptions:** At the start of our research, we projected to do field measurements on New Year for five years. We completed this for 2016/17, 2017/18, 2018/19

[2] https://berec.europa.eu/eng/document_register/subject_matter/berec/public_consultations/8012-consultation-paper-on-the-evaluation-of-the-application-of-regulation-eu-20152120-and-the-berec-net-neutrality-guidelines

and 2019/2020. Unfortunately, Covid-19, and the ensuing lockdowns, meant that there were no New Year Firework celebrations in London in 2020/21 and 2021/22. Consequently, this has forced our research to be based on only four years worth of data instead of five years.

- **Software and hardware hiccups:** Regardless of how much testing we did, software and hardware hiccups were a regular part of our field measurements. For example, while we officially say that we completed 800km of road test, the reality is that we drove for over 1000 kilometers and had to discard the data from some journeys because either the software or hardware failed enroute.
- **Personal challenges & tragedy:** As a part time PhD student, juggling my PhD with a full time job and family life was very challenging. There were periods when, due to work or family commitments, the PhD had to take a back seat. Worse still, I lost my wife in the first year of the PhD after a brief illness. The trauma, sense of loss/failure and demotivation at becoming a widower caused me significant psychological distress. I thank my supervisors, colleagues, friends and family for supporting me to persevere with the PhD studies.

Chapter 2

Background & Context

O kwesịrị ka a mata aha egwu tupu a gbawa ya (Igbo Proverb)

(It is good to know the brand of the music before dancing to it)

2.1 Introduction

The task of exploring QoS guarantees for SCS that rely on mobile connectivity is multi-disciplinary as it includes mobile network providers, telecoms regulators, enterprises and policymakers in charge of critical infrastructure. This multi-stakeholder view also means that the literature and references are diverse. In this chapter, I summarise the background of this multi-disciplinary and multi-stakeholder landscape to identify novel perspectives that provide holistic solutions to improve reliability for SCS. The chapter starts with an overview of SCS to understand their requirements and how society is supporting them. The second section focuses on the reliability & resilience in 4G/5G networks to set the scene on what to expect from our field measurements. In the third section, I provide a historical review of D-QoS mechanisms to understand their technical and commercial trajectory. The fourth section is a review of the literature on mobile network measurements to clarify the state of the art and the results that shape today's attitude to the reliability of mobile networks. The last two sections focus on providing a background to the two specific use cases that we evaluate in this report: Digital Public Information Services and Connectivity for Connected Cars.

2.1.1 Key questions answered

In this chapter, I review the background and the literature that inform my work. The starting point is to unbundle the title of this work to identify the different components and building blocks that make it up. From the title, three building blocks are identifiable: Society-Critical Systems; Differentiated Connectivity; and 5G era. This chapter provides a compendium of insights on these building blocks and lays the foundation for my exploration of all the research questions.

2.1.2 Chapter layout

This chapter is subdivided into six sections. Section 2.2 introduces the SCS concept and explores how it brings together five types of critical systems across society. Section 2.3 aims at explaining the key indicators for understanding mobile network performance and reliablity, and how they have been used in this PhD report. Section 2.4 focuses on the 5G era and what it means for mobile network performance. Section 2.5 introduces the differentiated connectivity paradigm together with its promises and challenges, including net neutrality. Section 2.6 focuses on how the use of redundant connectivity is an important part of differentiated connectivity. Lastly Section 2.7 wraps up the chapter by introducing the differentiated connectivity paradigm when applied to connected cars.

2.1.3 Publications linked to chapter

This chapter has been informed by all the papers listed in Section 1.7.

2.2 Society Critical Systems (SCS)

2.2.1 Introduction

SCS systems are characterised by their need for high reliability to avoid the significant negative consequences from their failure or malfunction[4]. In this work, I define an *SCS as any system whose failure or malfunction can lead to death or serious injury to people, loss or severe damage to property, significant tangible or intangible economic losses, or damage to the environment.* A crucial principle of all SCS is that they must work correctly or fail in a predictable or safe way. This leads to a choice of design options or *reliability regimes* to determine the course of action to take in the event of a failure. These include ‘fail safe’, ‘fail secure’, ‘fail passive’, ‘fail operational’, ‘fail soft’ and ‘fault tolerant’. SCS are constituent parts of the critical infrastructure for most countries. In that regard, political oversight comes from the respective ‘Critical Infrastructure Protection’ (CIP) in the country (e.g. European Commission’s directive EU COM(2006) 786[1]).

There are broadly five types of SCS:

1. Safety Critical Systems (SaCS)
2. Mission Critical Systems (MCS)
3. Business Critical Systems (BCS)
4. Critical Information Systems (CIS)
5. Security Critical Systems (SeCS)

2.2.2 Safety Critical Systems (SaCS)

These are systems whose failure or malfunction can lead to loss of life, serious injury to people or damage to property and the environment [5]. Within the context of 5G era use cases, this will include examples such as connected road vehicles (whether fully or partially autonomous), Smart Cities (including cameras, signage/displays, sensors), robots & autonomous guided vehicles (in factories, mines, ports), Smart Healthcare (including Smart Ambulances, remote assisted consultation/diagnostics/surgery), drones/unmanned aerial vehicles etc. SaCS focus on providing *functional safety* and regard safety as a higher priority attribute to reliability. This is illustrated by the example in [26]: a bus that does not start is unreliable but completely safe; conversely, if the brakes on the bus fails at high

[1]https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52006DC0786&from=EN

speed, the bus is both unreliable and unsafe. In practice, concerns that mobile connectivity (including 5G) is unable to support such functional safety levels is forcing designers to be pragmatic (see Connected Cars example[2]).

SaCS originated in the manufacturing and process industries and are focused mostly on hardware, software and overall system safety for electronic and electro-mechanical equipment. Their requirements have been progressively formalised and developed since the 1970s in the UK and across Europe: BS 5304 “Code of practice: Safeguarding of machinery” for UK and BS EN 954-1:1997 “Safety of machinery" for Europe[27]. Originally, both BS 5304 and EN 954-1 were based on non-programmable elements. However, from the 1980s, increasing use of computer-based controllers and Programmable Electronic Systems (PESs) led to increased focus in the UK and Europe on programmable elements. This culminated in the International Electrotechnical Commission (IEC) 61508 standard, titled "Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems (E/E/PE, or E/E/PES)"[28] which was first published in 1998 and revised in 2010. IEC 61508 applies to all industries although there are other standards derived from it for specific industries (e.g. ISO 26262 for Automotive Electric/Electronic Systems, IEC 62279 for software used in railway operations, IEC 61511 for instrumentation in process industries)[29]. Broadly, SaCS would come under the discipline of *Safety Engineering*, and regulatory oversight will be provided by safety organisations such as the UK’s Health and Safety Executive (HSE).

Safety-Critical Systems and IEC 61508

IEC 61508’s overarching philosophy is that functional safety is best addressed in the entire safety lifecycle[30]. To drive this, it promotes the concept of the ‘Safety Life Cycle’ as an engineering process to avoid design errors and omissions. IEC 61508 approaches risk of failure as a probability. It opines that the probability of risk can be reduced, but zero risk can never be reached. As such, it promotes the principle of "ALARP" (as low as reasonably possible) for non-tolerable risk. The IEC 61508 principles are codified into the *Safety Integrity Level (SIL)* metric that is used as a design benchmark for each safety function. SIL levels are determined through a risk assessment that evaluates the systematic capability of design quality, architectural constraints regarding minimum levels of safety redundancy, and the probability of Dangerous Failure Analysis. Depending on the latter, there are four SIL levels that provide guidance on the design and provision of functional safety. These are summarised in Table 2.1. IEC 61508 calls for the use of either qualitative or quantitative hazard and risk analysis techniques. This work uses one of the suggested qualitative risk classification frameworks (Figure 2.1) to evaluate SCS use cases.

[2]https://www.bbc.co.uk/news/business-45048264

Table 2.1 IEC 61508: Safety Integrity Levels (SIL) benchmarks

SIL	Low demand mode: average probability of failure on demand	High demand or continuous mode: probability of dangerous failure per hour
1	$\geq 10^{-2}$ to $< 10^{-1}$	$\geq 10^{-6}$ to $< 10^{-5}$
2	$\geq 10^{-3}$ to $< 10^{-2}$	$\geq 10^{-7}$ to $< 10^{-6}$
3	$\geq 10^{-4}$ to $< 10^{-3}$	$\geq 10^{-8}$ to $< 10^{-7}$
4	$\geq 10^{-5}$ to $< 10^{-4}$	$\geq 10^{-9}$ to $< 10^{-8}$

	Frequency of Occurrence	**Consequence**			
	Definitions & (failures per year)	**Catastrophic** *Multiple loss of life*	**Critical** *Loss of a single life*	**Marginal** *Major injuries to one or more persons*	**Negligible** *Minor injuries at worst*
1	**Frequent** *Many times in lifetime (> 10^{-3})*	I	I	I	II
2	**Probably** *Several times in lifetime (10^{-3} to 10^{-4})*	I	I	II	III
3	**Occasional** *Once in lifetime (10^{-4} to 10^{-5})*	I	II	III	III
4	**Remote** *Unlikely in lifetime (10^{-5} to 10^{-6})*	II	III	III	IV
5	**Improbable** *Very unlikely to occur (10^{-6} to 10^{-7})*	III	III	IV	IV
6	**Incredible** *Cannot believe that it could occur (< 10^{-7})*	IV	IV	IV	IV

Where:
- *Class I: Unacceptable in any circumstance*
- *Class II: Undesirable: tolerable only if risk reduction is impracticable or if the costs are grossly disproportionate to the improvement gained*
- *Class III: Tolerable if the cost of risk reduction would exceed the improvement*
- *Class IV*: Acceptable as it stands, though it may need to be monitored

Fig. 2.1 IEC 61508: Qualitative Risk Classification Framework

2.2.3 Mission Critical Systems (MCS)/Business Critical Systems (BCS)

MCS are fundamentally essential to the success of a specific operation and their failure or malfunction will generally lead to an inability to complete the task[6]. When an MCS is needed for a business to carry out its normal operations, it becomes a BCS. Failure or interruption in a BCS can lead to tangible and intangible economic losses, customer dissatisfaction and inefficiencies in productivity[7]. Broadly, MCS/BCS can include any component, equipment, personnel, process, procedure or software that is critical to a mission or business. But in this work, the focus is on hardware and software for electronic and electro-mechanical equipment. Within the context of 5G era, many of the use cases promoted for industry verticals, especially under Industry 4.0, could transform 5G connectivity into BCS. A vision of Industry 4.0 is that 5G connectivity binds cyber-physical systems, Internet of Things (IoT) nodes, on-demand & elastic computing resources at the cloud/edge and Artificial Intelligence (AI) driven cognitive computing to transform the industrial landscape[2]. For example, in a Smart manufacturing factory, 5G can be used as the Connectivity Technology (CT) fabric for the factory's Operational Technology (OT) for controlling physical devices, and to integrate with the Information Technology (IT) systems for managing factory data.

The integration of CT, OT and IT into the same 5G ecosystem in an industrial scenario opens a new front on how to assure *Business Continuity* in the event of a connectivity failure or malfunction. Business Continuity is focused on how to help organisations to avoid disruption and on how to recover quickly in the event of a disruption to continue the delivery of products or services at pre-defined acceptable levels[31]. While there were originally many national and industry-specific guidelines, ISO 22301 has become the key framework, since 2012, to plan, establish, implement, operate, monitor, review, maintain and continually improve a Business Continuity Management (BCM) system[32]. Business Continuity is related to risk management - the former evolved from IT and disaster recovery, while the latter - presented as Enterprise Risk Management (ERM) and codified under ISO 31000 - has its roots in insurance, loss control and compliance.

MCS/BCS, Business Continuity and ISO 22301

If 5G becomes an MCS/BCS, it intensifies the vulnerability of a business to operational incidents and interruptions to operation activities due to failures or unreliability of connectivity. This has implications for corporate governance. Put simply, the 5G-enabled CT/OT/IT elevates CT from the IT manager's remit to a C-level decision. The preparedness of business leaders to take on this additional responsibility is an often overlooked consideration in the promotion of 5G for industry verticals - an interview of 30 enterprise leaders in [2] reported that some enterprises are not sure of the 5G value proposition and may not rely on it for their product roadmap. Likewise, for business leaders who traditionally own their OT or maintain very stringent service level agreements (SLAs) with their suppliers, it will be challenging to rely on an unreliable connectivity provider for their BCS OT. This explains why 5G Network Slicing is likely to struggle to gain traction while potential industrial target customers opt for Private 5G networks[25].

2.2.4 Critical Information Systems (CIS)

CIS are society's trusted and verified information/news sources to ensure that critical and important information is effectively communicated[8]. Traditionally, this has been the job of the national broadcaster (e.g. BBC in the UK). But the nature of information gathering and dissemination has changed dramatically over the past 30 years with the rise of digital media, increasing disinterest in traditional media houses, growing cacophony of social media channels and questionable integrity of some information sources. In this emerging digital landscape, it has become imperative to evolve the distribution medium for critical/important information that is currently prioritised on TV, radio and newspaper (e.g. warning/emergency/safety information, traffic/weather news, governmental information). A particular characteristic of CIS is that it delivers localised and actionable information directly to specific individuals or groups. To ensure uniformity of all messages on all platforms, the Common Alerting Protocol (CAP) has been developed as the global standard on how to coordinate and harmonise public warning and emergency alerts across alerting technologies. It is based on an XML data format and provides an open, non-proprietary digital message format for all types of alerts and notifications[33].

Using the CAP format, mobile phones, either via SMS or through Public Warning Systems (PWS) (e.g. Cell Broadcast or Wireless Emergency Alerts) are now well established for text-only information dissemination [34] and Smartphones are now offering a channel for multimedia CIS messages[35]. The UK finally launched its PWS 'Emergency Alerts' in early 2022 in response to the challenges the government faced in sending out SMS to the public at the start of the Covid-19 pandemic[3].

Broadly, there are two emerging limitations of PWS. The first is that, due to concerns about spamming individuals with irrelevant warnings, the threshold to use them is set very high [36, 37]. For example, in the US, the US Federal Emergency Management Agency (FEMA) was forced in early 2019, to significantly reduce the number of flood warnings to only 'considerable' or 'catastrophic' damage warnings because of a large number of

[3]https://metro.co.uk/2021/06/29/emergency-alert-system-uk-what-is-it-and-when-will-it-be-used-2-14845506/

complaints from the public[4]. The second limitation is that PWS is still a push service, whereas there is a growing preference by individuals to 'pull' content in the same way they do for on-demand media content. Some governments have responded to this with general emergency alert apps (e.g. '112 Suomi' in Finland, 'NINA' in Germany, 'Alertable' by several Canadian provinces and '311' by New York, USA) or alert apps for specific incidents (e.g. 'ShakeAlert' for earthquakes in the US). [38] provides a general survey of smartphone systems for emergency management. We group these government owned/approved apps as 999 (or 911/112) apps[8] and provide selected screenshots from them in Figure 2.2. To the best of our knowledge, ours is the first example in the literature detailing such a 999 app.

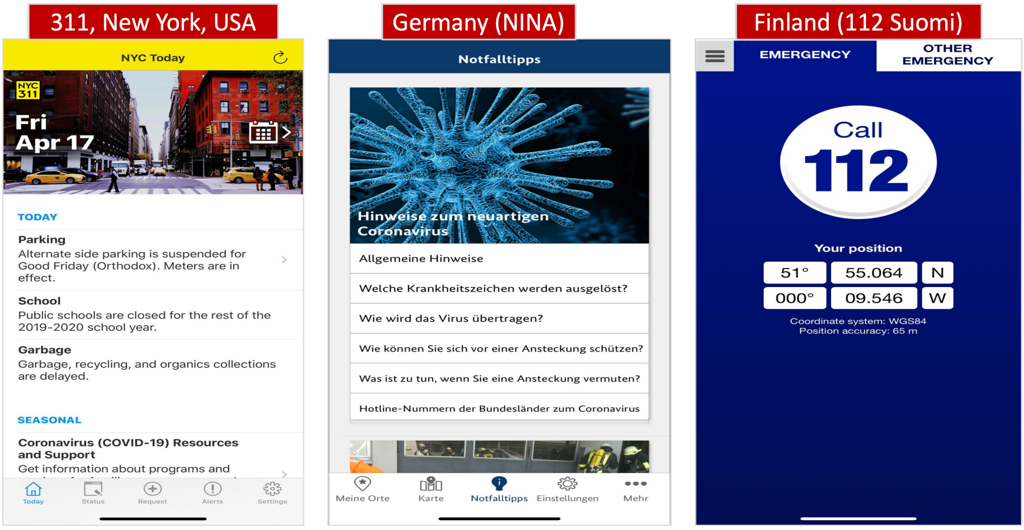


Fig. 2.2 Selected screenshots from three government owned emergency alert apps - in this work we call these 999 apps

As these 999 apps gain traction, we identify five key limitations (the first is mostly technical and the rest mostly socio-political):

- No claim nor expectation that these apps will always work if there is mobile network congestion in the location of the user.
- No local/national/global coordinated approach to validate and standardise a reliable 999 emergency alert app.
- No local/national/global effort to educate users on what to expect when using these apps (cf. most adults in a society are trained on how to use 999/911/112 numbers).

[4] https://www.weather.gov/media/wrn/FFW-IBW-factsheet.pdf

- No provision to zero rate the app to always make it accessible in 'Data off' scenario (i.e when the user does not have sufficient credit or has switched off their data) [39].
- Little awareness nor incentive for non-residents or visitors to download and use such apps in these locations.

Reliability assurance for a 999 Emergency Alert app

On the mobile network, there are broadly five mechanisms to provide reliability assurance for 999 apps and other CIS, even during congestion periods:

1. Prioritise CIS traffic: To assure their availability and reachability, prioritisation of CIS traffic can be enforced in the network by reserving special lanes for them or treating the CIS data packets as special packets.
2. Delay non-CIS traffic: Deprioritisation of non-CIS traffic or any traffic without an appropriate priority label can during congestion periods can be used to improve the availability and reachability of CIS.
3. Impose a speed/throughput cap for all or some traffic: Network operators can impose a speed/throughput cap on all services during congestion periods in order to improve throughput for some or all services (cf. average speed limits on 'Smart Motorways' during busy periods to increase the capacity of the road network).
4. Pre-fetch CIS traffic during periods of low network demand: CIS information can be pre-fetched and cached at the core/radio network or at the user device.
5. Changing how some traffic is billed: For customers still using metered Internet services, the cost of using non-CIS during periods of congestion can be varied in order to discourage their use and free up network resources for CIS.

A core theme in this work is to explore Point (1) that CIS such as the 999 emergency alert app ought to be prioritised in the same way that PWS is prioritised on the mobile network or traffic/weather alerts on radio and TV.

2.2.5 Security Critical Systems (SeCS)

SeCS protect assets (including physical and digital assets) from threats, losses and attacks, regardless of whether they are intentional, accidental or malicious[9]. SeCS have become crucial in many domains because of the growing threat and risk of cybersecurity. However, while SeCS may rely on 5G era connectivity, the reliability of the underlying connectivity is generally not a critical part of an SeCS. Hence, SeCS is out of scope in this work.

2.3 Understanding Mobile Network Performance & Reliability

2.3.1 Reliability & resilience in mobile connectivity systems

Reliability and resilience are fundamental design considerations for all complex systems in biology, computer science, ecology, economics, environmental science, engineering etc[40]. [41] clarifies that reliability is the ability of a system to perform as designed while resilience is the ability of a system to bounce back after a perturbation.

Reliability is a key mobile network KPI and there are two completely different expectations on it. The first is from the original ITU-R “Minimum requirements related to technical performance for IMT-2020 radio interface(s)" which specifies a minimum reliability requirement of $1\text{-}10^{-5}$ success probability of transmitting a layer 2/3 packet within a required maximum time [21]. Operationally, it is rare to see any mobile network being defined by this. Instead, system reliability is typically the user-experienced reliability for a sporadic end-user. For this, two key parameters are typical used to determine instantaneous network performance and over time, the user-experienced reliability of the network:

- Throughput, Bandwidth, Speed
- Latency, RTT, RTT Variance (Jitter)

2.3.2 Speed, Throughput and Bandwidth

Speed or throughput is the most popular and definitive performance indicator for mobile networks. It has shaped the narrative and expected performance of networks from 2G to 5G, helping to determine the services that are created for each mobile generation era[2]. Most comparisons of network performance use speed as the metric of choice and the same is the case with most measurement platforms and apps. Speed is also the metric of choice for policymakers for network performance analysis and benchmarking.

By definition, speed/throughput is the actual number of data packets that can be transferred from a source to a destination within a specified time. It is analogous to speed/velocity in mechanical systems and is often measured in megabits per second (mbps) or Megabytes per second (Mbps). In contrast, bandwidth is the maximum transfer capacity of the network and is governed by Shannon’s capacity formula.

In this PhD, we use the Android NetworkCapabilities API - getLinkUpstreamBandwidthKbps() and getLinkDownstreamBandwidthKbps() methods which return the speed of the connection in kbps and these are converted to Megabytes per second (Mbps). This PhD uses throughput and speed interchangeably.

2.3.3 Latency, RTT, RTT Variance (Jitter)

Latency is a key performance indicator for 5G era services and, perhaps, the most important performance indicator for ultra reliable and low latency (URLLC) services in the 5G era[2]. By definition, latency is the length of time it takes a data packet to travel between two end points. It is a crucial metric in networking because it impacts the user experience. Latency is impacted by the physical distance between the two end nodes, the transmission medium (e.g. 5G for wireless or fibre for wired) and the congestion in the network.

To overcome the challenge of checking the time stamp at both the sending and receiving nodes, latency is typically determined by the round trip time (RTT) which measures the length of time it takes for a data packet to travel from a sending node to an end node and back. Ping is the most common way to measure RTT and is implemented using the Internet Control Message Protocol (ICMP) protocol in the transport layer to TCP port 80. However, as external ping is often viewed by network administrators as a nuisance, ping is increasingly blocked or delayed by many websites or is redirected to Network Address Translation (NAT) ports.

Instead, in this PhD, we use java sockets and the InetAddress class to measure RTT by measuring the end-to-end latency from our devices to the webservers of selected websites from both TCP port 80 and TCP Echo Port 7. This ensures that we receive a response from all websites. We measure the RTT Variance or Jitter by taking multiple RTT measurements and then determining the standard deviation.

This PhD uses latency and RTT interchangeably.

2.4 5G-era Mobile Network Performance

2.4.1 Introduction

The 5G era has begun and signals the start of an era when 5G mobile connectivity, back-filled by 4G connectivity, is expected to support existing and new SCS use cases. Consequently, understanding the performance of the mobile network in typical and atypical scenarios is the first step towards determining if the networks can provide reliability guarantees for SCS. Generally, the overall telecommunications infrastructure (including the Internet) is often regarded as an SCS because of its importance to routine consumer and business engagements plus supporting emergency services. As a result, governments react strongly to any failure of the telecoms infrastructure. Examples include: in July 2022, Canadian authorities reprimanded Rogers as a software upgrade at Rogers triggered a 19 hours network outage[5]; in 2019, Dutch authorities severely criticised KPN for a 4-hour network failure that knocked out emergency telephones[6]; in December 2018, a software failure knocked out O2 UK's network, disrupting connected digital displays (for buses) and parking payment meters[7]; in November 2018, a fire at one of KT's Central Office in Seoul, South Korea caused widespread disruption as point-of-sale (PoS) units of many retailers failed[8]. However, while a broad remit for the resilience of the entire infrastructure has been sufficient for pre-5G use scenarios, it is unlikely to be sufficient for the SCS use cases that are being imagined for the 5G era.

2.4.2 5G expected performance

The expected improvements in 5G performance will come from improvements in both the access and the core network. For the access network, the ITU Radiocommunication (ITU-R) Study Group proposed the minimum requirements for the 5G radio interface in [21]. For the user, these technical requirements should translate to speed/throughput of between 10 - 100 times faster than 4G and latency of up to 10 times smaller than 4G [2]. Achieving this will however depend on which of the five 5G deployment options chosen by the operator (Option 2 & Option 5 for Standalone; Option 3, Option 4 & Option 7 for Non Standalone). While Standalone (SA) 5G should deliver the best performance when the 5G New Radio (NR) is connected to the 5G Core (5GC), all early commercial public 5G networks (as at December 2021) are based on Non Standalone (NSA) 5G which rely on connecting the 5G NR to the 4G Enhanced Packet Core (EPC)[42]. Figure 2.3 compares Option 2 for SA and Option 3 for NSA.

Meanwhile many current 4G networks are being upgraded to LTE-Advanced Pro, leveraging MIMO antennas to boost throughput [43]. These disparate deployment scenarios mean that it is difficult to establish a comparative benchmark between 4G and 5G because the intermediate steps of LTE-Advanced Pro and 5G NSA muddle the comparison. Based on [2, 21, 44] & [45], I summarise the performance benchmarks for 5G and 4G used in this work in Table 6.2.

[5] https://www.bbc.co.uk/news/world-us-canada-62174477

[6] https://uk.finance.yahoo.com/news/dutch-telecom-kpns-ceo-leave-074142457.html

[7] https://www.wired.co.uk/article/o2-down-network-problems

[8] https://www.voanews.com/east-asia/seouls-telecom-outage-highlights-need-redundancy-connected-world

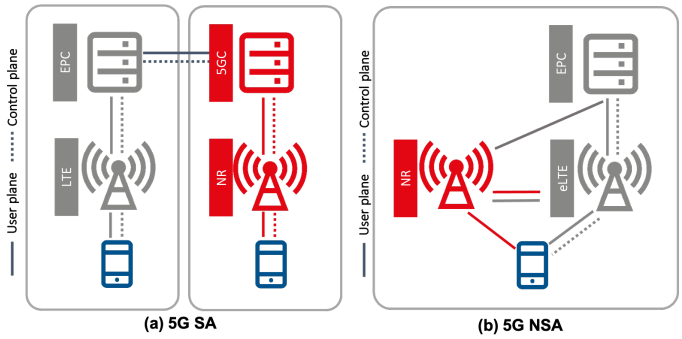


Fig. 2.3 5G NSA vs SA. In (a) 5G SA (3GPP Option 2), device is connected to 5G new radio (NR) and the 5G icon is shown by default. In (b) 5G NSA (3GPP Option 3), device is connected to both the 4G & 5G via Dual Connectivity & 5G icon decision is determined by GSMA recommendations[46]

Table 2.2 Benchmark performance for 4G & 5G

Metric	4G	5G
Uplink speed	25 Mbps	50 Mbps
Downlink speed	50 Mbps	100 Mbps
Packet Loss	0%	0%
Round Trip Times (RTT)	100 ms	50 ms
RTT Variance	20 ms	10 ms
Page Load Times (PLT)	1000 ms	500 ms

2.4.3 5G performance drivers

For the core network, improvements in computerisation and the adoption of a fully virtualised 5G core will boost performance. This explains the expectations for a better 5G experience once the 5GC is deployed in the 5G SA deployment configuration. For the access network, improvements will require developments in three spectrum-related areas (see Figure 2.4) plus the transport/backhaul network. These are described below:

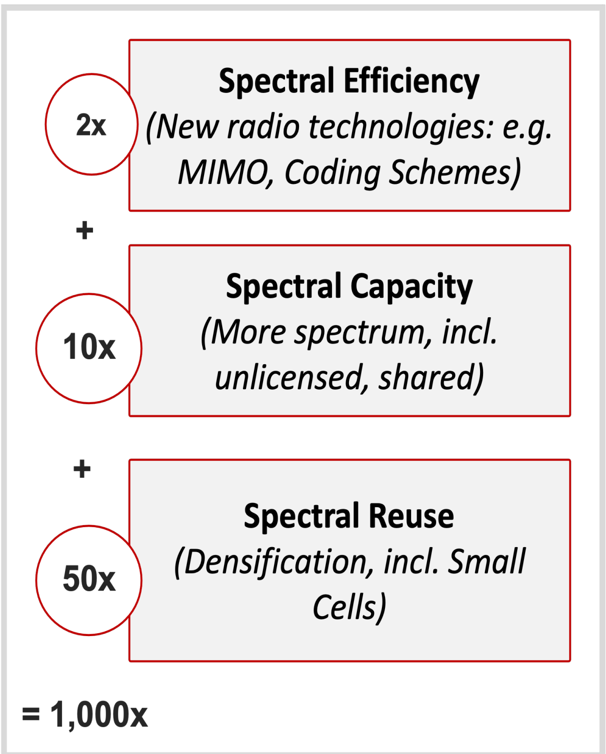


Fig. 2.4 Three spectral outcomes that define 5G capacity. [13] notes that these can yield a 1000x improvement over 4G.

- **Spectral Efficiency**: This will come from using more efficient radio technologies to eke out further gains closer to the Shannon bound, especially by using more than one communications channel. These innovations include massive MIMO (massive input massive output), beamforming and efficient coordination of adjacent cells to minimize interference[47]. However, [2] notes that the gains from 4G to 5G will be less than the gains from 3G to 4G because the 4G spectral efficiency is already close to the theoretical maximum. This explains why early work on 6G is looking into the use of semantic communications[48] and machine learning[49].

- **Spectral Capacity**: Simplistically, more spectrum bandwidth equals greater network capacity.

$$C = Blog_2\left(1 + \frac{S}{N}\right) \tag{2.1}$$

  Where C is the channel capacity, B is the spectrum bandwidth used and $\frac{S}{N}$ is the signal to noise ratio of the channel.

  This explains the concerted push to unlock and issue up to 100MHz in the sub 6GHz C band and up to 1000MHz in the mmWave band (above 24GHz band) for each public 5G network. For example, [2] shows that cell throughput in 5G networks when using 30MHz is about a third of the throughput when using 100MHz. The GSMA has codified ten '5G spectrum positions' for the industry to drive all stakeholders to achieve these goals[50]. Ideally, these spectrum bands should be contiguous but up to five disjointed 20MHz spectrum bands can be aggregated using carrier aggregation technology[51].

- **Spectral Reuse**: A key attribute of the mobile network is that it is arranged in cells making it feasible to re-use frequencies to increase both coverage and capacity. A high reuse factor equals a denser network. Hence, simplistically, the higher the re-use factor, the greater the network capacity. Combined with the poor propagation characteristics of mmWave, the drive to higher spectral reuse is the primary reason for the growing use of small cells for 5G networks. ABI Research forecasts that overall number of mobile cells will grow from 11.8million in 2017 to 18.4 million by 2025; small cells will grow by 25% from 0.7 million in 2017 to 4.3million by 2025[52].

- **Backhaul / Transport**: 5G networks will need much larger backhaul capacity to evacuate the data traffic from the expected higher network capacity. Failing to do so creates the potential risk of a network bottleneck. Fibre optics cables is the de facto option for increasing backhaul capacity for mobile cell sites. However, given that fibre will not always be feasible or viable, alternative solutions will also play a role. These include satellites (especially Low Earth Orbit (LEO) constellations) and wireless links - microwave (7-40GHz), V-band (60GHz) or E-band (70/80GHz).

2.4.4 5G actual performance

There are both press commentaries and growing peer-reviewed studies on the actual performance of 5G. The former can be effusive while the latter is a bit more circumspect. For the former, there are already several press articles and industry commentators saying that 5G

NSA achieves much faster throughput than 4G. For example, [53] claims that UK network operators were achieving real world download speeds of between 150Mbps to 1Gbps compared to 20-30Mbps for 4G. Likewise, measurements in 5G test-beds suggest significant performance improvements over 4G[54]. However, the emerging peer-reviewed publication scene suggests a less than stellar performance for 5G, so far, in the field.

In the first of its kind paper looking at 5G performance in the field in the US (both mmWave and sub 6GHz), [55] makes three observation: first, NSA 5G offers only a little latency improvement over 4G due to the reuse of 4G infrastructure; second, mmWave 5G offers higher throughput than 4G ($\sim$ 10x improvement) but exhibits significantly higher variation in throughput because of the nature of mmWave signals; third, application QoE does not benefit fully from the high throughput of mmWave 5G because there are many cross-layer factors that require joint optimisations. Based on measurements in China (sub 6 GHz only), [56] corroborated the latency and application QoE shortcomings of 5G. In addition, they note that 5G quality drops sharply in going from outdoors to indoors and that the 5G radio hardware has an alarmingly high power consumption of 2 - 3× over 4G. Based on measurements in the UK (sub 6 GHz), we observe in [57] that the worst cases for latency and uplink/downlink speeds were minimised in 5G compared to 4G but the best case performance was the same on 4G and 5G devices.

2.4.5 5G perception and ‘Fake’ 5G icons

Regardless of the delta between expected and actual 5G performance, there were reports in the press in late 2019 and early 2020, especially in the US, that mobile operators were offering a ‘fake’ 5G icon[58]. These reports triggered user/press concerns about ’fake’ 5G icon or notifications in a way that was not seen for 1G/2G/3G/4G. While the choice of icon to display is ultimately a commercial decision for each operator, the specification and design is determined at an industry level. In this case, the industry level discussions included publicly-recorded disagreements, debates and delays[58], forcing 3GPP into a rearguard action to retrospectively provide clarification[46]. Between 2018 - early 2020, 3GPP and the GSMA sought to provide industry guidance on what and when the 5G icon should be used and efforts were made to investigate compliance with the agreed 3GPP/GSMA position on the 5G status indicator[46].

2.4.6 4-stage framework for determining 5G icon status

Given the concerns about 'fake' 5G icons, we introduce a four-stage analytical framework to provide guidance of how decisions on 5G icons are made for open-market or off-the-shelf Android devices.

1. Network-level decision

There are two major decisions, by the network provider, on whether 5G is available or not for use in any given market: firstly, the 5G network has to be available and then secondly, users have to be notified of its availability.

5G 'availability' covers the actual deployment of a 5G NR (New Radio) radio, in the appropriate frequency bands that will be supported by devices in a market. It also includes the interrelationship of the 5G NR with the LTE radio, the 4G EPC and the 5G core network (5GC), if available. For 5G SA, this decision is straightforward as the 5G NR, running on a 5G base station (gNb), is connected to a 5GC. However, for 5G NSA, the decision is based on 'dual connectivity' (DC), a mechanism introduced in 3GPP release 12 to permit connectivity to both a macro LTE base station and a small cell[59]. In 5G NSA option 3, using DC means that a device can connect simultaneously to both LTE and 5G NR; LTE manages the control plane while the NR manages the data plane.

'Notification of availability' of 5G is set by the *upperLayerIndication* bit, as contained in 3GPP TS 36.331 specifications on Radio Resource Control (RRC)[60] and clarified in the discussion paper R2-1801529[61]. Under the guidance, each mobile network operator - i.e. Public Land Mobile Network (PLMN) - adds a 1 bit NR indicator to the System Information Block type 2 (SIB2) to indicate that a 5G cell is co-located and that a user has entered a coverage area that offers 5G capabilities. The *upperLayerIndication* is set as *true* when 5G co-location is available or, otherwise, is omitted.

2. User Equipment (UE) level decision

A UE will support 5G if it has the appropriate 5G chipset (e.g. Exynos 9820 Octa and Qualcomm Snapdragon X55 5G) and supports any one of the C/mmWave 5G NR frequency bands. The network uses the 'UE capability enquiry and information' mechanism to determine if UE can support 5G[60]. This is done during the LTE/5G attach procedure when the network sends an RRC 'UE Capability Enquiry' message to the UE and receives the 'UE Capability Information' from the UE detailing all the information that the UE is capable of.

Once established, the UE can then decide on the 5G icon based on the *upperLayerIndication* status. However, while the network informs its 5G status via the *upperLayerIndication*

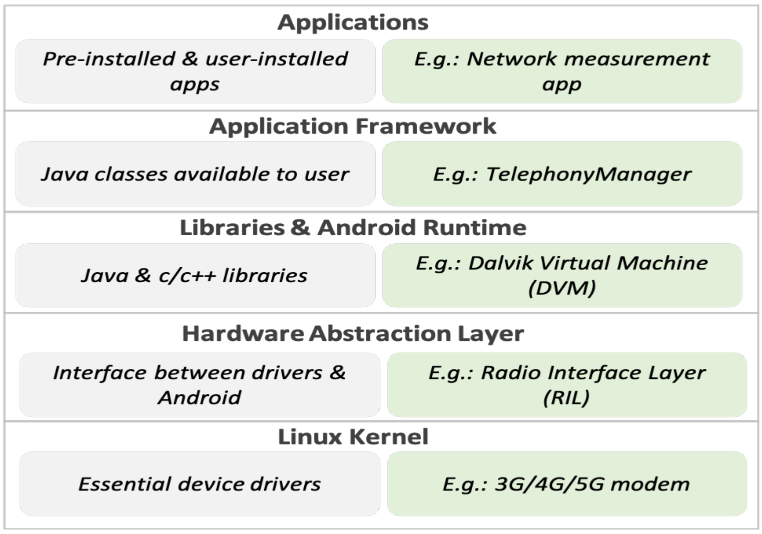


Fig. 2.5 Android Architecture Stack. Android provides a TelephonyManager API package which reports a value of 20 for 'Network_Type' when the OS is connected to a 5G radio.

bit, the UE relies on a not-so-definite decision framework for its decision. This is because 3GPP has left it to individual operators to decide how to implement *upperLayerIndication*. After several iterations, and after early 5G devices have shipped causing confusion about the 5G icon for users, in a Liaison Statement (LS) to 3GPP in January 2020, the GSMA's 5G Status Indicator (5GSI) taskforce proposed that when a UE is connected in an active mode to a 5G NR, the 5G icon should be shown[46]. Conversely, when the UE is in idle mode, it will only show a 5G icon if it is in an LTE cell that is capable of 5G NSA and has been notified that it is in an area where 5G is available. Figure 2.6 summarises the GSMA 5GSI guidance.

3. Operating System (OS) level decision

The OS provides a linkage between the low level hardware functions in a UE and the higher lever application stacks. We focus on the Android OS, the predominant mobile device OS/platform in the world ($>$70% market share) and use its 5-layer architecture to understand how the OS will interpret 5G availability - Figure 2.5.

A UE with the requisite 5G chipset will have vendor-specific modem drivers managed by the Linux kernel. In the hardware abstraction layer, Android runs a Radio Interface Layer (RIL) daemon to provide an interface with these device drivers[62]. Within the application framework, Android provides a TelephonyManager package or sets of classes which exposes APIs that an app can use to interact with the UE hardware and system information[63]. While there are several parameters that report on 5G connectivity in the TelephonyManager

	Use 5G icon	Do not use 5G icon
Idle	The UE shall be able to display a 5G icon when it is in idle mode in an LTE cell capable of NSA and is notified that it is in an area where 5G is available.	The UE shall not display a 5G icon when all the following conditions are met: a) The UE is camped on a base-station capable of NSA; and b) The UE has been provided with a list of NR frequency band(s) deployed in the area; and c) The UE supports none of the NR frequency bands that are deployed in the area or cannot use NR for any other reason known to the UE.
Active	The UE shall display a 5G icon when the UE is in active mode and is using NR. Regarding the length of time the 5G icon is displayed, operators should seek to strike a balance between a frequently-changing icon and miscommunicating the connectivity status.	

Fig. 2.6 Summary of GSMA recommendation to 3GPP on when to show a 5G icon[46]. The guidance recommends how the User Equipment (UE) will make a decision to show a 5G icon based on whether it is in idle mode or active mode.

package, the ‘Network_Type_NR’, added in Android 10 - API Level 29, returns a numerical value of 20 to indicate that the UE is connected to a 5G New Radio.

In practice, while the UE’s system information log will report 5G connectivity based on *upperLayerIndication*, the RIL reports 5G based on the connectivity established in the modem. Accordingly, Android OS reports a ‘Network_Type_NR’ = 20 only when the RIL reports connection to a 5G NR.

4. Application level decision

While there maybe temporal, situational and contextual variations in network quality and performance[64], the default capability of the underlying network connection will impact application behaviour and performance. Thus, we expect that, for applications that require connectivity, performance will progressively improve from 3G to 4G to 5G. We focus on the two performance measures that are predominantly used in the literature to describe network quality: latency/delay and speed/throughput. Applications connected to 5G ought to perform better than those connected to the best performing 3G/4G networks and should experience latencies of $<$10ms and speeds in excess of 100Mbps[2]. However, and as shown by [55] in their first look at 5G performance in the field, 5G NSA offers only a marginally better latency than 4G.

2.4.7 Determining factors for network performance

Regardless of whether it is 5G or 4G, in most locations, the mobile network will have been adequately designed and dimensioned to cope with peak demand on a typical day. But this is often not enough to assure reliability because the nature of demand on the network is stochastic and varies greatly across temporal [64] [65], spatial[57] and situational/contextual [66] [67] [68] scenarios. This can be seen in Figure 2.7 based on measurements taken on a typical day in Central London shows. In fact, [69] notes that download speeds can drop by 50% or more during peak hours. The expectation is that this variation will be worse on atypical days and crowded events where unpredictability in user numbers and behaviours does not help in network planning, even with the use of ad-hoc, portable base stations - e.g. Cells on Wheels (COWS). On a more granular level, there are five factors that cause variation in network performance: congestion caused by other users, radio technology, coverage, handovers and measurement device [68].

Congestion caused by other users is already a well-established determining factor for network performance as it varies greatly in time, location and situations. The underlying principle is that there is a finite capacity for the network as determined by the Shannon bound

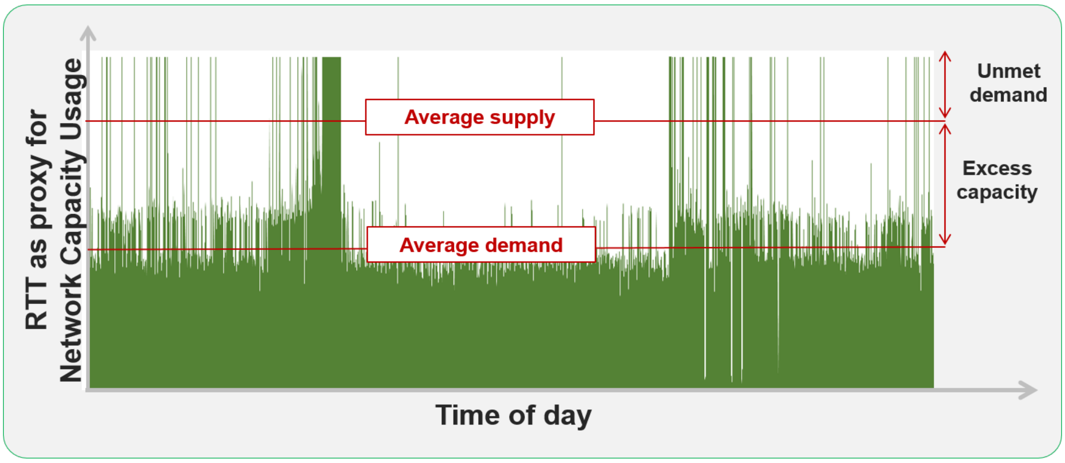


Fig. 2.7 Network performance varies stochastically across temporal, spatial and situational/contextual scenarios. Chart is based on an actual 24 hours Round Trip Time (RTT) measurements in a fixed location in London, UK from 12:00 07/01/2017 - 12:00 08/01/2017)

which determines how the bandwidth is shared amongst all the users: the average downlink bandwidth throughput per user is inversely proportional to the number of users in a cell. This can be used by network planners to determine overall number of subscribers that can be supported in a given cell and the desired average downlink throughput per subscriber. [70] explains how this is applied in practical LTE network design and deployment. In general, and focusing on 4G, the number of subscribers g supported in a cell site [70] is given by:

$$g = \frac{cf}{((1+d)e)} \tag{2.2}$$

where:

- a = downlink cell average capacity (assume 33Mbps/sector for LTE)
- b = designed downlink cell loading (typically 70%)
- c = a x b = designed downlink cell capacity in Mbps per sector
- d = peak to average ratio (typically 20%)
- e = average downlink bandwidth throughput per subscriber
- f = number of sectors in the cell site

Besides congestion, the other factors are equally important in understanding variation in network performance. For radio technology in use, logically, as mobile technology evolves from 3G to 4G to 5G, there ought to be an initial boost to network performance[13]. Given the reality of mobility of the user, coverage is a fundamental determinant of network availability and performance. Also, in moving from one cell site to the other, the effectiveness of handovers can deprecate the performance experienced by the user.

2.4.8 Improving network performance for SCS

Given its variability, there are technical, commercial and policy proposals to improve network performance. [71] noted that the reliability of the mobile network is much lower than imagined but that this can be improved significantly if devices can multi-home (i.e network sharing). [72] explores how to achieve end-to-end reliability of mission-critical traffic in 5G networks. [73] propose a 5G network slice for CCs. [74] suggest to place content in Multi-access Edge Computing (MEC) for the Internet of Vehicles. [75] calls for a regulatory mandate to create 999-style (or 112/911) data lanes for SCS traffic. Yet despite all these efforts, the assumption in this work is that mature 5G networks will similarly struggle with network performance from congestion because of the limits of the electromagnetic spectrum and because of the cost of densifying the network [76].

Regarding congestion, most of the classical solutions to dealing with congestion focus on improving speed and expanding the capacity of the network to cost-effectively reduce congestion [66]. In this work, I question whether these are still the only answers to delivering improved reliability and providing guarantees for SCS. Some literature expand on this point. On speed, [77] argues that "speed isn't everything", and faster speeds may no longer bring a better experience for end users as before. For [78], latency, as measured by the Round Trip Time (RTT), is of more profound importance; reducing RTT from, say, 150ms to 100ms, would improve page load times more than increasing a user's throughput from say 4Mbps to 1Gbps. [79] notes that not all customers want to, or can afford to pay for, faster broadband speeds. On capacity, its role in network performance is less certain given that the re-emergence of unlimited mobile data bundles in several developed markets (e.g. Austria, Denmark, Finland etc.) suggests that there is already surplus 4G capacity.

2.4.9 Measuring network performance

There are several papers in the literature that showcase insights from network measurement studies. For 3G, [80], the first paper in the literature to investigate mobile network behaviour during crowded events, focused on traffic behaviour and network performance for a tier-1 US network during two high-profile crowded events in 2012. Also [17] showed that there is significant temporal and spatial variations in throughput in different parts of a 3G data network, albeit with predictable aggregate behavior. For 4G, and focusing on usage in atypical scenarios, [18] studied performance during Superbowl 2013 in the US, finding causes and triggers for performance unpredictability. For 5G, [57] measured performance at the five busiest train stations in the UK in 2020 and noted that early 5G networks did not outperform 4G at the same locations.

One approach to getting network performance data is to get it from the network providers. Both [18] and [80] analyse the traffic dynamics from the perspective of the network operator, using data provided by the network operator. While such *a posteriori* data offers huge insights into the nature of network performance, it is often not suitable for end-users to understand what is going on in real-time. Crucially too, the data is not independent enough to be used to determine if the promise of reliability guarantees or SLAs between network operators and users are been met. Besides, given competition concerns, it is near-impossible for different operators to agree to contribute network performance data for a cross-operator comparison, unless it is demanded by a regulator.

An alternative to using *a posteriori* data is to use a proprietary measurement tool (e.g. a smartphone app) or use any of the crowdsourced measurement apps. [57] is an example of measurements done with an own app. Similarly, with the proliferation of smartphones and measurement apps, network measurement can be crowdsourced using tools/apps such as Netrader[68], MopEye[81], MobiPerf[82], Ookla, OpenSignal and RilAnalyzer[83] etc. These provide high-level insights on network performance across geographies and circumstances. [84] provides a comprehensive survey of mobile measurement testbeds, tools and services while [85] highlights the limitations of using apps for network measurement.

2.5 Differentiated Connectivity

2.5.1 Rationale and options

If network performance, even with 5G, is proven to be inadequate for SCS, then the telecoms industry needs a proportionate response to remedy the situation. One tried and tested approach is to treat selected users or traffic differently. This is neither new nor unique. Since the 1970s, fixed and mobile telecoms operators have sought ways to offer differentiated performance to different customers. The logic is a classical economics approach - if channel capacity is a finite resource, then a zero sum game approach means that each user or service gets the differentiated performance it is prepared to pay for, when they need it[86]. Given the limited capacity of the available electromagnetic spectrum, mobile data networks have an even stronger need to do so for commercial, operational or regulatory reasons. Commercially, the idea is to demand a premium price for the users enjoying a superior performance. Operationally, it is an option when speed and capacity improvements prove unfeasible or non-viable to solve congestion problems or to offer optimal performance to selected users. Regulatorily, differentiated performance underpins public safety services such as 999 or connectivity to Emergency Services Personnel or First Responders.

In practice, selective treatment for operational and regulatory purposes are generally accepted, but regulators and campaigners who are committed to maintaining the principle of Net Neutrality discourage selective treatment of users or traffic for commercial reasons. A core theme of our thesis is that *connectivity for SCS needs a differentiated performance because they straddle across commercial, operational and regulatory reasons.*

There are generally two approaches to offering differentiated performance to different users or services: Differentiated QoS (D-QoS) mechanisms and Smart Data Pricing (SDP).

- **Differentiated Quality of Service (D-QoS) Mechanisms:** A range of technical standards and implementations that propose to treat different users/packets differently or to create special lanes for different users/packets. D-QoS began as an effort to provide telephone-like guarantees over a best effort data network [87]. The motivation for this approach is that it will provide a guaranteed connection over the best effort internet link and that customers will be prepared to pay a premium for it. Examples of D-QoS include Asynchronous Transfer Mode (ATM), IP Type of Service (IP ToS) and QoS Class Identifier (QCI). Some of these D-QoS standards are targeted at the mass market, while some are utilised in limited scenarios (e.g. for enterprise customers).

- **Smart Data Pricing (SDP):** Several proposals to use differentiated pricing, both statically and dynamically, to maximise utilisation and revenue in data networks. SDP has been investigated extensively from 2010 [88] and several trials were performed [89]. Its primary proposal is to enable Internet Service Providers (ISPs) to price different traffic classes differently based on time, type of traffic/service, delay tolerance, QoS, etc. For example, Zero rating - the practice of giving customers free access to selected Internet content - is fairly established as a commercial prioritisation proposition. Yet, there has been little adoption of SDP [90].

While there are numerous research papers on the technicalities of most D-QoS standards, at at 2022, there is little evidence from academic and industrial research on their commercial performance. [25] is the first attempt in the literature to do so. In contrast, [91] provides a comprehensive review of the technicalities of implementation, and the commercial performance of SDP. This work focuses only on the use of D-QoS.

2.5.2 Net Neutrality

Net neutrality, the principle that all internet traffic should be treated equally, looms large in any discussion on how to provide reliability assurance to SCS. Net neutrality is a philosophical, political, economic and technical issue. Philosophically, it challenges the original design intention of the internet as a libertarian platform where all are treated equally. Politically, it raises the unpalatable political risk that the poor will receive an inferior experience. Economically, it is the antithesis to efforts to apply economics principles to managing demand in mobile networks [92]. Its critics generally argue that it has stifled innovation and investment in the industry[2]. Some technical commentators argue further that the principle stems from a fundamental misreading of the behavior of stochastic networks [93] and propose how to align the principle with the architecture of mobile networks [94]. However, to its advocates (including most regulators, policymakers and the weight of public opinion), equal treatment of all traffic is the best way to foster innovation [95].

The impact of such perceptions is partly responsible for why Smart Data Pricing (SDP) policies have failed to take off [90]. SDP has been investigated widely in the past decade and has spawned an extensive body of work on how to use pricing to manage demand and supply in internet networks [91] and for managing congestion [88]. Even when some types of SDP – e.g. Sponsored Data [96] – have been introduced in the market, operational and logistical challenges have often also prevented them from gaining traction or scaling up. Likewise, it is unlikely that attempts, such as in [92], arguing to impose similar net neutrality constraints on Content Delivery Networks (CDNs) and Search Engines, are helpful to mobile operators, nor will be endorsed by policymakers.

Essentially though, I note that net neutrality is mostly a concern when using prioritisation for selected users or user traffic for commercial reasons and is less of an issue with regards to the use of redundant multi-connectivity options or use of private infrastructure or using prioritisation for public safety. This understanding supports our thesis about using a broad group of officially-approved D-QoS approaches to provide reliability assurances to SCS.

1. Current status of Net Neutrality

Much of the policy debate that could impact D-QoS in the 5G era has focused on the possibility that mobile operators could violate Net Neutrality by giving preference to their own services. However, by tightening the focus to mostly SCS, this work seeks to defang the debate and reposition it to a discussion on how to improve society rather than on the commercial gain of operators. Here, I summarise the Net Neutrality position in the EU and US, to highlight how they make provisions that support the use of D-QoS for SCS. These two jurisdictions have shaped global policy on net neutrality most.

- **European Union:** Net neutrality in the EU is guided by the basic framework laid down by Article 3 of EU Regulation 2015/2120 on Open Internet Regulation[97]. The Body of European Regulators for Electronic Communications (BEREC) guidelines permits operators to prioritise specialised services if it is necessary to guarantee a specific level of quality [98]. This is known as the 'necessity requirement' and examples of services mentioned include VoLTE, live broadcasting IPTV and real time health services (e.g. remote surgery). A second rule, the 'capacity requirement' mandates that network capacity must be sufficient such that internet experience for other services is not degraded. Both requirements support using D-QoS for SCS as has been clarified by BEREC following a public consultation in April 2018 which we contributed to[99].
- **USA:** At the core of US regulation is whether, under the Telecommunications Act 1996, telecoms operators should be classified as Title I 'information services' or Title II 'common carrier services'. Under the Act, net neutrality will apply to Title II but not Title I services. Recently, this classification has been evolving in sync with the political cycle. Originally, the Federal Communications Commission (FCC) classified telecoms operators as Title II under the 'Open Internet Order 2015'. But this was suspended in June 2018 [9] during the Republican-led administration. However, with the return of the Democrats to power, the government has sought to restore the pre-2018 position again. The previous Order brought Broadband Internet Access Service (BIAS) under Title II

[9] https://www.fcc.gov/restoring-internet-freedom

but then created the non-BIAS services class. Some examples of non-BIAS services include IPTV service, VoIP, e-readers, educational services, and expectedly, SCS.

2. Lessons from other utility-like providers

The mobile industry can learn from other utility-like industries that have historically used, or have recently standardised the use of non-neutral schemes for the greater good of society, to balance demand and supply, or for commercial gain. An example of the first group can be seen in how emergency vehicles are given priority on roads under the assumption that they are on a duty tour that is of importance to society. Also travel updates on FM radio using Traffic Message Channel in Europe[100] where traffic messages are allowed to override current radio broadcast because society considers it important for drivers. An example of the second group is the electricity industry, which provides insights on how peak and off-peak tariffs are used to *balance demand* on the network. This also applies to the growing interest in the development of the Smart Grid for electricity to integrate data and information flow between suppliers and smart user devices to improve the efficiency of the electricity grid.

An example of the third group is the postal service which has been allowed to prioritise posts based on differential prices using so-called First-class and second-class stamps whilst being held liable for very stringent performance. Under the UK Postal Services Act 2011, and Royal Mail's Quarterly Quality of Service and Complaints Reports, the regulator, Ofcom, can fine Royal Mail up to 10% of its annual revenues if it fails to meet a recognized benchmark for 'best effort' – e.g. 98.8% of post using second class stamps should be delivered within three days. A non-neutral commercial service can also pose moral questions: for example, giving preferential treatment for higher-paying rail or airline passengers while everyone around suffers. Table 2.3 provides several other examples of prioritisation in other industries.

3. Adapting to Net Neutrality

I envisage that a zero-sum game assumption about a scarce resource would continue to influence perceptions about network neutrality. Hence, this work argues that a pragmatic and selective violation of net neutrality for SCS will cost-effectively improve reliability assurance only when warranted. I acknowledge that this view is controversial, especially given the negative reputation of mobile operators in society. However, there are steps that can be taken by mobile operators to allay most of the concerns of net neutrality advocates. These include:

- **User-determined and opt-out capability:** Through user studies, [101] demonstrate that users want some services to be given preference and proposed 'network cookies'

Table 2.3 How other utility-like industries treat users differently

Industry	Approach
Broadcasting	Overide of normal radio/TV broadcast with traffic update messages
Electricity	Smart Grid: information exchange between utilities and smart meters/devices to align demand and supply
Postal service	Different ticket classes for different service qualities
Rail	Different ticket classes for different days, times and service classes
Roads	High-occupancy toll and Express toll roads: allow selected drivers to use less busy lanes
Roads	Smart Motorway: average speed enforcement during peak hours, using the hard shoulder during peak hours or charging drivers per mile driven or for time of day
Roads	Expectation that all drivers make way for emergency services
Roads	Congestion charging: charges for entering a designated area or for pollution charges
Telecoms	Time of day pricing to move voice and data demand to nightime when there is ample unused capacity

as a means to express these preferences. This is already the default setting on most services for users to choose between mobile data and WiFi.

- **Automated negotiation of service classes:** Mobile operators may develop an API that would exchange information directly with applications on a user device. There is already a template for this for WiFi which enables apps to automatically respond when the user is connected to WiFi [102].

- **Increased network transparency:** Network operators can only succeed in violating net neutrality if they are very clear and transparent on what they are doing and why they are doing it. [103] supports this view, arguing that the current efforts to detect neutrality violations or enforce neutrality through technical means should be discarded in favour of increasing network transparency.

- **Focused at the mobile network edge:** Given that operators can currently deploy technical solutions at the network edge without much controversy, it is likely that selective violation of net neutrality would gain greater acceptance if it is limited at the network edge too. Operators already use mechanisms such as deep packet inspection, traffic shaping, handover to alternative cells/sectors, network type changes etc as a solution to the constraints faced in the radio environment. Selective violation of net neutrality can be added to this toolkit.

2.6 Redundancy Analysis for Connectivity

2.6.1 Redundancy for reliability & resilience

Generally, redundancy is the tried-and-tested means of improving reliability in biological and man-made engineered systems. Within engineering, the use of redundancy to assure reliability and resilience is an established approach and is applicable across all layers of the stack in the designed system[104]. [105] argues that “redundancy is the single most important engineering tool for designing, implementing and proving reliability in all complex, safety-critical technologies".

Best practice in design is to ensure that the redundant units are independent of each other[106], providing diversity or distributed redundancy akin to the ‘degeneracy’ seen in biological systems[107]. Using the example of the crash of the European Space Agency (ESA) Arianne 5 rocket on 4 June 1996, [105] illustrates the importance of ‘design diversity’ to achieve independence. For most ultra-reliable computer infrastructure (e.g datacentres), redundant paths and multi-homing are now the default operational setup.

2.6.2 Redundancy for performance-constrained reliability

For mobile networks, it is necessary to determine the performance indicator for which the use of the redundant backup will be triggered. Mobile network users can make this decision based on any of the constituent attributes of system reliability or cost - e.g. use of multi-SIM devices [24] and Software defined wide area network (SD-WAN) solutions[108]. [71] explains that these reliability attributes include network reliability (e.g. failures, availability, radio conditions), data plane reliability (e.g. packet loss, loss runs, large events) and performance reliability (e.g. latency, HTTP throughput, SIP success rate).

Classic reliability theory focuses mostly on network reliability and uses ‘availability’ as the KPI of interest to ascertain if the system is functioning or not: i.e a system is functioning if, and only if (iff), a connectivity path exists between the input and output [109]. But a connectivity path can be available yet grossly unable to deliver the performance required - necessitating path dependent reliability analysis [110].

2.6.3 Redundancy setup for mobile services

The mobile industry offers operator-redundancy in selected scenarios using ‘National Roaming’ where customers can be connected through the network of a different provider if the contracting operator’s network is unavailable. This is used for services such as 999 calls.

While national roaming should theoretically improve resilience, it is unclear whether a switchover will happen for under-performance (but not outright failure) on one or several performance indicators (cf: train companies in the UK will provide a replacement bus service when the train line is closed but will not do so if trains are only running with a delay). In our work, we suggest that national roaming is a continuation of the 'Always Best Connected' concept[111] of the 3G era where mobile network operators expected to decide how users are connected to mobile or wi-fi. Yet, over the past 15 years, the trend has been unmistakable in the opposite direction with users (or user-controlled devices) making the decision.

Users managing redundancy for themselves through the use of multi-SIM devices is fairly common in developing countries as a mechanism for price and quality arbitrage[24]. In most developed countries, a multi-SIM approach has typically been regarded as unnecessary until Google launched Google Fi in April 2015 in the US. However, [112] note that Google Fi does not deliver effective multi-access switching. Instead, they propose iCellular as a client-side service to let commodity mobile devices customize their own mobile network access. In reality, despite initial market excitement, momentum for Google Fi has waned and the service, or similar, has not been launched outside the US because it offers little to relatively affluent customers in countries with decent mobile coverage. Our work on multi-connectivity for connected cars[113] focuses on a use case that has safety (and as such regulatory) consequences instead of only price comparison.

Operationally, the use of embedded SIMs (e-SIM)[114] and blockchain-based Accountable Just-in-time (AJIT) smart contracts[115] will make user-managed redundancy easier. But there will be new problems in the way data packets are handled. Studies such as [116] have already shown that TCP behaves poorly in high-mobility environments. They note that, because of larger RTT and RTT Variance, TCP throughput in high speed rail is 3x worse than static and 2x worse than driving scenarios. Swapping networks brings additional complication. [117] notes that multi-path TCP improves performance across mobile providers while [118] explored how different TCP variants improved performance over 4G in a high speed, multi provider scenario.

2.6.4 Global feasibility of redundant mobile connectivity

In practice, achieving independent parallel redundant mobile connections should be mostly feasible across the world because, apart from Djibouti, Eritrea & North Korea, all countries have at least two operational mobile infrastructure. However, most publicly confirmed pilots or commercial contracts between a mobile network operator and a provider of safety-critical system (e.g. public safety LTE for emergency services[20] such as AT&T's Firstnet in the US or Testra's LANES in Australia) have touted the exclusivity of the relationship, with at best supply-side managed fall back in exceptional circumstances. Such operator managed redundancy is akin to hospitals (e.g. under the UK's Health Technical Memorandum 06-01[119]) expecting their utility provider to also provide the legally-mandated standby power generator.

2.7 Differentiated Connectivity for Connected Cars

2.7.1 Setting the context

Connected Cars (CCs) and their associated vehicle-to-everything (V2X) use cases, are an example of an SCS use case that requires stringent reliability for safety and non-safety applications. As such, a CC is a good candidate for differentiated connectivity. CCs are growing in importance in society and use mobile connectivity for real-time navigational support with maps, in-car infotainment (incl. in-car WiFi hotspots), updating firmware over-the-air (FOTA), remote car diagnostics and support (e.g. by the car manufacturer), monitoring of driving habits (e.g. by insurance companies) and for emergency response support (e.g. eCall in Europe [120]). Cellular V2X (cV2X) use cases can be Vehicle-to-Vehicle (V2V), Vehicle-to-Pedestrian (V2P), Vehicle-to-Infrastructure (V2I) and Vehicle-to-Network (V2N).

In futuristic models of society, the vision is that all cars are connected, creating an ecosystem of cars, transport infrastructure and ancillary maintenance, operational and infotainment services[2] and a market for data that is generated by CCs. The 5G-PPP whitepaper [121] provides the joint vision of the automotive and telecommunication industries. This vision has informed the activities of the 5G Automotive Association (5GAA), bringing together automotive manufacturers and ICT providers to develop and support interoperable and reliable cellular V2X services. Based on developments since 3GPP Release 14 [122], the major new capabilities envisioned for CCs in the future are autonomous driving and inter-connectivity with other cars, pedestrians and road infrastructure. While these are most suited for Sidelink, current market deployments are still predominantly reliant on V2N connectivity for both safety and non-safety use cases.

2.7.2 Performance expectations for Connected Cars

To realise the vision of an integrated and reliable ecosystem around CCs, there are performance expectations for safety and non-safety V2X scenarios (Table 2.4). Many of the current use cases are for non-safety scenarios and will remain so in the future. These include high data rate infotainment, in-car WiFi hotspots, navigational map updates, remote diagnostics and support. [73] note, however, that 3GPP is focused on four safety-related V2X scenarios:

1. Safety: Focusing on the use of extended sensors to gather and exchange information among cars, road infrastructure and pedestrians. The aim is to improve safety and reduce the number of road traffic accidents.

Table 2.4 3GPP V2X categories and main KPIs (Adapted from [73])

Focus	V2X category	Comms type	Latency	Throughput	Reliability
Safety	Safety in driving	V2V, V2P, V2I	100 ms	Not a concern	Not yet explicit
Safety	Traffic efficiency (esp. Platooning)	V2V, V2P, V2I	100 ms	Not a concern	Not yet explicit
Safety	Tele-operated driving	V2N	20 ms	25 Mbps uplink for video & sensors data, 1 Mbps for downlink	99.999%
Safety	Advanced driving	V2V, V2I, V2N	1 ms	10 Mbps for uplink and downlink	Nearly 100%
Non-Safety	Car internet and infotainment	V2N	100 ms	0.5 Mbps for browsing, up to 15 Mbps for video	Not a concern
Non-Safety	Remote diagnostics and support	V2I, V2N	Not a concern	Not a concern	Not a concern

2. Traffic efficiency: Covering platooning to enable a selection of cars to dynamically form groups, positioning themselves very close to each other, and moving together at high speeds as one unit. This mimics the railways and the aim is to increase road network capacity.
3. Tele-operated driving: For scenarios where a remote driver or V2X Application Server can remotely take control of a car (e.g. for dangerous driving conditions).
4. Advanced driving: Covering both semi-autonomous and fully-autonomous driving, and is facilitated via the exchange of data between cars and road infrastructure. This is the *holy grail* of the future of driving.

2.7.3 State of market for Connected Cars

As the number of CCs in the field grows, it becomes relevant to evaluate if the visions and expectations for the CC ecosystem are being met. Continued growth in the number of CCs confirms that the vision is broadly supported. So we focus on two outcomes: scaling the CC ecosystem and achieving expected mobile network performance for CCs.

On scaling, the original vision for scaling CCs assumed that companies in the telecommunication and automotive industries will be the primary providers, developers and gate keepers for CC connectivity and solutions[121]. In practice, since 2014, Apple's CarPlay and Google's Android Auto have become the dominant solution for infotainment services on CCs[123]. Today, in most countries, Apple and Google have disintermediated car manufacturers to become the primary gatekeepers of non-safety services for CCs and the car owner/driver brings along their existing smartphone connectivity.

On performance, [124] provides an extensive examination of the state of the CC ecosystem, noting the need to collaboratively address some long held assumptions to ensure that mobile connectivity remains relevant over the typical 12 to 15 years design lifecycle of the car industry. They caution that any chosen mobile connectivity option "must deliver reliable, seamless, uninterrupted coverage in all countries and markets where the vehicles are sold and driven". This view is supported by [125] who noted that many stakeholders in the CC ecosystem work in isolation on hard-to-scale, island and exclusive relationship between two players.

2.7.4 Overcoming challenges to mobile connectivity

Given the challenges of *performance* above, it can be seen that the current setup of mobile connectivity for CCs is insufficient to deliver full benefits for the CC ecosystem (the challenge of scaling is mostly of commercial and environmental importance and so is out of scope of this PhD). For performance for safety cV2X scenarios, CCs today experience the same performance as other network users and understanding this is a well trodden research area.

Given this variability in network performance, several authors have proposed technical, commercial and policy remedies to assure reliability that can benefit CCs. [71] observed that the reliability of the mobile network is much lower than imagined but that this can be improved significantly if devices can multi-home (i.e network sharing). [72] explores how to achieve end-to-end reliability of mission-critical traffic in 5G networks. [73] propose a V2X 5G network slice. [74] propose to optimize content placement in Multi-access Edge Computing (MEC) for the Internet of Vehicles. [75] pushes for a regulatory mandate for a 999-style (or 112/911) prioritisation that can be used for CC control signals.

For a CC, geographical variation in connectivity performance is a major concern given that in the course of its journey, a CC will be handed over from one base station to the other. [126] observes that this operational reality leaves only a short time window for large data transfers. For systems using the default single-path TCP, disruptions in the single connectivity path is a major source of performance degradation and has informed the development of Multi-path TCP as a means of providing redundant multi connectivity paths [127].

2.7.5 Comparing V2X and smartphones

Although CCs and smartphones experience the same network conditions today, there are still differences in how they respond to those conditions. Using data from one million CCs in the USA (from a single manufacturer) and over one billion mobile radio connections, [126] suggests that CCs are different to smartphones in three ways:

1. Comparison with smartphones: Connectivity for CCs and smartphones exhibit similar weekly and diurnal patterns. However, as CCs spend less time connected (overall and to each base station), the window to deliver large volumes of data is small and hence require different management approaches (e.g. for FOTA updates).

2. Comparison to IoT devices: Connectivity for CCs and IoT devices similarly only use few carriers, connect only to a subset of the network, and spend only a short time on the network and per session.

3. Unique traits: CCs are unique in the way they connect to different cells on different days, and predictably, the connectivity pattern matches commute-time patterns. Accordingly, ISPs could use prediction models to efficiently deliver content, manage mobility and assure QoS during the frequent handovers under high speed.

2.7.6 Improving reliability of V2X

Given the different remedies that have been proposed to improve reliability, and taking cognizance of the peculiarities of CCs vs smartphone, we summarise five broad approaches to improving and assuring the reliability of cV2X connectivity. Apart from the first approach, the remaining four approaches offer a differentiated connectivity experience to CCs.

- Generic generational improvements: Going from 4G to 5G will provide a performance boost to connectivity to support many safety and non-safety V2N use cases[13].

- Sidelink: Introduced in 3GPP Release 16 to support V2V, V2P and V2I use cases such as maintaining safety while driving, interacting with roadside infrastructure, avoiding accidents with pedestrians and for platooning. It uses the dedicated 5.9 GHz spectrum for Intelligent Transport Systems (ITS), to allow CCs to communicate directly with one another or other nearly entities, without the need for the wider, mobile network[128].

- Prioritisation: This is recommended for use when routine network upgrades are insufficient for reliability/security. 4G QoS Class Identifier (QCI), 5G QoS Indicator (5QI) and 5G Network Slicing are the key mechanisms for delivering prioritisation for V2X use cases[73] and 3GPP has already prescribed QCI/5QI 75/79 for them.

- Private 4G/5G networks: For scenarios where there is no public mobile network (e.g. supporting Rio Tinto's Autonomous Haulage Systems in Australian mines[129]) or where customers prefer to own/manage the communications infrastructure (e.g as seen in the long term vision articulated by Highways England on the possibility of building its own 5G network for UK motorways[130]).

- Multi-operator connectivity: Given the inability of any single network provider to provide QoS assurance, some customers may prefer to embrace multi-operator connectivity[131], either by using multiple redundant SIM cards, using an MVNO provider/Universal SIM card with multi-operator wholesale partners (e.g. Transatel or Twilio) or via roaming agreements (e.g. as several car makers are doing in Europe)[132]. This option relies on the engineering benefits of redundancy and is the focus of our work on evaluating D-QoS in Chapter 6.

2.8 Chapter Summary

This chapter has provided a background to the core theme and the key research questions of this work. It started with an exploration of SCS to understand their connectivity requirements and validate why they matter to society. Then, the focus shifted to the key research questions: firstly, an understanding of the nature of variations in mobile network performance; secondly, an exploration of the concept of D-QoS to understand if and when they can be used in case mobile network performance is adequate for SCS and also the role of redundancy in improving QoS assurance; lastly, the chapter introduced D-QoS for connected cars.

In the next chapter, I focus on the methodology for this research report. I will describe the research steps, the field measurements and the tools and benchmarks used for the measurements.

Chapter 3

Methodology

Ụka akpara akpa, isi ka eji ekwe ya (Igbo Proverb)

(Prior agreement is acknowledged by nodding of the head)

3.1 Introduction

Mobile connectivity has been a major driver of economic and social progress across societies. Since the introduction of data services in the 2G era, mobile communication has been enabling a plethora of SCS use cases, whether for IoT or for PWS. This means that the reliability of the networks, and the mechanism of how the reliability impacts SCS use cases, is generally understood. In this chapter, I introduce the methodology that will be used in this report to evaluate the research thesis and answer the research questions. The chapter starts with the overall research plan, describing the research steps and the field experiments that have been conducted. Then, the chapter describes the experiment setup and details, including how the measurements were setup, the software and hardware used, the parameters measured, the mobile network operators used and finally, the benchmark for the measurements.

3.1.1 Key questions answered

This chapter sets out the methodology and research plan used for the research in this report. It is based on using an experimental design/methodology to answer the generic research question "Can the application of X to Y produce an improved result (in the context of Z)?" In this context, the generic research question is:

> *In addition to standard technological improvements, can the application of Differentiated QoS (D-QoS) [X] to the mobile connectivity for Society Critical*

Systems (SCS) [Y] deliver improved reliability given the inadequacy of the current mobile networks to deliver QoS guarantees at all times? [Z].

3.1.2 Chapter layout

The rest of this chapter is subdivided into two sections. Section 3.2 introduces the research plan, highlighting the five research steps and the four field experiments that have informed this PhD report. Section 3.3 focuses on the experiement setup, detailing the measurement setup, the software and hardware used, the measurement parameters, the mobile network operators involved and the benchmark for the measurements.

3.1.3 Publications linked to chapter

This chapter has been informed by all the papers listed in Section 1.7.

3.2 Research Plan

3.2.1 Research Steps

There are five main steps towards answering the research questions in this PhD report. These, covering quantitative field experiments and qualitative desk research, are:

1. **Step 1**: Set out the hypothesis on why the adoption of D-QoS for the connectivity for SCS is a necessary approach for a proper functioning society in the digital age. Step 1 is covered in Chapter 1 and Chapter 2.

2. **Step 2**: Validate the need for D-QoS for SCS by investigating the reliability of live 3G/4G/5G networks via field research, to understand currently-achievable performance of live mobile networks in both typical and atypical scenarios. Step 2 is covered in Chapter 4.

3. **Step 3**: Review the history of D-QoS in the field to understand how similar rationale has driven earlier efforts in the past 50 years to use D-QoS to shape data services in the communications industry. Step 3 is covered in Chapter 5.

4. **Step 4**: Evaluate the performance improvement from using D-QoS for a live and real-time SCS use case to demonstrate the practicability of using D-QoS. Step 4 is covered in Chapter 6.

5. **Step 5**: Conceptualize a framework for action to guide industry stakeholders on how to proceed in adopting D-QoS for SCS use cases. Step 5 is covered in Chapter 7.

3.2.2 Field Experiments

There are four sets of field experiments in this research report. The first three experiments are used in Chapter 4 to answer the question on the adequacy of public mobile networks to provide QoS guarantees for SCS. The fourth experiment is used in Chapter 6 to evaluate the practical implementation of D-QoS for an SCS use case.

1. **Experiment 1:** Focusing on 3G and 4G networks in London UK from 2016/17 to 2019/20. The aim is to understand the performance of pre-5G networks during the New Year eve fireworks in London (atypical day) and on another day two weeks later (typical day).

2. **Experiment 2:** Focusing on 4G and 5G networks at the five busiest train stations in the UK in January 2020. The aim is to understand the performance of early 5G networks in high footfall locations.

3. **Experiment 3:** Focusing on 4G and 5G networks at four different locations in UK and South Korea respectively in 2021. The aim is to understand the performance of 5G networks in an internationally comparable manner.

4. **Experiment 4:** Focusing on 4G networks along 800 kilometers of major and minor roads in South East England in 2020. The aim is to understand whether a D-QoS approach will improve connectivity performance for a connected car.

3.3 Experiment Setup and Details

3.3.1 Measurement Setup

A logical methodology to measuring mobile network performance on SCS would have been to analyze real traffic data from UK cellular operators to understand the impact of network performance on access to different services.

However, this was not possible because of inability to gain access to the necessary data from UK mobile operators. As a result, the only option for us was to collect our own data. To start with, we valuated several publicly-available measurement apps: Netrader [68], [81], [82] and [83] but decided against using any of them because of their inability to provide us with the granular results nor the ability for us to tweak them to suit our research plan.

Instead, we created our own simplified “Netperf-KCL” app and this was used for all our field experiments to capture details for selected websites. Netperf-KCL was installed on identical Android devices which were connected to multiple mobile operators and the setup is as shown in Figure 3.1. Additional, and different, measurements were done for Field Experiment 4 and these are described in Chapter 6.

3.3.2 Measurement Software

Netperf-KCL was the primary measurement software for all our experiments and is based on an adaptation of Multiping-for-Android app[133]. It enables us to conduct 24 different measurements for 14 websites every 5 minutes. These include RTT on both TCP port 80 and TCP echo port 7, PLT, packet loss, RTT Variance, uplink speed, downlink speed, network type (e.g. LTE, HSPA+ etc), base station ID, base station location, device CPU parameters,

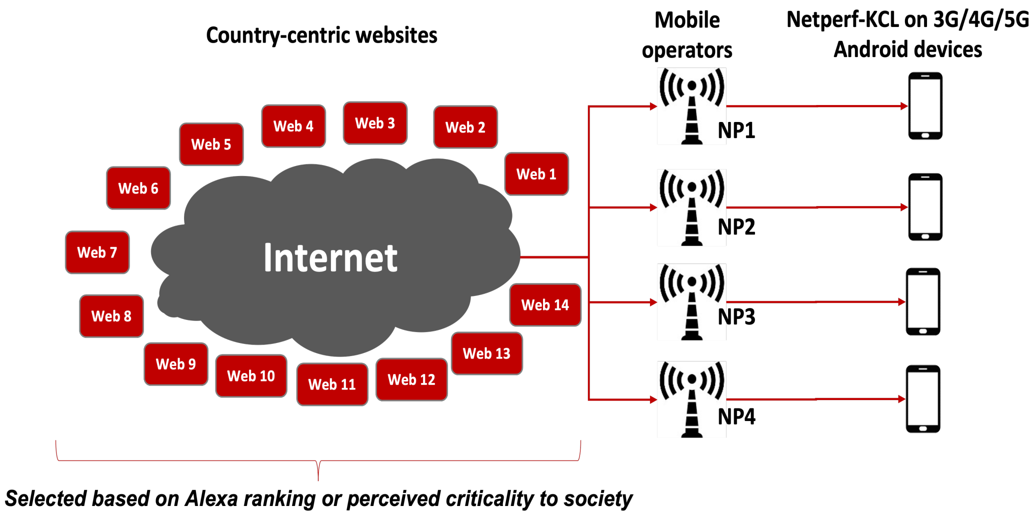


Fig. 3.1 Setup for field measurements on the UK & South Korean mobile networks

device RAM usage. The choice of websites was based on their Alexa Ranking as at December 2016 and on perceived importance to the UK digital society.

The 14 websites were selected based on whether they represent a service that is critical for society (e.g. www.police.uk) or based on their Alexa ranking at end 2016 (e.g. www.youtube.com) - see Table 3.1. These were originally selected for the UK in December 2016 and used for Experiments 1/2/3. The equivalent websites in South Korea were selected in January 2021 for the Korean part of Experiment 3.

3.3.3 Measurement Hardware

For 3G, we used Samsung SIII devices (released in May 2012). The CPU is the Quad-core 1.4 GHz Cortex-A9, running on the Exynos 4412 Quad (32 nm) chipset. The devices can run up to Android version 4.3 (Jelly Bean) and support Android SDK up to API level 16.

For 4G, we used Xiaomi Mi 4i devices (released in April 2015). The CPU of the devices is the Octa-core (4x1.7 GHz Cortex-A53 & 4x1.0 GHz Cortex-A53), running on the Qualcomm MSM8939 Snapdragon 615 (28 nm) chipset. The devices run Android version 5.0.2 (Lollipop) and support Android SDK of up to API level 23.

For 5G, we used the Samsung Galaxy A42 5G (released in November 2020). The CPU is the Octa-core (2x2.2 GHz Kryo 570 & 6x1.8 GHz Kryo 570), running on the Qualcomm SM7225 Snapdragon 750 5G (8 nm) chipset. The device runs Android version 10.0, supports Android SDK version of up to API level 29 and has a 5G-enabled SIM card.

Table 3.1 List of websites for experiment and rationale for their selection

	Website (UK)	Alexa ranking (Dec 2016)	Reason for including	South Korean Equivalent
1	gov.uk	28	The portal for most of UK government services	gov.kr
2	nhs.uk	148	Access to the NHS is already prioritised through the 999 and 111 emergency services	nhis.or.kr
3	nationalrail.co.uk	-	Provides information for a critical national infrastructure in the UK which is already prioritised over radio	letskorail.com
4	tfl.gov.uk	128	Provides information for all transport services in London which is already prioritised over TV and radio	seoulmetro.co.kr
5	police.uk	-	Access to the police is already prioritised through the 999 emergency service	police.go.kr
6	heathrow.co.uk	-	Provides information on arrivals & departures of a critical national infrastructure in the UK	airport.kr
7	highways.gov.uk	-	Provides information on the current status of roads across the UK and this information is already prioritised for broadcast on radio	roadplus.co.kr
8	bbc.co.uk	7	The highest ranked news website in the UK and the country's national broadcaster	kbs.co.kr
9	dailymail.co.uk	18	A major UK newspaper that is widely acclaimed to have successfully transitioned to a digital business	yna.co.kr
10	google.co.uk	1	Based on its Alexa ranking	naver.com
11	facebook.com	4	Based on its Alexa ranking and the popularity of social media	facebook.com
12	youtube.com	2	Based on its Alexa ranking and the popularity of online video	youtube.com
13	amazon.co.uk	5	Based on its Alexa ranking and the growing importance of online shopping in the UK	coupang.com
14	dropbox.co.uk	51	One of the most popular cloud storage websites	drive.google.com

3.3.4 Measurement parameters

While the design of Netperf-KCL provides insights on 24 parameters, we use only 11 measurement parameters in our analysis. The 11 measurements we conducted using Netperf-KCL include RTT on both TCP port 80 and TCP echo port 7, Page Load Times (PLT), packet loss, RTT Variance, uplink speed, downlink speed, network type (e.g. LTE, HSPA+ etc), base station ID, base station location and video playback analysis. The 'what', 'why', 'when' and 'how' of each of the parameters is summarised in Table 3.2.

Our primary measurement parameter was RTT - in keeping with [134]'s view that RTT can be used to understand how websites respond to customers in varied scenarios.

3.3.5 Mobile Network Operators

Throughout our work, we describe the four UK mobile network operators as NP 1, NP 2, NP 3 and NP 4 and the three South Korean operators as SNP1, SNP2 and SNP3. Whenever we used an MVNO in the UK, we describe it as NP 5. This approach gives anonymity to the network operators we are working with and allows us to focus our work on evaluating the right requirements for SCS and not on the political/legal minefield of comparing UK mobile network performance against each other.

Table 3.2 Measurement parameters: what, why, when and how

	What	**Why**	**When**	**How**
1	Round Trip Times (RTT)	To measure end-to-end latency from our devices to the webservers of the 14 websites	From 2016/17	Use InetAddress class to connect to TCP port 80 and TCP Echo Port 7 of the websites, using Java sockets, every 5 minutes. The shorter of the two delays was selected. Connecting to both ports ensures reach even for websites that may block ping on port 80.
2	Base Station CellID	To understand how many base stations or sectors that a device connects to, even when stationary	From 2016/17	Using the cell ID function in Android Telephony Manager.
3	Signal Strength	To understand variations in signal strength	From 2016/17	Using the signal strength function in Android Telephony Manager. Not used for further analysis because of lack of insight.
4	Network Type	To understand variations in the underlying mobile technology used	From 2016/17	Using the network type function in Android Telephony Manager.
5	Page Load Times	To understand how long it takes to load the contents of each service	From 2017/18	Calculated from how long it took Android webview to download a webpage.
6	Uplink speed	To understand the realizable uplink speeds	From 2019/20	Using the Android NetworkCapabilities API - getLinkUpstreamBandwidthKbps() method which return the speed of the connection in kbps and these are converted to Megabytes per second (Mbps).
7	Downlink speed	To understand the realizable downlink speeds	From 2019/20	Using the Android NetworkCapabilities API - getLinkDownstreamBandwidthKbps() method which return the speed of the connection in kbps and these are converted to Mbps.

8	Congestion Flag	To understand how the Android system assessed congestion	From 2019/20	Using the congestion flag function in Android Telephony Manager. Not used fo r further analysis because of lack of insight.
9	Packet Loss	To understand how many packets sent	From 2019/20	By observing packet losses recorded during the ping measurements.
10	RTT Variance	To understand the variation in the latency	From 2019/20	By measuring RTT five times and taking the standard deviation.
11	Video playback	To evaluate the impact of connection quality on video playback	From 2019/20	Download and playback a 30 seconds Youtube video, every five minutes.

3.3.6 Benchmarks for analysis

Table 3.3 summarises the benchmarks for evaluating the results from the measurements. These are largely informed by two factors: what is already obtainable on operational 4G networks globally (e.g based on Opensignal reports[135]) and the connectivity expectations for 5G use cases[2]. Specifically for RTT, we are using 50ms as the 5G benchmark even though that ought to be the ‘designed target for mature 4G networks’ (cf. in 2008, [136]). Also, in using only Korea or UK focused websites, we have minimised the geographical delay - longest UK to UK distance is 1,398 kilometers which can only add <5ms delay given that a distance of 300 kilometers adds 1ms delay.

Table 3.3 Performance thresholds for the parameters in experiments

Metric	4G	5G
Round Trip Times (RTT)	100 ms	50 ms
RTT Variance	20 ms	10 ms
Page Load Times (PLT)	1000 ms	500 ms
Uplink speed	25 Mbps	50 Mbps
Downlink speed	50 Mbps	100 Mbps
Packet Loss	0%	0%
Availability	RTT > 3,000 ms	RTT > 3,000 ms

3.4 Chapter Summary

This chapter has described the methodology used for this research report. It started with an enumeration of the research steps and the field experiments to provide a summary of the different activities used. Then the focus shifted to the measurement setup, the hardware and software (Netperf-KCL) used, the measurement parameters that inform the analysis in the report and the benchmarks for the parameters.

In the next chapter, I focus on Step 2 of the research plan to address **Research Question 1 on the adequacy of the mobile network to provide QoS guarantees to SCS**. I will report on our measurement studies on mobile network performances covering 3G, 4G and 5G; in different locations in the UK and South Korea; and for typical and atypical scenarios.

Chapter 4

Understanding Mobile Network Performance

Ogologo abughị na nwa m etola (Igbo Proverb)

(Being tall does not make a child an adult)

4.1 Introduction

Mobile connectivity has been getting better since it became a mass market service in the 1980s. From 1G to 2G, 3G to 4G, and now 5G, the benefits of a steadily improving mobile connectivity has been of immense value to society. The GSMA reports that, in 2021, the mobile industry contributed $4.5 trillion to the global economy, making up about 5% of global GDP[137]. It is therefore understandable that the mobile network is considered as a critical infrastructure in most countries[138]. While the generic generational improvements in mobile technology has unlocked more and more use cases, the impression is still that mobile connectivity is not yet sufficient for safety and life threatening use cases. Examples include the decision to create separate public safety 4G/5G networks[20] and the insistence of car manufacturers not to rely on mobile connectivity for autonomous vehicles[1].

4.1.1 Key questions answered

In this chapter, I set out to investigate how true it is that today's public mobile connectivity is inadequate for SCS. In doing do, I focus on three sets of measurements done on 30 days

[1] https://www.bbc.co.uk/news/business-45048264

over six years, on 3G/4G/5G networks, in the UK and South Korea, on typical and atypical scenarios. The insights from the measurements will provide answers to the research question:

Research Goal 1: *Investigate the adequacy of public mobile networks to provide QoS guarantees for society critical services*

> **RQ 1.1:** How reliable are 3G and 4G networks to deliver guaranteed QoS guarantees to end users?
>
> **RQ 1.2:** Will extra capacity and design improvements in public 5G networks be sufficient to provide consistent and enforceable QoS guarantees for SCS?

4.1.2 Chapter layout

The rest of this chapter is subdivided into three sections. Section 4.2 focuses on Experiment 1, looking to performance in pre-5G networks in the UK. Section 4.3 focuses on Experiment 2, exploring performance in an early 5G network in the UK. Section 4.4 focuses on Experiment 3, and will bring insights from measurements on 5G networks in UK and South Korea. Each of Sections 4.2 - 4.4 is presented in the 'scientific writing format' featuring the subsections: Aim; Methodology; Results; Discussion and Conclusion.

4.1.3 Publications linked to chapter

The paper presenting the details used in Section 4.2 was presented at the 45th IEEE International Conference on Telecommunications and Signal Processing (IEEE TPS) 2022. The results in Section 4.3 were presented at the IEEE Infocomm Global Internet Workshop 2021[57]. Finally, the results in Section 4.4 are contained in a paper submitted to ACM Internet Measurement Conference 2022.

4.2 Experiment 1: Performance in Pre-5G UK Networks

Pre-5G networks are pervasive globally. In most European countries, population coverage for 2G, 3G and 4G is already well above 90%. While 2G/3G coverage were originally the de facto means of providing connectivity, this has now been replaced by 4G. For example, in the UK, Ofcom reports that each of the four UK mobile operators provides 4G coverage to 99% of UK premises by 2021[139]. These 4G networks are very much likely to exist well into the 2030s, providing backfill as 5G networks are progressively rolled out[2]. It is therefore imperative to understand the nature of performance on 4G networks as they will provide a lot of the connectivity for SCS in the 5G era.

4.2.1 Aim

The primary aim of this measurement study is to provide an answer to **RQ 1.1** on whether 3G/4G networks can deliver QoS guarantees to end users. Breaking it down further, the measurements set out to understand three questions:

1. How does user-perceived network performance vary in typical and atypical scenarios? This work focuses heavily on RTT because it is a good proxy for congestion [134] and [66] has already observed that the RTT during crowded events degrades by 1.5 – 7 times compared to the average on typical days.

2. Do all services experience similar degradation in the event of poor network performance? If net neutrality is being upheld, service degradation will be democratic with no recognition of the criticality of some services.

3. To what extent is the cellular network responsible for the observed service degradation? Effort should be made in the design and operations of the measurement to ensure that causality can be attributed appropriately.

4.2.2 Methodology

The open question in RQ 1.1 provides the background to the field measurements we conducted in Central London, covering typical and atypical days over four years (2016/17 - 2019/20).

1. Measurement philosophy & design

Our core measurement philosophy is to understand the nature of mobile network performance, over multiple years, in typical and atypical scenarios and using a repeatable scientific approach. Accordingly, we controlled for the five major factors (as identified in [19]) that affect bandwidth availability for users - the ultimate determinant of network performance: radio technology, devices, coverage, handovers and congestion. On radio technology & devices, we always used the same devices running on the same radio technology. On coverage & handovers, we used a fixed location with verified network availability and no user-triggered handovers. On congestion, we chose the New Year eve fireworks celebrations as a repeatable annual event, where attendance, as reported in local media, was always over 100,000 spectators. Additionally, we took measurements at busy train stations in London.

2. Date, time & location

Our study design was originally for six years (2016/17, 2017/18, 2018/19, 2019/20, 2020/21 & 2021/22) but Covid-19 cancelled the final two years. The measurements were conducted over two 24 hour periods on each of the years. The first period, our atypical scenario, is the 24 hours between 12:00noon (31 December) to 12:00noon (1 January). The second period, our typical scenario, happens two weeks after over a similar 24 hour period on 14/15 January. The location is in the Engineering lab at King's College London Strand campus (KCL Strand) which is directly opposite to the location of the New Year fireworks in Central London. In 2019/20, we also took measurements, for the same time periods, at London Bridge - UK's fourth busiest train station - for additional comparison. In 2017, we also took measurements at London Cannon Street train station during a strike by London Underground workers. The dates, times and locations are summarised in Table 4.1.

Table 4.1 Exp 1: Dates and rationale for mobile network measurements in London

	Dates	Location	Rationale	Technology
1a	31/12/2016 (12:00) – 01/01/2017 (12:00)	KCL Strand	New Year eve: estimated 110,000 people watched the fireworks	3G
1b	07/01/2017 (12:00) – 08/01/2017 (12:00)	KCL Strand	To compare & contrast with New Year eve performance	3G
2a	09/01/2017 (10:00) – 10/01/2017 (19:00)	Cannon Street Station	Major transportation disruption because of London Underground workers' strike: 22.7 million entries and exits in 2016/17 for National Rail; 9.4 million for London Underground in 2017	3G
2b	16/01/2017 (10:00) – 17/01/2017 (19:00)	Cannon Street Station	Normal day with no transportation disruption	3G
3a	31/12/2017 (12:00) – 01/01/2018 (12:00)	KCL Strand	New Year eve: estimated 100,000 people watched the fireworks	4G
3b	14/01/2018 (12:00) – 15/01/2018 (12:00)	KCL Strand	To compare & contrast with New Year eve performance	4G
4a	31/12/2018 (12:00) – 01/01/2019 (12:00)	KCL Strand	New Year eve: estimated 100,000 people watched the fireworks	4G
4b	14/01/2019 (12:00) – 15/01/2019 (12:00)	KCL Strand	To compare & contrast with New Year eve performance	4G
5a	31/12/2019 (12:00) – 01/01/2020 (12:00)	KCL Strand & London Bridge Station	New Year eve: estimated 100,000 people watched the fireworks	4G
5b	14/01/2020 (12:00) – 15/01/2020 (12:00)	KCL Strand & London Bridge Station	To compare & contrast with New Year eve performance	4G

Human traffic at different measurement locations

1. *New Year fireworks crowd: 2017[2]; 2018[3]; 2019[4]; 2020[5]*
2. *Cannon Street Station passenger numbers: National Rail[6]; London Underground[7]*

[2] http://news.sky.com/story/crowds-gather-across-uk-to-bring-in-the-new-year-with-fireworks-10713895
[3] https://www.bbc.co.uk/news/uk-42528945
[4] https://www.express.co.uk/news/uk/1065594/London-nye-fireworks-london-eye-new-year-2019
[5] https://www.mirror.co.uk/news/uk-news/breaking-new-years-eve-2019-21192748
[6] http://orr.gov.uk/statistics/published-stats/station-usage-estimates
[7] https://data.london.gov.uk/dataset/london-underground-performance-reports?resource=b6ab04fc-9062-4291-b514-7fa218073b4c

3. Measurement devices

For 3G, we used five Samsung SIII devices (released in May 2012). The CPU is the Quad-core 1.4 GHz Cortex-A9, running on the Exynos 4412 Quad (32 nm) chipset. The devices can run up to Android version 4.3 (Jelly Bean) and support Android SDK up to API level 16. For 4G, we used six Xiaomi Mi 4i devices (released in April 2015). The CPU of the devices is the Octa-core (4x1.7 GHz Cortex-A53 & 4x1.0 GHz Cortex-A53), running on the Qualcomm MSM8939 Snapdragon 615 (28 nm) chipset. The devices run Android version 5.0.2 (Lollipop) and support Android SDK of up to API level 23.

4.2.3 Results

In this subsection, I summarise the results from 10 days of measurements, across five years, in Central London, focusing on the results that elucidate the reality of providing reliability/QoS assurance to SCS. Overall, we captured 544,560 readings across 14 categories.

For simplicity, this report uses ‘A’ to represent the Atypical scenario and ‘T’ for the Typical scenario. This is combined with the NP 1 - 5 naming style for each of the UK operators. Hence, ‘A_1’ is the atypical scenario results for operator NP 1 while ‘T_3’ is the typical scenario result for operator NP 3.

1. RTT: Overall performance

The first result is the overall picture of RTT measurements on either 3G or 4G device, on either atypical or typical days, on either of the four network operators, and on either 2017/2018/2019/2020. Figure 4.6 shows the results from an overall system view for 544,560 RTT readings. Assuming that the readings are bifurcated into two equal halves of atypical vs typical scenario, ***the headline is that only 58% of 4G measurements achieved a 50ms target (see Table6.2)***.

The choice of 50ms as the benchmark is pertinent because it is worse than the median RTT recorded in 2020 for NP1 - 4. For example, for NP1, the 24 hour atypical RTT median on New Year was 88ms in 2018, 30ms in 2019 and 31ms in 2020. For NP2, this was 537ms in 2018, 100ms in 2019 and 35ms in 2020. For NP3, it was 73ms in 2018, 36ms in 2019 and 37ms in 2020. For NP4, it was 122ms in 2018, 66ms in 2019 and 32ms in 2020. It is therefore evident that 4G network performance have improved consistently since 2018 - the readings for 2017 were on 3G.

Given that 50ms is a suitable expectation for latency on 4G networks today and many services (including SCS) will have been designed to work best at such performance, the reality is that only 58% of 4G measurements would have achieved such a target. If latency of 50ms is used as the proxy for overall network performance, then it means that any QoS assurance for SCS connectivity will only be achieved 58% of the time.

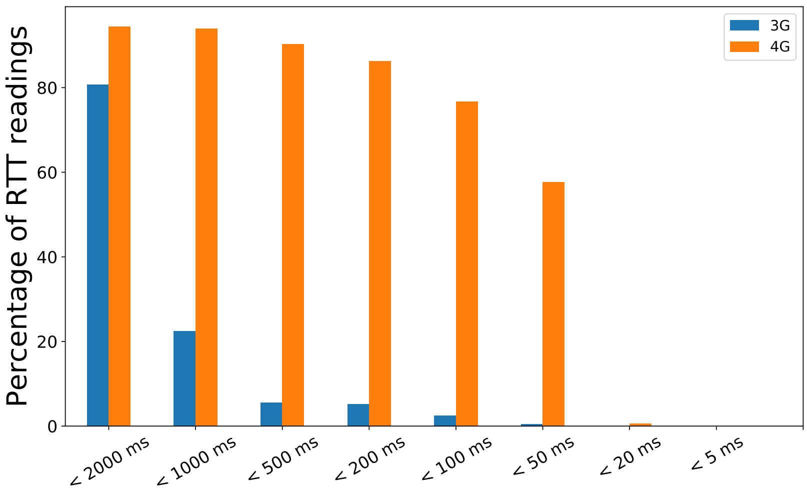
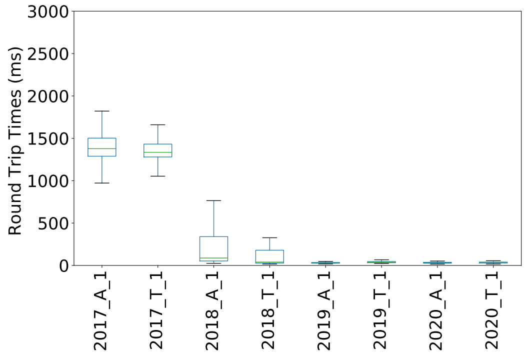


Fig. 4.1 Overall RTT readings for 3G and 4G devices. Only 58% of 4G device measurements achieved <50ms latency

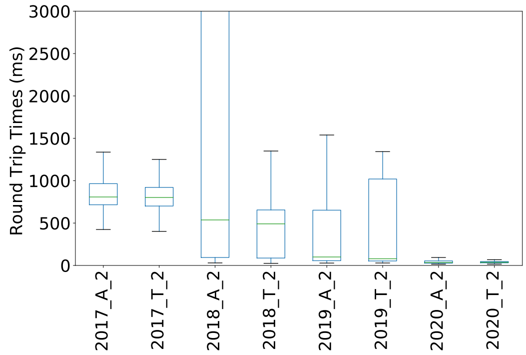
2. RTT: atypical vs typical days (NP 1 - NP 4 and 2017-2020)

Figures 4.2 and 4.3 show the RTT measurements for atypical and typical days for each operator over the four year period. In our measurements, ***the RTT recorded on atypical days is generally worse than the measurements on typical days***. This is reflected in the CDFs in Figures 4.4 and 4.5, where the typical day CDFs are mostly superior to atypical days across all latency levels. The annual results are for readings taken between 12:00 31/12 to 12:00 01/01 and also 12:00 14/01 to 12:00 15/01 and applies to all four operators. For example, in 2017, using 3G devices, for NP1 the median was 1378ms on atypical scenario and 1335ms for typical scenario; for NP2, this was 808ms vs 802ms; for NP3, it was 1580ms vs 1522ms and for NP4, it was 1243ms vs 1573ms (the anomaly). In 2018, using 4G devices, for NP1 the median was 88ms vs 40ms; for NP2, this was 537ms vs 491ms, for NP3, it was 73ms vs 64ms and for NP4, it was 122ms vs 94ms. Results for 2019 and 2020 are similar. Generally, on an annual basis, the average median RTT, over 24 hours on atypical vs typical day was 1.0x for 2017 on 3G, 1.2x for 2018 on 4G, 1.1x for 2019 on 4G and 0.9x for 2020 on 4G. Across the four years, the comparison for NP 1 on atypical day was 1.0x the RTT on typical day, 1.1x for NP 2, 1.0x for NP 3 and 1.1x for NP 4. Overall, there is a 1.1x difference between atypical vs typical scenarios.

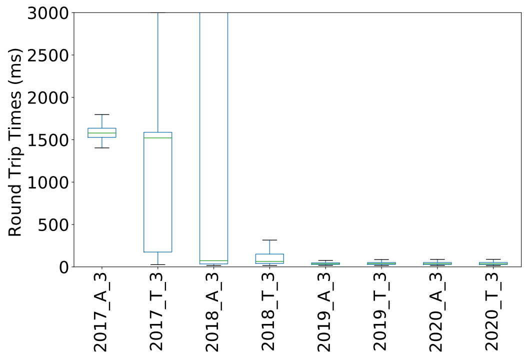
Our observation was expected given the volume of human traffic on the atypical days and the fact that telecoms networks are designed and dimensioned to cope with peak demand on a typical day. When atypical days and crowded events are to occur, predictability in user numbers and behaviours helps in network planning, including the use of ad-hoc, portable base stations - e.g. Cells on Wheels (COWS). However, in many situations where SCS is in use, such predictability may not be possible.



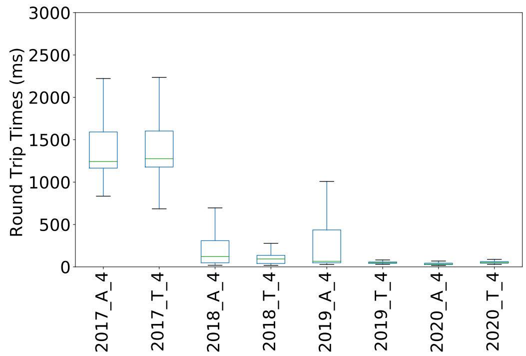
(a) NP1 RTT results



(b) NP2 RTT results



(c) NP3 RTT results



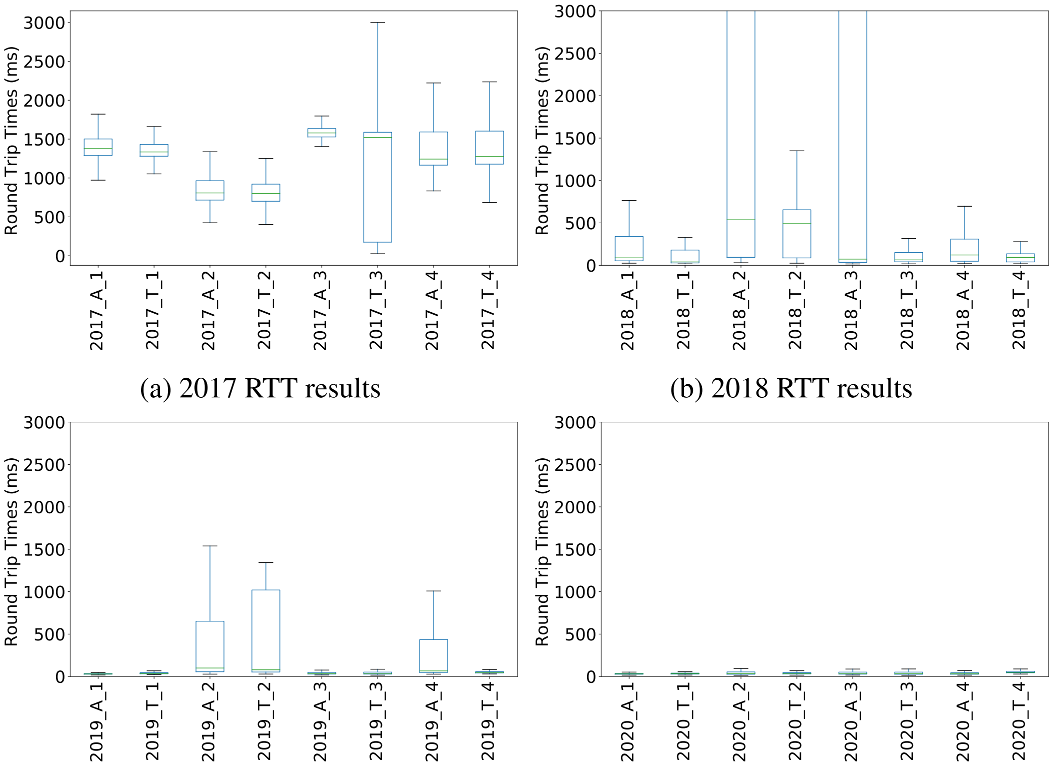
(d) NP4 RTT results

Fig. 4.2 RTT results per operator (NP 1- 4) showing how each operator's network performance changed between 2017 - 2020 on atypical (New Year; 12:00 31/12 to 12:00 01/01) and typical (12:00 14/01 to 12:00 15/01) scenarios. 2017 was done using 3G devices while rest were on 4G devices. Same y-axis scale used for consistency across all charts.

3. RTT: Performance during New Year fireworks

While we have already established that atypical scenarios usually have worse network performance, here, I drill down into the 6 hour segment (20:00 - 02:00) during which the crowd was at its peak. This is to establish, with a higher degree of certainty, how much increased number of users impacted network performance.

For example, in 2017, for NP1, this was 1416ms vs 1332ms; for NP2, it was 1116ms vs 789ms, for NP3, it was 1605ms vs 1549ms and for NP4, it was 1609ms vs 1519ms. In 2019, for NP1, it was 30ms vs 36ms (the anomaly); for NP2, this was 78ms vs 61ms; for



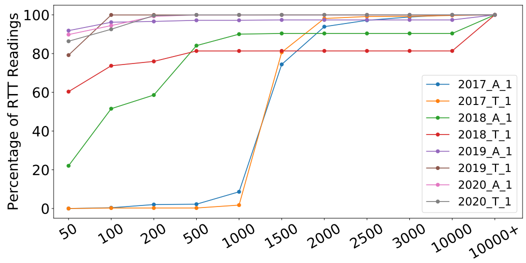
(a) 2017 RTT results

(b) 2018 RTT results

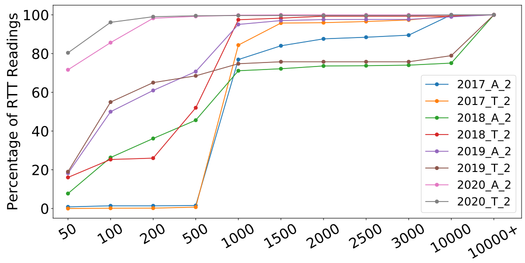
(c) 2019 RTT results

(d) 2020 RTT results

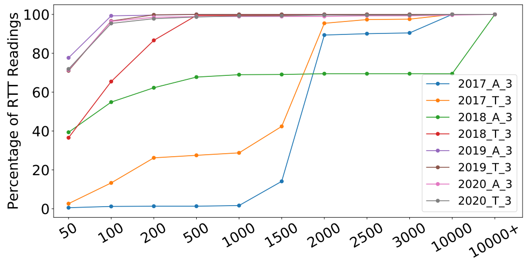
Fig. 4.3 Annual RTT results (2017 to 2020) showing how each of the Operators NP1 - NP4 performed compared to its competitors on atypical and typical scenarios. Same y-axis scale has been used to maintain consistency across all charts.



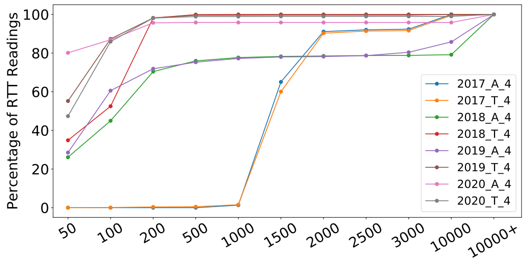
(a) NP1 Results



(b) NP2 Results

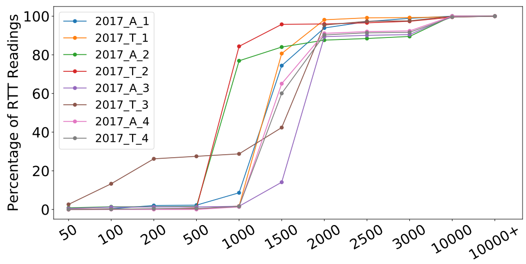


(c) NP3 Results

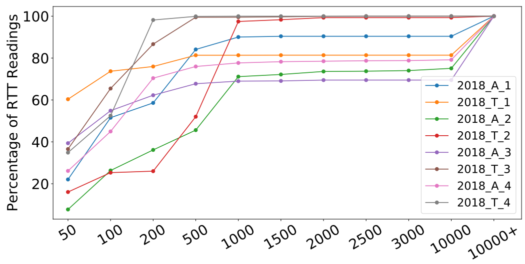


(d) NP4 Results

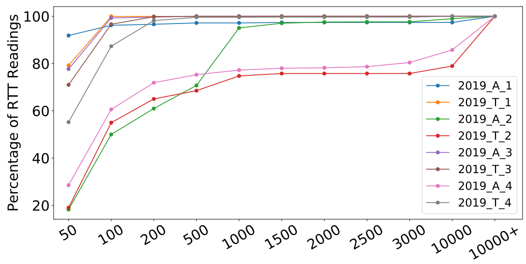
Fig. 4.4 RTT CDF for NP 1 - 4 for 24 hours, showing better performance on typical days over atypical days



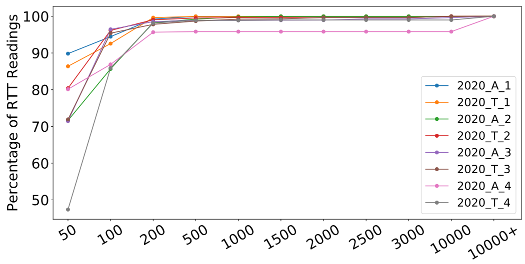
(a) 2017 Results



(b) 2018 Results



(c) 2019 Results



(d) 2020 Results

Fig. 4.5 RTT CDF for all Operators (2017 - 2020) showing significant variation in performance across operators

NP3, it was 38ms vs 31ms and for NP4, it was 77ms vs 46ms. Results for 2018 and 2020 are similar. Generally, on an annual basis, the average median RTT, over 24 hours on atypical vs typical day was 1.1x for 2017 on 3G, 4.2x for 2018 on 4G, 1.3x for 2019 on 4G and 1.0x for 2020 on 4G. Across the four years, the comparison for NP 1 on atypical day was 1.1x the RTT on typical day, 1.3x for NP 2, 1.2x for NP 3 and 1.1x for NP 4. Overall, ***there is a 1.2x difference between atypical vs typical scenarios.***

This suggests that, whether for 3G or 4G, there is a noticeable degradation in performance during and around the New Year fireworks celebrations, providing strong evidence that congestion from more network users impacted overall performance for all users.

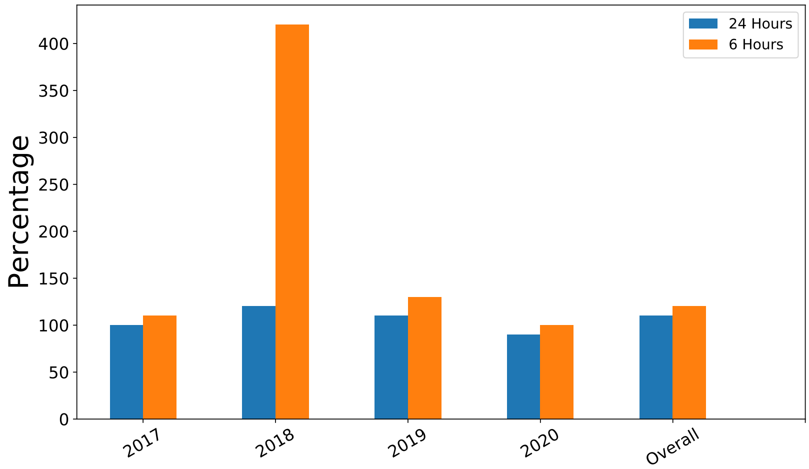


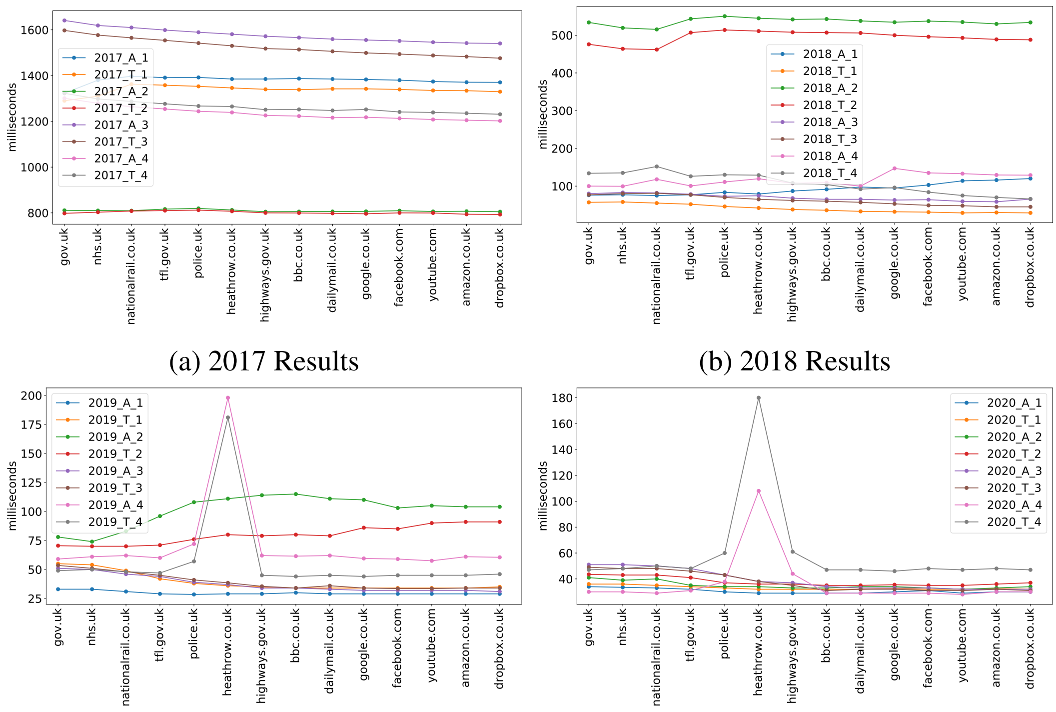
Fig. 4.6 Atypical vs Typical Scenario Median RTT over 24hrs (12:00 - 12:00) vs 6hrs (20:00 - 02:00). Chart shows that for both time segments, atypical performance was generally worse. However, the inferior performance was more noted during the 6hr slot at the peak of the New Year fireworks

4. RTT: Median performance per website

A key benchmark in the measurements was to understand how the different websites performed. This is akin to evaluating how different services, including SCS, will perform compared to other services that are running concurrently. This is important in order to control for server-side under-performance for any of the websites. Figure 4.7 summarises the results for the four years showing the median RTT for each website. Broadly, we observe that ***RTT to the 14 websites is broadly similar on each operators's network, confirming that performance or under-performance is democratic inline with net neutrality expectations***. For example, despite some visible outliers in 2019 and 2020 for www.heathrow.com, the

Median Absolute Deviation (MAD) was 4ms in 2017, 8ms in 2018, 1ms in 2019 and 1ms in 2020.

Given that the location, devices and websites (i.e server side) are common for each operator in Figures 4.7a-4.7c, it is further evidence on how it is challenging to entrust reliability/QoS assurance for SCS to a single operator's public network.



(a) 2017 Results (b) 2018 Results

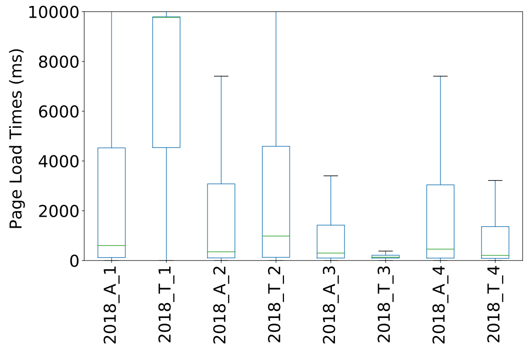
(c) 2019 Results (d) 2020 Results

Fig. 4.7 Median RTT for 14 UK websites (2017 - 2020). Figure shows that the experience of different websites is broadly democratic with a Median Absolute Deviation of 4ms in 2017, 8ms in 2018, 1ms in 2019 and 1ms in 2020

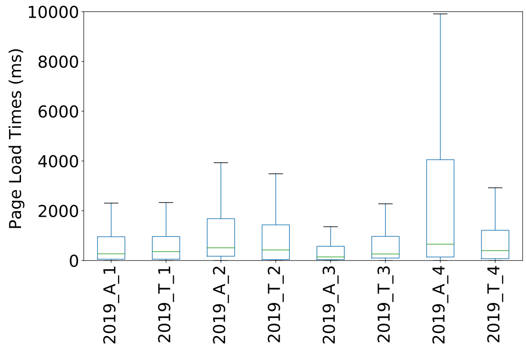
5. PLT: atypical vs typical days (2018 - 2020)

We set out to measure PLT in the expectation that it will inform us about actual application behavior. As with RTT, our starting assumption was that atypical days will have worse performance than typical days. However, unlike RTT where there was a clear improvement in performance on typical days vs atypical days, ***the observations on PLT were more mixed without any immediately discernible trend to support the assumption.*** Figure 4.8 illustrates this. In 2018, only NP3 and NP4 conformed to our expectation of a worse performance on atypical days. In 2019, this only happened for NP2 and NP4. In 2020, it happened to none of the operators. We also note that one of the websites performed significantly worse, in 2020 typical day, across all operators.

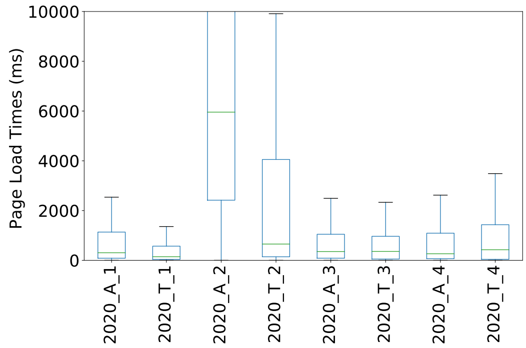
Based on this, we highlight a truism that regardless of any improvements or optimisations at the access network, end-to-end application-level SLAs will need to track source server performance in addition to network performance.



(a) 2018 Results



(b) 2019 Results



(c) 2020 Results

Fig. 4.8 Page Load Times (2018 - 2020). Results are mixed and there is no immediately discernible trend to support that atypical days are worse than typical days across all the years.

6. Base Station cell IDs used (2017 - 2020)

In an ideal world where there is adequate capacity at each access network, a stationary device should be able to connect to only a single base station to achieve its connectivity needs. As our measurements were in a fixed location, we anticipated that all readings for each operator will come from the same base station or base station sector. In order words, we expected the cell ID to be singular and unique. Figure 4.9 is the longitudinal count of base stations and sectors observed in our measurements. It shows that ***most of the networks used varying numbers of base stations or sectors on different days and with apparent randomness***. Counter-intuitively, more base stations were in use on typical days compared to atypical days, negating our expectation that network operators would use ad-hoc, portable base stations to cope with congestion on atypical days. On the share of readings from the top used base station, we notice that, for most operators across the four years, usually >50% of readings came from the top used base station.

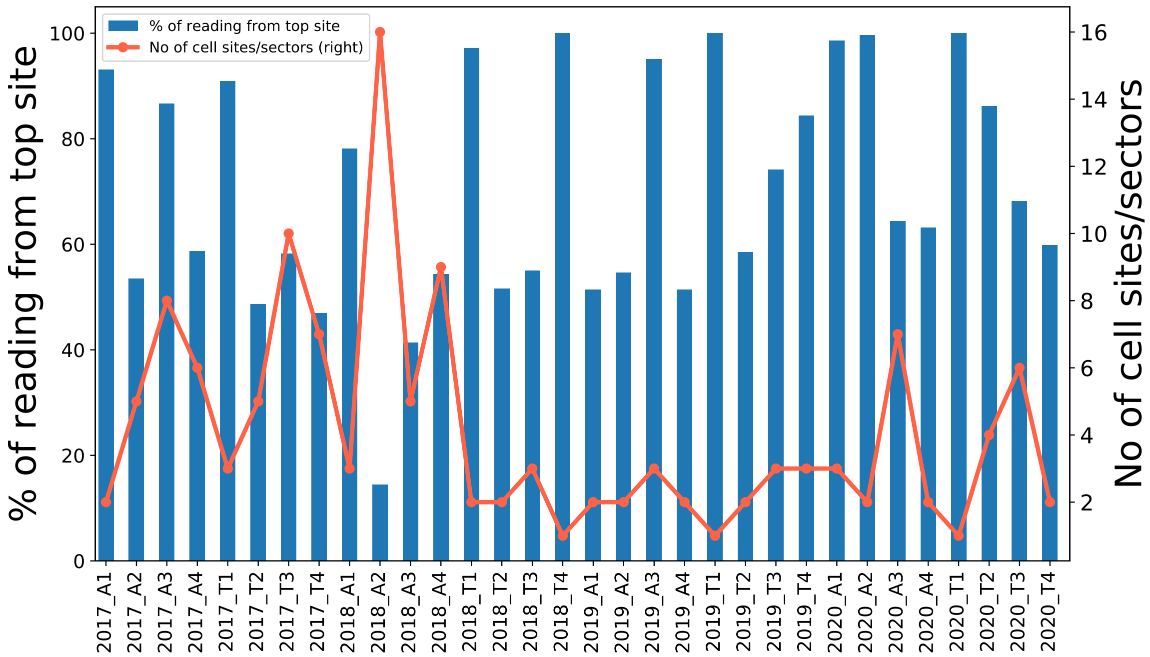
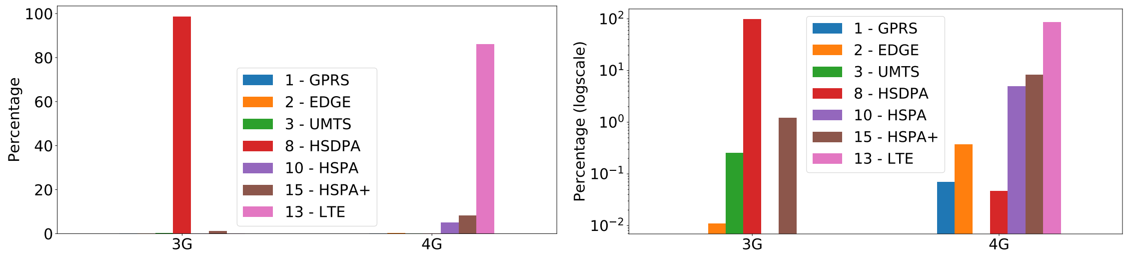


Fig. 4.9 Number of base stations/sectors in use, and the share of readings from top base station (2017 - 2020)

7. Network type variation (2017 - 2020)

A brute approach to assuring reliability is to use a high-level device in the expectation that the device would always be connected to the high-level network technology when it is available. As most devices are backward compatible upto at least one previous technology generation, the expectation is that there can always be fallback to a lower technology if required. We tested this out by capturing network type using the Android Telephony framework. Figure 4.10a shows the results on a normal percentage scale while Figure 4.10b shows the results on a log scale. We observe that on the Samsung SIII devices 3G devices, 98.5% of measurements were on HSDPA (code 8 on Android Telephony framework) with the rest on EDGE, UMTS and HSPA+; on the Xiaomi M1 4G devices, 86% of measurements were on LTE (code 13 on Android Telephony framework) with the rest on GPRS, EDGE, HSDPA, HSPA and HSPA+. It is indeed remarkable that the 4G devices recorded readings on GPRS and EDGE which can evidently not deliver the reliability expectations on 4G. We remark that ***device type cannot be***

used to provide reliability assurance to SCS because there is always variation in network type regardless of the device-preferred connectivity.



(a) Percentage Scale (b) Percentage Logscale

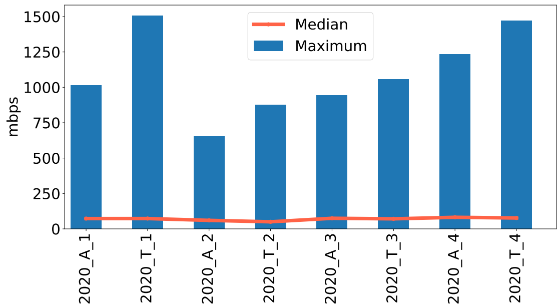
Fig. 4.10 Network type captured on the 3G & 4G devices (2017 - 2020). Figures show that only 86% of measurements on the 4G device, in a location where 4G availability is confirmed, recorded LTE. The rest were mostly on HSPA and HSPA+ although nearly 0.5% were on the 2G network types of GPRS and EDGE.

8. Uplink & Downlink Speeds achieved (2020)

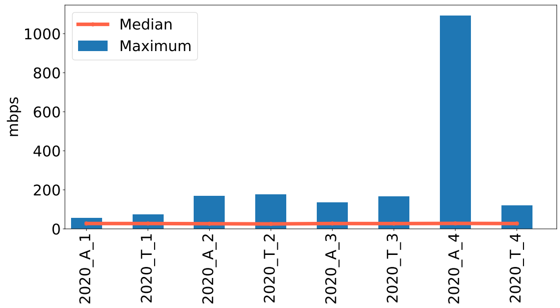
In 2020, we additionally measured uplink and downlink speeds on 4G devices to understand what is realizable speeds on the 4G networks on both atypical and typical days. From Section 2.3.1, we note that in the practical design and configuration of LTE networks, the achievable average downlink throughput per subscriber is inversely proportional to the number of subscribers [70]. Hence, for the atypical day with significantly more users, we expected downlink and uplink speeds to be lower. Figures 4.23 and 4.24 reflects the median and the maximum speeds achieved on the four networks. Our ***results are mixed and do not provide a consistent basis for decision making on reliability assurance for connectivity.*** On the one hand, the maximum speeds conformed to our expectations whereby the speeds on typical days were higher than the speeds on atypical days (except the outlier of NP 4 on atypical day). However, counter-intuitively, the median speeds on atypical days is higher.

9. City centre vs busy train station (2020)

We expanded our measurements in 2020 to London Bridge, the UK's fourth busiest train station, to understand how network performance at a busy train station compares to performance in the city centre. Our starting hypothesis is that, at best, ***performance on the atypical day, at the city centre will be worse than the performance at the train station and the reverse on typical day***. Figure 4.12 shows that it is even worse, with a comparison of 6 hours of RTT data from 20:00 to 02:00 for city centre (NP1 & 3) and train station (NP6 & 7). We see that



(a) Downlink Speeds



(b) Uplink Speeds

Fig. 4.11 Downlink and Uplink speeds (2020). The maximum speeds on the typical day was higher than the atypical day speeds (except for the outlier NP4 uplink). However, counter-intuitively, the median speeds on all 18 options was higher on the atypical day than on the typical day.

on both atypical and typical days, NP6 & 7 at the train station had a worse performance than NP1 & 3 at the city centre. On atypical day, share of RTT readings below 100ms were 93% for NP 1, 92% for NP 3, 76% for NP 6 and 88% for NP7. On typical days, it was 90%, 95%, 81% and 89% respectively. We note that there were many more base stations at the train station - evidently to improve performance - and yet the performances were worse.

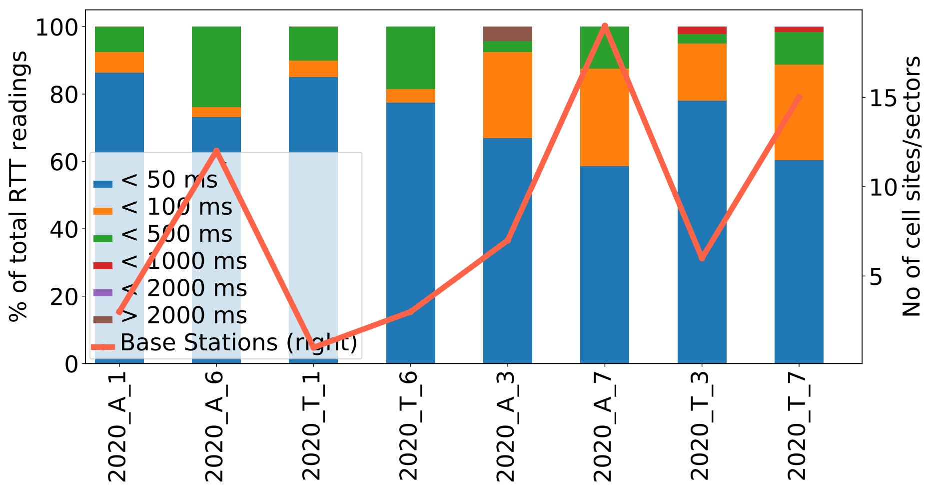


Fig. 4.12 RTT in City Centre (NP 1/3) vs London Bridge train station (NP 6/7) (2020).

10. MNO vs MVNO (2017 & 2020)

In 2017 (for 3G) and 2020 (for 4G), we took additional measurements at the city centre location on a mobile virtual network operator (MVNO) - NP5. We selected an MVNO that uses the network of NP1 to provide a basis for comparing an MVNOs service and that of its host mobile network operator (MNO). Given their reliance on their host MNOs, especially for radio access infrastructure, our hypothesis is that the MVNO should not have a superior performance to its host MNO. As expected, Figure 4.13 shows that both the mobile network operator (MNO) and the MVNO share base stations. Next, we compare the RTT measurements over the 6 hour window from 20:00 - 02:00 for the MVNO and MNO. We show in Figure 4.14 that ***the MNO had a noticeable better performance than the MVNO on both 3G in 2017 and 4G in 2020***. This is similar to the observations in [140] but contrary to the conclusion of [141], highlighting the lack of consensus from previous studies on how MVNO performance compares to host MNO performance.

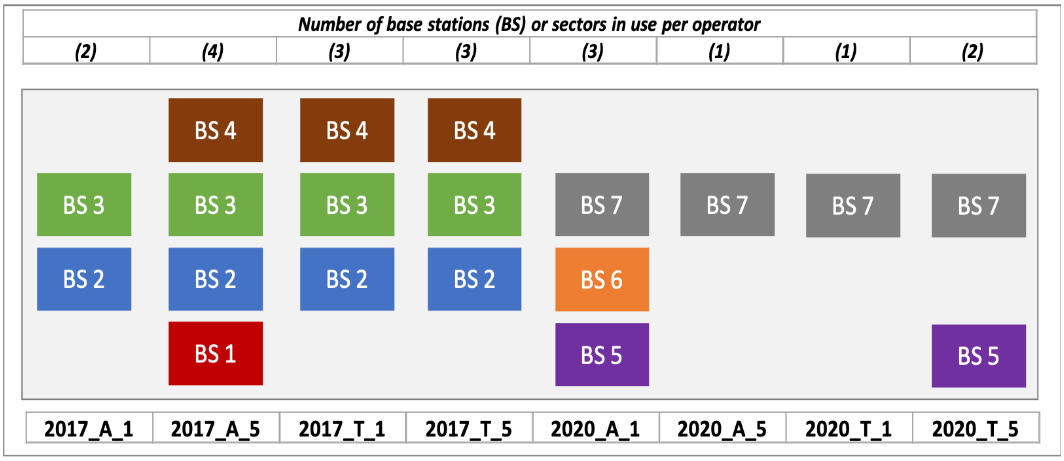


Fig. 4.13 Comparing the base stations used by an MVNO (NP5) and its host MNO (2017 & 2020).

4.2.4 Discussion

1. Performance inconsistencies in pre-5G networks

The primary insight from Experiment 1 is that pre-5G public mobile networks struggle with performance inconsistencies whether on atypical or typical days. Using RTT as the benchmark parameter, we show that only 58% of 4G measurements would have achieved a 50ms target, at a time when most 4G networks in mature markets boast of delivering a better performance. We confirm that this underperformance can be as high as 1.2x worse on atypical days as crowds can produce human congestion above and beyond the designed and operational capacity of the local mobile network. While network operators can plan ahead for predictable crowded events, such as New Year fireworks celebrations, many of the scenarios where SCS will be used may not be predictable enough for the operators to take

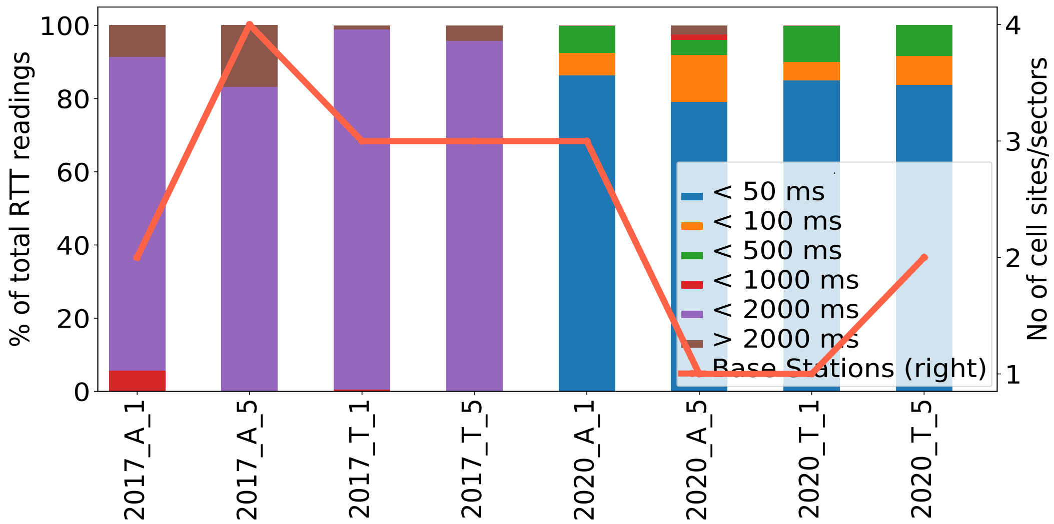


Fig. 4.14 Comparing the RTT performance of an MNO and an MVNO hosted on the same MNO (2017 & 2020).

mitigating actions. This is evident in the city centre vs train station analysis where, despite the seeming predictability of human traffic on the train station, performance was still poor.

2. Democratic impact of poor performance

As should be expected given the prevailing net neutrality paradigm, applications and services receive generally the same treatment from the network. We confirm this with an application level analysis in Figure 4.7. This means that the impact of underperformance in connectivity will be largely democratic and affect all services equally. It therefore puts the onus on the owners/providers of the service to optimise their service accordingly in order to sustain effective operational performance even during connectivity underperformance. The analysis of PLT shows how the different sizes of the homepages of the different websites would generally impact on overall application experience. We expect that services from large internet companies will generally use the best/latest technical approaches to optimise performance. However, many SCS may be provided by technically not-so-savvy government agencies or under-resourced bodies, potentially making them more susceptible to underperformance in the connectivity fabric.

3. Insufficient remedial action by network operators

In our work, we investigated several industry approaches to improving the reliability of the connectivity.

Firstly, we observed significant diversity in the number of base stations used despite conducting all the measurements at a single, stationary/fixed location. Standard industry practice is to add more mobile base stations in order to offload traffic from congested networks. Yet, the performance indicators do not offer us conclusive confidence that this availability of multiple base stations to connect to, will ameliorate the impact of congestion. One reason for this outcome could be that the congestion has migrated from the access network to the transport/backbone network, therefore transferring the bottleneck from the base station to the backhaul infrastructure.

Secondly, our measurements progressed from 3G devices in 2017 to 4G devices in 2018/2019/2020 with the expectation that the measurements on the 4G devices would experience significantly better performance. While this was largely the case, we could also see that some measurements on the 4G device was actually done using the 2G technologies of GPRS and EDGE. We opine, from this, that device type should not be used to provide reliability assurance as there is always variation in network type regardless of the device-preferred connectivity.

Thirdly, the option of relying on special MVNOs to deliver improved performance is probably a mirage. In popular imagination, an MVNO can be the best 'mobile operator' in the country. This was evident in the 2000s when Virgin Mobile UK, an MVNO on T-Mobile's network, consistently topped the charts as the best 'mobile operator' in the country[8]. However, our results show that this MVNO comparative advantage is not only contractually improbable, but also realistically infeasible.

4. Potential for perverse incentives for network operators

Our measurements and analysis centred on RTT instead of speed/bandwidth/throughput. RTT is important and increasingly sought after for enterprise applications where machines are not limited by human's slow response time. In reality, speed is by far the most common way of measuring performance of mobile networks by users, regulators and consumer groups. Consequently, mobile operators take strong interest in optimising their networks to deliver improved speeds. This may partly explain why the maximum speeds were even higher on atypical days when it is more likely that regulators would be monitoring network performance. However, by focusing on RTT, we have been able to uncover better insights into the nature of mobile network performance beyond what we could have got by focusing on speed analysis.

[8]https://www.marketingweek.com/virgin-mobile-tops-polls-of-uk-mobile-phone-operators/

4.2.5 Conclusion and next step

I conclude from the analysis in this section that pre-5G networks are unlikely to provide reliability/QoS guarantees for SCS. For example, focusing only on overall RTT performance, we see that only 58% of RTT measurements on 4G would achieve a 50ms latency target. This provides the answer to ***RQ 1.1*** on how reliable 3G and 4G networks are to deliver guaranteed QoS guarantees to end users.

This raises the question on whether these measurements on mature public 3G and 4G networks offer any insight on what could likely happen on mature, public 5G networks. My answer is yes. I believe that the insights are applicable because ***if mature 3G and 4G networks cannot consistently deliver to their expected specifications, then a mature 5G network is unlikely to perform differently***.

The next logical step is to investigate performance on actual 5G networks to experimentally evaluate if they can consistently deliver guaranteed reliability/QoS for SCS.

4.3 Experiment 2: Performance in an early UK 5G Network

5G networks are now proliferating around the world since the first networks were launched in South Korea and US in Q4 2018. By end 2021, the Global mobile Suppliers Association (GSA) reported that a total of 200 operators in 78 countries/territories have launched commercial 3GPP-compliant 5G services[142]. This includes 5G networks for both mobile services or for fixed wireless access, and regardless of the spectrum in use.

In reality, a lot is riding on 5G networks for mobile services to support existing and new use cases, including SCS. If the 5G vision outlined in [1, 2] is to be realised, then 5G networks will need to deliver connectivity with predictable QoS/reliability assurance. This realisation is partly why there are growing calls for 'consistency' in network performance and experience for 5G. For example, [143] has argued that the pursuit of higher speeds for 5G is misguided and that the focus ought to be on providing a *consistent* rather than *speedier* cellular experience. My task is this section is to explore the nature of performance in early 5G networks to understand their suitability for the connectivity for SCS.

4.3.1 Aim

The primary aim of this measurement study is to provide an answer to **RQ 1.2** on whether the extra capacity and design improvements in public 5G networks will be sufficient to provide consistent and enforceable QoS guarantees for SCS. We identify that early 5G networks were going to be used to supplement 4G capacity at congestion points such as train stations. Hence, we focused the measurement at the five busiest train stations in the UK and set out to understand two questions:

1. What does it mean for a user to be connected to 5G? This became necessary as early press commentary suggested that the 5G icon did not always indicate connectivity *through* 5G either because the 5G icon was being used irresponsibly[58] or as a consequence of the decision in 2018 to accelerate roll out of 5G by decoupling the access and core network developments[57].

2. How does user-perceived network performance in early 5G networks compare with 4G? This is an important comparative exercise given initial expectations that 5G will be more than 10x better than 4G in speed and latency[1, 2].

4.3.2 Methodology

The open question in RQ 1.2 provides the background to the field measurements we conducted at the five busiest train stations in the UK over five days in January 2020. This study was on only one network operator and we measured 4G and 5G from the same mobile network operator.

1. Measurement philosophy & design

Our core measurement philosophy is to understand the comparative performance of mature 4G and early 5G networks. We designed to control or investigate the same five major factors (as identified in [19]) that affect bandwidth availability for users: radio technology, devices, coverage, handovers and congestion. On radio technology & devices, we used different 4G and 5G devices, in the hope of connecting to the respective 4G and 5G networks. On coverage & handovers, we used a fixed location with verified 4G & 5G network availability and no user-triggered handovers. On congestion, while the volume of human traffic is the same, we recognize that the 5G network is likely to be lightly loaded as only very few persons would have a 5G device in early 2020. Therefore, this ought to improve the performance we receive on the 5G device.

2. Date, time & location

Our study design was for five days at the five busiest train stations in the UK in the week starting Monday, 20th January 2020. Coincidentally, according to the UK's Office of Rail and Road (ORR) data for 2017/18, these were all in London: Waterloo, Victoria, Liverpool Street, London Bridge and Euston[144]. However, we opted for Kings Cross/St Pancras instead of Euston as St Pancras (no 9) and Kings Cross (no 10) are co-located, with a combined traffic that puts them in no 3. Lastly, on 27th January and for comparison, we also did measurements in Stevenage Hertfordshire, 30 miles from London, where 5G was not available. On each of the five working days, between 17:30 - 19:00pm, we sat at the train concourse of one of the stations and measured network performance as if we were an average commuter during the busy, pre-Covid-19, rush hour. We assume that this is a time when commuters will be enjoying video-rich infotainment activities. In total, we had 450 minutes (7.5 hours) of measurement data at the train stations and 540 minutes (i.e. 9 hours) across all our measurements.

3. Measurement devices

We conducted the measurements using the Netperf-KCL app installed on a single 4G and 5G device connected to the network of the same mobile operator. For 4G, we used a Samsung Galaxy S6 Edge+, version G928F that was released in August 2015, runs Android version 7.0 and supports Android SDK version of up to API level 24. The CPU is the Octa-core (4x2.1 GHz Cortex-A57 & 4x1.5 GHz Cortex-A53), which runs on the Exynos 7420 Octa (14nm) chipset. For 5G, we used a Samsung S10 5G, version SM-G977B that was released in April 2019, runs Android version 9.0, supports Android SDK version of up to API level 28 and has a 5G-enabled SIM card. Its CPU is the Octa-core (2x2.73 GHz Mongoose M4 & 2x2.31 GHz Cortex-A75 & 4x1.95 GHz Cortex-A55), running on the Exynos 9820 (8nm) chipset.

4.3.3 Results

In this subsection, I summarise the results from the train stations in Central London and the suburban Stevenage location. Overall, we captured 4,408 readings across 14 categories, 2,184 on the 4G device and 2,224 on the 5G device during our measurements in January 2021.

1. Device 5G icon indicator

We confirm that the 5G icon was present for most or all of the time during the 450 minutes of measurements in London as we did not visually observe any moment when it was off. At the minimum, based on Section 2.3.6, this suggests that ***all the 4G base stations at the train stations have been upgraded to support 5G NSA and that the network operator has opted to set the* upperLayerIndication *bit to true***.

2. Android's Network Type indicator

Following the visual observation of the 5G icon, we evaluated the readings from the 5G device to establish if there was actually 5G connectivity. Intriguingly, out of its 2,224 readings, 100% of them reported an Android network connection to only LTE as the 'Network_Type' api returned code 13 - for LTE - for all the readings. In other words, out of the 1,804 readings in London locations where there was a 5G icon on the device and 420 readings in Stevenage where there was no 5G icon, we only saw 'Network Type = 13 = LTE' and absolutely no Network Type = 20. We eventually uncovered the reason for this contradiction as we confirmed from [55] that ***all 5G phones released before October 2019 cannot be***

upgraded to Android 10, making them unable to show the 5G network type, due to Android API support.

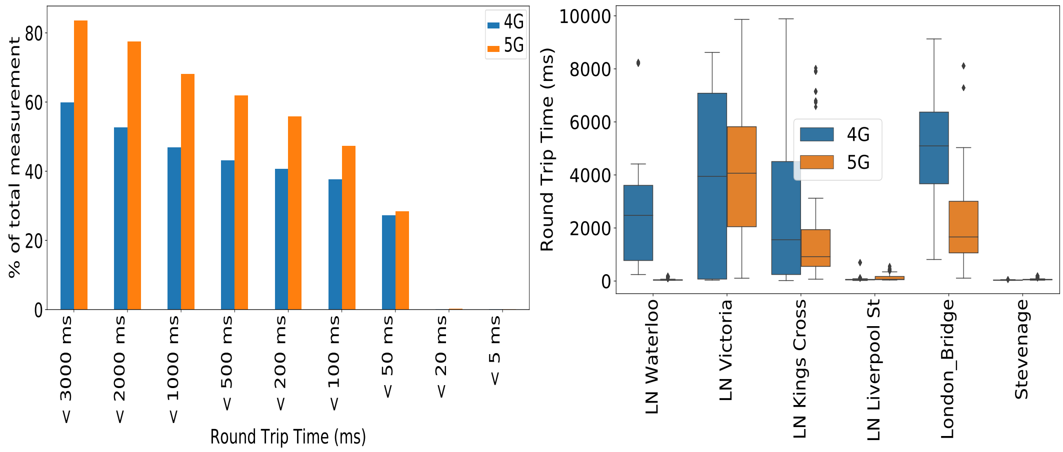
3. RTT variation

Figure 4.15 summarises the analysis of the RTT readings on the 4G and 5G devices. Overall, there is a significant difference in the median RTT: 5G device median was 131 ms whereas the 4G device median was 1,275 ms, an 877% improvement. Delving into the data, we note that the 5G device advantage happened mostly at higher latency targets. For example, for the share of RTT <50ms, we note that this was 28% on the 5G device vs 27% on the 4G device. Likewise, 47% of RTT measurements on the 5G device were <100ms compared to 38% on the 4G device. At 500ms, this was 62% vs 43%; at 1000ms, 68% vs 47%; at 2000ms, 77% vs 53%; and at 3000ms, 84% vs 60%.

Given that 5G is expected to deliver significantly lower latency, our observation was that there were no demonstrably clear performance advantage with the 5G device on lower latencies. This partly supports the findings of [55] that latency improvement in 5G NSA was insignificant. Considering that the measurements were between an April 2019 5G device vs an August 2015 4G device, it raises the questions on if this was based on improved connectivity or improved device capability. However, assuming the difference is down to the network type, this ***suggests that the main benefit of 5G is to minimise very poor performance rather than to maximise very good performance.*** Generally, this should lead to an overall improvement for all services on 5G, including SCS.

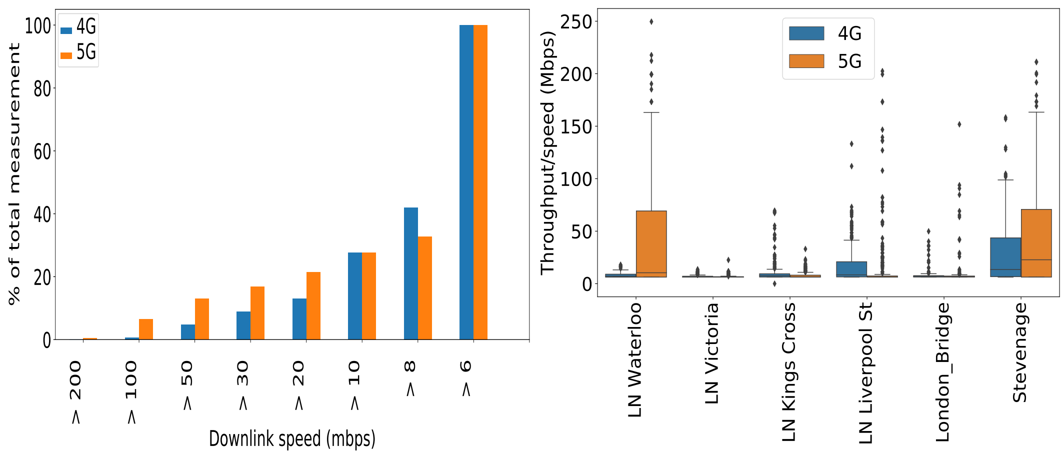
4. Downlink and Uplink Speeds

Figures 4.23 and 4.24 show the downlink and uplink speeds achieved in our results. Unlike RTT, here, the 4G device recorded higher median speeds than the 5G device. For downlink, it was 7.3 mbps vs 6.42 mbps (12% worse performance on the 5G device); for uplink, it was 3.23 mbps vs 3.18 mbps (2% worse performance on the 5G device). However, similar to RTT, both the speeds recorded on either the 5G or 4G device were comparable at lower Mbps levels until after the 70% percentile value. For example, for uplink, the 75% percentile values were 3.5 Mbps on both the 5G and 4G devices. For downlink, the 75% percentile values were 12.9 Mbps vs 10.5 Mbps. We note that the ***5G device outperformed 4G in Stevenage where there is no 5G and both devices were connected to same base station, suggesting a strong role for device improvements***.



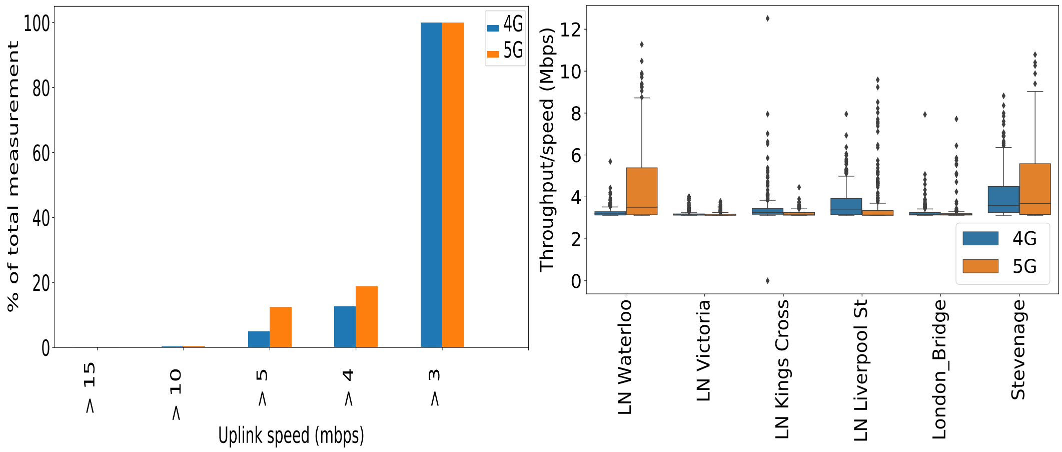
(a) RTT performances by device (b) RTT performance by locations

Fig. 4.15 RTT measurements analysis. (a) shows that performance is much better on the 5G device at higher RTT levels but similar for RTT <50ms. However, overall 5G device median = 131ms; overall 4G device median = 1275ms (b) shows that RTT varies significantly across locations



(a) Downlink speed performances by device (b) Downlink speed performance by locations

Fig. 4.16 Downlink speed measurements (a) shows more 4G device measurements achieved speeds >10Mbps although 5G device had more individual high speeds. This is reflected in overall 5G device median = 6.4Mbps vs 7.3Mbps on the 4G device (b) shows faster speeds on the 5G device in Stevenage where there was no 5G availability and both devices were connected to only 1 base station. This raises the question that on whether device improvements are driving better performances instead of network type



(a) Uplink speed performances by device (b) Uplink speed performance by locations

Fig. 4.17 Uplink speed analysis. Both (a) & (b) show better performance on 5G device. 5G & 4G device median are similar = 3.2Mbps

5. Base Station analysis

Figure 4.18 shows the diversity in the base station IDs that both the 4G and 5G devices connected to. As with Experiment 1, we observe connection to multiple base station IDs, with both devices connecting to a total of 15 base station IDs in the six locations. However, while these were mostly shared by both the 4G and 5G devices, and the 5G device always connected to at least one unique ID per London location, ***the share of measurements from those unique base stations was minimal and so we cannot establish a meaningful correlation***.

6. Video Playback analysis

While RTT and speed measurements offer insights on the performance of a typical application, the insights from downloading and playing back a 30 seconds Youtube video, every five minutes, offered specific insights on video performance. Figure 4.27 summarises the playback time of the videos on the 5G vs 4G devices, showing clearly that the Youtube application playback was worse off on the 5G device - a 19% inferior performance in playback time. In fact, the 5G device only had better performance in Waterloo and Victoria stations. Baring any unknown or unforeseen circumstances, ***it is difficult to see how an inferior application performance can be justified for the 5G device.***

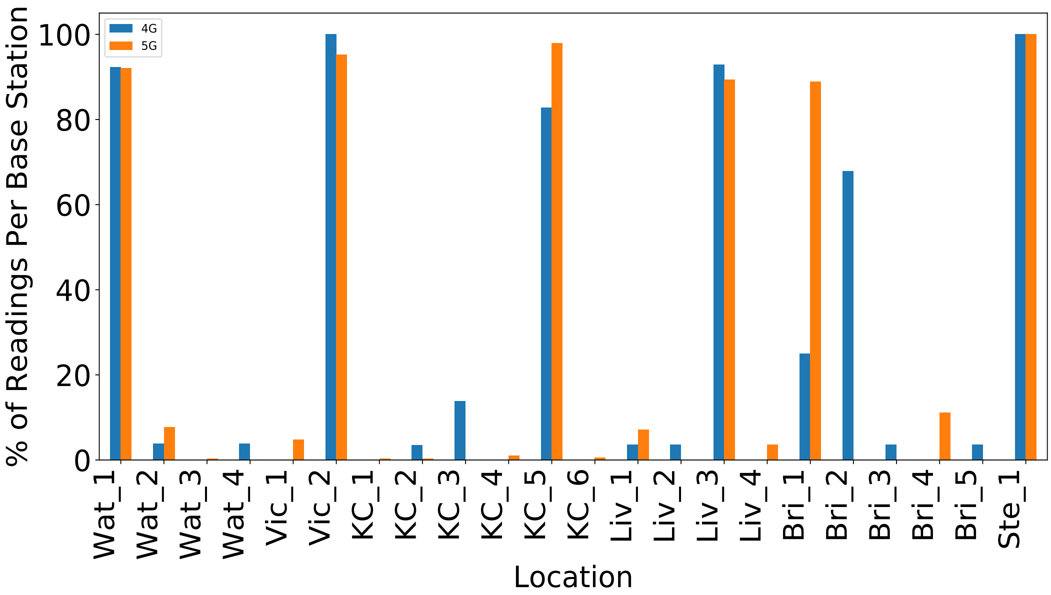


Fig. 4.18 Base stations connected to per location, based on Android's TelephonyManager API. Wat = LN Waterloo; Vic = London Victoria; KC = London Kings Cross; Liv = London Liverpool Street; Bri = London Bridge; Ste = Stevenage

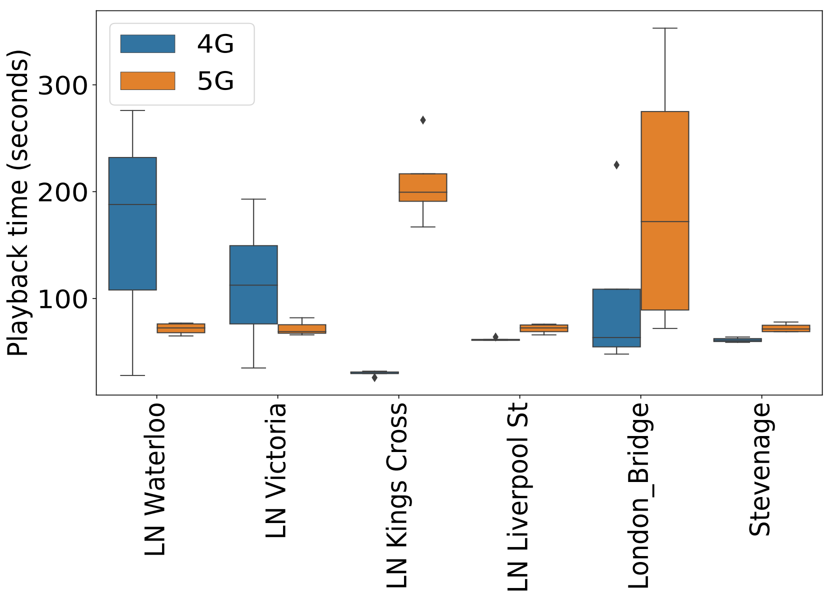


Fig. 4.19 Video playback time per location. Barring any unforeseen reasons, the 19% overall inferior performance on 5G is unjustifiable.

4.3.4 Discussion

1. Unproven source of performance improvement: device vs network

Regrettably, we did not see clear evidence to conclude that early 5G networks in the UK where out performing mature 4G networks. From our results, any superior performance could have come from being connected to a 5G network or the superior capabilities of the much newer 5G device. This means that any SCS use cases cannot immediately assume that connectivity to a new technology will offer an instant improvement in performance, without taking into account other older equipment within the unit. It might be that this is a case of 'new wines wont necessarily be fine in an old wine skin'. Regrettably, we didn't use 4G and 5G SIM cards on the same device to firmly eliminate device performance as a variable in our experiments.

2. 5G will get better with time

From the history of 1G to 2G to 3G to 4G, it is clear that 5G will eventually deliver significantly superior performances over 4G. This may not have happened yet because in the early days of a new network generation, it takes time for operators to fully tune the network for optimal performance. Or it may be that a new network configuration is needed to realise the promised benefit - cf: 3G UMTS vs 3G HSDPA; 5G NSA vs 5G SA.

3. Focus of 5G improvements should be on consistency instead of Speed

As the task of improving 5G continues, it is important for the focus to be on making the 5G experience consistent rather than simply turbocharging its speed. Our latency and downlink readings have already hinted that 5G is improving the consistency of the performance by helping to minimise very poor performance rather than maximising very good performance. This is important for SCS, where a consistent performance is prioritised and supports the call in [143] to focus 5G on ensuring a 'consistent' experience rather than a 'speedier' experience. Accordingly, industry stakeholders should be enjoined to prioritise 5G improvements that deliver a more consistent experience to assure reliability for SCS.

4. Better understanding of consequential changes in standardisation

The big decision to decouple, for the first time, RAN and core network deployment in order to bring forward the launch date for 5G globally does not seem to have worked out well. In February 2017, industry stakeholders decided to accelerate the standardisation of 5G NR and create the 5G NSA option[145]. This acceleration meant that 5G first arrived in late 2018 instead of early 2020. However, in hindsight, it has led to three adverse outcomes:

- **Increased operational complexity:** Decoupling the RAN from the core means that new 5G NR to be connected to the legacy 4G EPC, allowing early deployment of 5G to commence[42]. However, it meant that operators are now faced with five non-mutually exclusive deployment scenarios to choose from.
- **Confusion on when to use 5G icon:** The acceleration of 5G launch date left little time to resolve the 5G icon question or for some product developers to adjust their schedules to accommodate 5G. For the former, this led to disagreements on what the icon should signify (e.g. showing 5G, 5G+ etc) and how many bits to use to reflect this[58]. For the latter, this is reflected in 5G support on devices. For example, iPhones did not support 5G until the release of iPhone 12 in October 2020. Likewise, Android 10, scheduled for release in October 2019, would have been timely for 5G device rollout in early 2020 but not for devices rolled out earlier.
- **An undercooked 5G:** As our measurements have shown, there is no knock out outperformance by the early NSA 5G networks, negating the rationale for its hasty rollout.

5. Plan ahead and better for 6G

Our work, focusing on the transition period from 4G to 5G, provides guidance on how to better prepare for the transition from 5G to 6G. In particular, it is important to ensure that the industry's decade-long technology generation transition is not undercooked when its performance is compared to the previous generation. Likewise, the icon confusion ought to be avoided at all costs - it didnt happen for 2G and 3G, but has now happened for 4G[146] and 5G. Whether the icon confusion comes from a lack of clarity or consensus on how to act, or it is a deliberate attempt by industry players to misrepresent the connection status, the industry should ensure it is not repeated for 6G. Perhaps it also suggests that the industry should go back to a pure/traditional full system deployment instead of the 5G approach of decoupling the RAN and core deployment.

4.3.5 Conclusion and next step

I conclude from this analysis that early 5G networks, based on the NSA standard, do not deliver a compelling performance advantage over mature 4G networks. For example, the share of RTT <50ms results was 28% on the 5G device vs 27% on the 4G device. Given the conclusion from Experiment 1, this suggests that the ***answer to RQ 1.2 is similar to the answer for RQ 1.1: 3G, 4G and early NSA 5G networks cannot consistently provide reliability/QoS assurance for SCS.***

The next logical step is to observe 5G performance as the technology stabilises. I note from [147] that 5G customer adoption has not really moved beyond the early adopter status by end of 2021 in most countries outside South Korea and China. As such, an experiment to comparatively understand how 5G networks perform in South Korea and UK can be helpful to evaluate if they can consistently deliver guaranteed reliability/QoS for SCS.

4.4 Experiment 3: Performance in UK & South Korea 5G networks

It's being over 3 years now since 5G launched. Since then, its pace of rollout and the performances it delivers have been subjected to scrutiny by industry stakeholders and users alike. While it has provided additional capacity to relieve some congested cell sites[148], early reports have raised concerns about its availability (e.g. 'fake' 5G icon)[57], usability (e.g. users needing to turn off 5G to make their phone usable)[149], performance (e.g. not really better than 4G)[55] and viability (e.g. doesn't offer much to telcos)[150].

Having shown in Experiment 2 that early 5G networks in the UK cannot consistently provide reliability/QoS assurance for SCS, it is imperative to investigate more mature 5G networks to understand if/how performance has changed. However, I note from [147] that the UK has not really being a bellwether market for 5G. This may be an irrelevant reality, but it could also be of material consequence in understanding the performance of 5G NSA networks. Consequently, the focus of Experiment 3 is to understand the performance of 5G in a multi-country study and as the technology matures.

4.4.1 Aim

Experiment 3 continues to seek answers to **RQ 1.2** on whether the extra capacity and design improvements in public 5G networks will be sufficient to provide consistent and enforceable QoS guarantees for SCS. Given that South Korea is one of the identified 5G market leaders[147], we focus Experiment 3 on multiple, but similar locations, in both UK and South Korea. On the assumption that these early 5G networks have matured, I set out to understand two questions:

1. How does user-perceived network performance in 5G networks compare with 4G? This is an important comparative exercise given initial expectations that 5G will be more than 10x better than 4G in speed and latency[1, 2].

2. How does 5G performance in UK compare to South Korea? By comparing a market leader (South Korea) and a middling 5G market (UK), this comparison begins to establish a picture of how 5G is performing internationally.

4.4.2 Methodology

The open question in RQ 1.2 provides the background to the field measurements we conducted at four types of locations in South Korea and UK in June / July 2021. The measurements were done on the networks of two mobile operators in each country, and for 4G vs 5G.

1. Measurement philosophy & design

Our core measurement philosophy is to understand the comparative performance of mature 4G and 5G networks in South Korea and UK. We used the same Netperf-KCL app and swapped the 14 UK websites with their Korean equivalents as described in 3.1.

We designed to control or investigate the same five major factors (as identified in [19]) that affect bandwidth availability for users, plus a multi-country comparison between UK and South Korea. On radio technology & devices, we used different 4G and 5G devices, in the hope of connecting to the respective 4G and 5G networks. On coverage & handovers, we used a fixed location with verified 4G & 5G network availability and no user-triggered handovers. On congestion, while the volume of human traffic is the same, we recognize that the 5G network, especially in the UK, is likely to be more lightly loaded. On international comparison, we conducted the measurements at the same type of locations and on the same times in the day (taking time zones into consideration).

2. Date, time & location

Our study design was to focus on four location categories in both South Korea and the UK. This included the largest departmental store, the busiest train station, a commercial location and a residential location. The expectation is that these four locations will present a good representation of 5G availability and performance in both countries. In each location, we measured for two hours each. Hence in total, we had 8 hours (480 minutes) of measurements in each country. Measurements were done in South Korea over two dates in June 2021. When the devices eventually arrived to the UK, we did the UK measurements on three dates in July 2021. The summary of the dates, locations and rationale is summarised in Table 4.2.

3. Measurement devices

We conducted the measurements in South Korea and UK using the Netperf-KCL app installed on the same 4G and 5G devices. In total, we used four devices (two 5G and two 4G) for the measurements and both 5G and 4G devices have SIM cards from the same network operator. The devices were acquired in South Korea and equipped with SIM cards from the two largest

Table 4.2 Exp 3: Dates, locations and rationale for measurements in South Korea & UK

	Category	**Dates**	**Local Time**	**Location**	**Rationale**
1a	Largest Korean departmental store	21 June 2021	11:00 - 13:00	Shinsegye Gangnam Branch, South Korea	These stores have huge footfall of shoppers, with or without the Covid-19 challenges. The measurements were carried out just outside the main front doors.
1b	Largest UK departmental store	12 July 2021	11:00 - 13:00	Selfridges Oxford Street London, UK	Same as above.
2a	Busiest Korean train station	21 June 2021	17:00 - 19:00	Seoul Station, South Korea	At the concourse of the busiest train station in the country where there is a lot of footfall.
2b	Busiest UK train station	12 July 2021	17:00 - 19:00	London Waterloo Station, UK	Same as above.
3a	Commercial location Korea	22 June 2021	11:00 - 13:00	KI Building, KAIST South Korea	At the premises of the leading engineering school in South Korea with evidence of 5G availability.
3b	Commercial location UK	14 July 2021	11:00 - 13:00	Services Park, South Mimms Business District, UK	Business district in the UK where the operators claim there is good 5G coverage.
4a	Suburban residential location Korea	22 June 2021	17:00 - 19:00	Hwaseong-si, South Korea	Residential districts where the operators claim there is 5G coverage.
4b	Suburban residential location UK	13 July 2021	17:00 - 19:00	Luton, UK	Same as above.

operators (described in this work as SNP1 and SNP2). At the end of the measurements in Korea, the devices were air-freighted to the UK and the Korean SIM cards were replaced with SIM cards from UK NP1 and NP3.

For 4G, we used the Samsung Galaxy A12 that was released in December 2020. Its CPU is the Octa-core (4x2.35 GHz Cortex-A53 & 4x1.8 GHz Cortex-A53), running on the Mediatek MT6765 Helio P35 (12nm) chipset. The device runs Android version 10 and supports Android SDK version of up to API level 29. For 5G, we used the Samsung Galaxy A42 5G that was released in November 2020. Its CPU is the Octa-core (2x2.2 GHz Kryo 570 & 6x1.8 GHz Kryo 570), running on the Qualcomm SM7225 Snapdragon 750 5G (8 nm) chipset. The device runs Android version 10.0, supports Android SDK version of up to API level 29 and has a 5G-enabled SIM card.

4. Indirect route to verifying 5G status

In the preliminary measurements to setup our experiment, it became clear that none of the 5G Android devices was recording a 'Network_Type_NR' value of 20. This is despite using the A42 devices released after October 2019[55] and the well established 3GPP and GSMA protocols for determining 5G status[46]. This suggested that, although the devices' engineering service mode shows 5G NR related data, the TelephonyManager API was still incapable of recording a 'Network_Type_NR' value of 20 and was returning code 13 for LTE instead. We therefore modified our Netperf-KCL app, using the approach detailed in NetMonster Core[151], to detect *NetworkType* instance for 5G NR. We validated that this works in both the UK and South Korea, with our Netperf-KCL app recording a value of '1' whenever the device is connected to 5G NR and '0' otherwise.

4.4.3 Results

In this subsection, I summarise the results from the four location categories in UK and South Korea, focusing on comparing the performance of a relatively stable 5G with a mature 4G. Overall, across five days, and for 16 hours of measurements, in eight locations and across two countries, we captured 615,000 data points.

1. Network Type: still ineffective for 5G

Out of the 32,059 'Network Type' readings on the 5G devices in both countries, and in locations that were clearly described as having 5G coverage, 98.6% of them reported an Android network connection to only LTE. That is, the 'Network_Type' api returned code 13 - for LTE - for the readings. The rest included 3G readings for HSDPA, HSPA and HSPA+.

Ordinarily, this contradiction with the visible 5G icon on the devices ought to raise questions about the authenticity of the 5G connection. However, because we already knew about the error from Experiment 2, we ***did not take the absence of Network Type 20 readings as an indication of 5G unavailability.*** Instead, we relied on the NetMonster Core approach to validate 5G availability.

Using the Netmonster Core approach for 5G status indication, we recorded 62% connection to 5G across all our measurements on the 5G devices - Figure 4.20. Diving into the data, we see that 100% of all the 5G devices readings in South Korea reported a 5G connection. This confirms that all the 4G base stations at the test locations have been upgraded to support 5G NSA and that the network operators in South Korea have opted to set the *upperLayerIndication* bit to true. In contrast, in the UK, we recorded 48% 5G compliance on NP3 but 0% on NP1. On NP3, we note that 5G compliance was recorded in each of the four test locations confirming that 5G ought to be present (London Waterloo 98%, Selfridges London 69%, South Mimms 11% and Luton 4%).

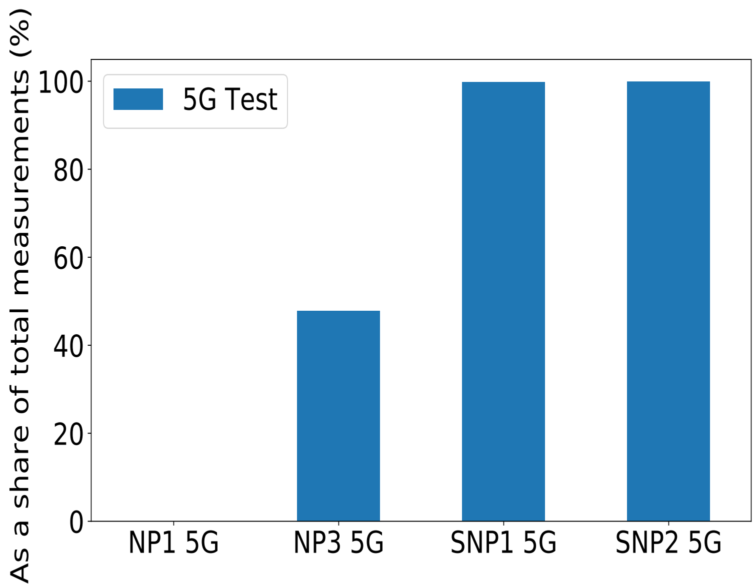


Fig. 4.20 Determining 5G Status with Netmonster Core: 100% of measurements on the Samsung Galaxy A42 5G in South Korea recorded a 5G status; in the UK, this was only 24%.

2. Overall performance: latency vs speed

The first result is the overall picture of performance on the 4G and 5G networks of the four operators in the two countries. This is an important performance criteria to determine whether the performance is consistent and stable. [143] argued that the focus of 5G development and policy making ought to be on ensuring a 'consistent' experience rather than a 'speedier' experience. Looking at RTT measurements in Figure 4.21, we did not see any evidence that 5G delivers a more consistent performance than 4G. About 91% of 5G measurements recorded latency of <200ms compared to 94% for 4G. For <100ms, the share of 5G RTT measurements meeting it is 83% whereas it is 90% for 4G. In contrast, we see that the 5G measurements significantly outperformed 4G on speed, with 96% of 5G measurements achieving downlink speeds of over 100Mbps unlike 87% for 4G. ***Our results suggest that 5G is over-performing 4G on speed but not on latency.*** While this over-performance on speed might look good in press and analyst reports, it is of little comfort to SCS use cases seeking reliability of their connectivity options.

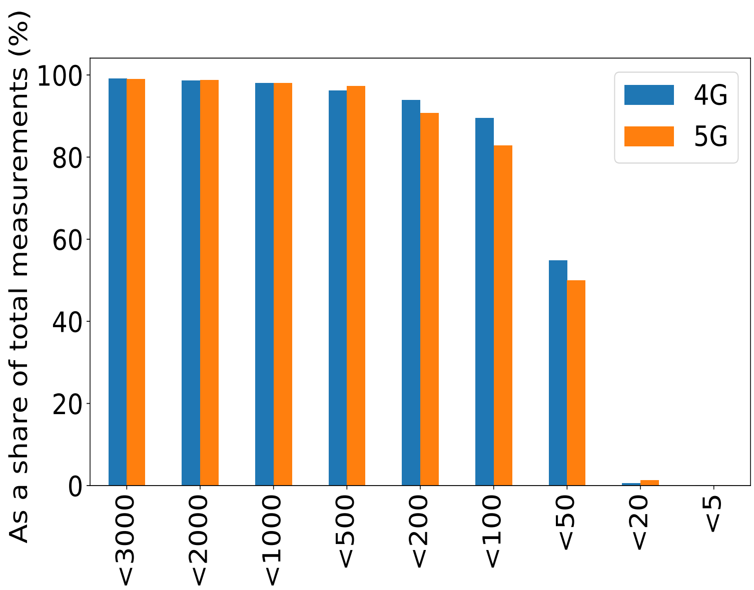


Fig. 4.21 Consistency in RTT performance. Our 5G measurements did not deliver a more consistent performance than 4G

3. RTT: 5G performing worse than 4G

Figure 4.22 shows the proportion of all RTT measurements below the 50ms end-to-end 5G latency benchmark was 52% compared to 86% that achieved the 4G benchmark of 100ms. We note that this was as a result of poorer latency performance in Korea where only 40% of measurements hit the 5G benchmark unlike 60% for the UK. Intriguingly, ***55% of readings on the 4G device had RTT below 50ms whereas only 50% of readings on the 5G device achieved same. This superior performance on 4G happened in both South Korea and UK***

showing clearly that current 5G networks are not delivering the expected low latency level and clearly not better latency than 4G networks. This corroborates the findings in [55–57]. It means that non of the ultra low latency services touted as 5G use cases (e.g. remote surgery and autonomous driving) can be supported by today's public 5G networks.

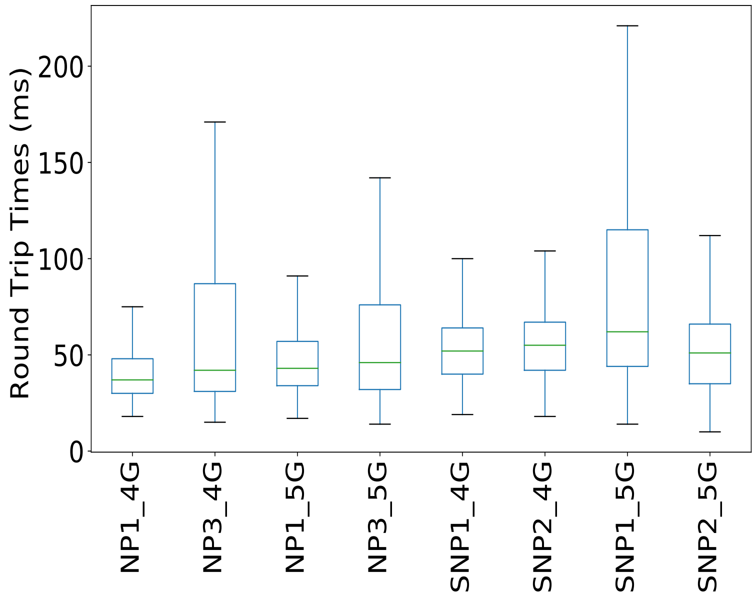


Fig. 4.22 RTT analysis: 5G under-performs, both against benchmark and 4G.

4. RTT Variance: 5G shows stability

Unlike RTT, the RTT Variance performance was better with 60% of all RTT Variance readings below 10ms and 80% below 20ms. We note that the Korean networks (both 4G and 5G) had significantly better RTT Variance (81% of 5G and 76% of 4G readings achieving 10ms). The results for RTT Variance are moderately encouraging, suggesting a low variation in latency on the 5G networks. This would make it easier to support services that require stable latency (e.g. AR/VR).

5. Downlink Speeds: possibly over-performing on 5G

Figure 4.23 shows that the proportion of all measurements that achieved the expected 5G downlink speed of 100Mbps was 90% compared to 92% that achieved the 4G benchmark speed of 50Mbps. On the 5G devices, thanks to strong results from Korea, 95% of readings achieved 100Mbps while 96% achieved 50Mbps. In Korea, 99% of 5G device measurements achieved 100Mbps while only 90% of UK 5G device met the 100Mbps benchmark. In fact, 99% of 5G device readings in Korea achieve 250Mbps downlink speeds. On the 4G devices, 85% achieved 100Mbps while 87% achieved 50Mbps. The main insight we draw from our results is that ***the 5G networks are meeting the benchmark and perhaps over-delivering on downlink speed.*** We suggest that this is because downlink speed is the defining benchmark for most internet measurement studies[143] and shapes much of the discourse on how superior 5G is over previous generations. Yet, [77] warns that speed is not everything as faster speeds may not always bring a better experience to end users while [143] argued for the focus of 5G to be on a more *consistent* experience rather than a *speedier* experience.

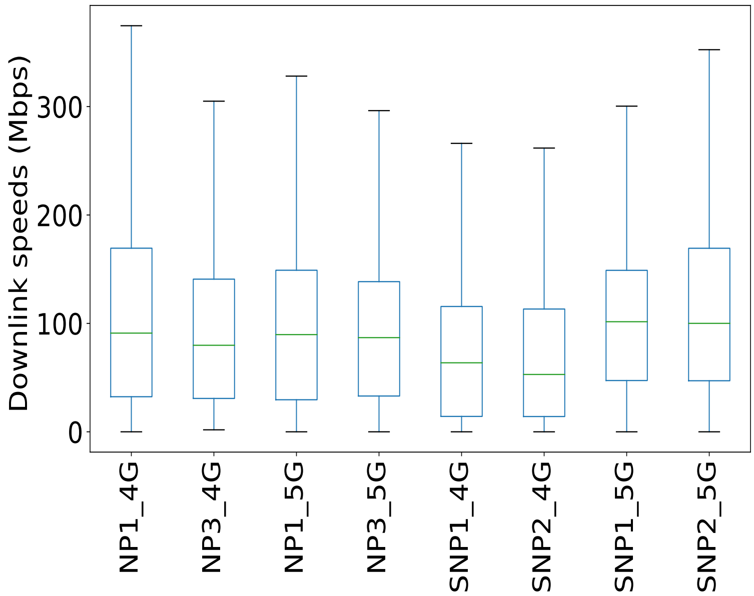


Fig. 4.23 Downlink analysis: Only parameter where 5G is consistently exceeding its benchmark and out-performing 4G. Overall, the share of 5G readings achieving the 100Mbps benchmark is 9 percentage points (pp) higher than the share of 4G achieving same.

6. Uplink Speeds - still inferior to downlink

Figure 4.24 shows that the proportion of all measurements that achieved the expected 5G uplink speed of 50Mbps was 23% compared to 65% that achieved the 4G benchmark speed of 25Mbps. On the 5G devices, only 28% of readings achieved 50Mbps while 78% achieved 25Mbps. On the 4G devices, this was 17% and 52%. We note that 40% of 5G device measurements in Korea achieved 50Mbps and 99% achieved 25Mbps. In contrast, in the UK, 16% achieved 50Mbps while 57% achieved 25Mbps. It is notable that unlike downlink, uplink is not meeting its benchmark. While downlink has been historically more important given the direction of internet downloads, SCS use cases and the boom in use of bi-directional video conferencing, especially during the pandemic, has drastically increased the importance of uplink for the 5G era[21]. We argue that by over-delivering - perhaps needlessly - on downlink and under-delivering on uplink, the public 5G networks are proving themselves incapable of providing reliability assurances to SCS, thereby driving many enterprises to opt for private 5G networks[152].

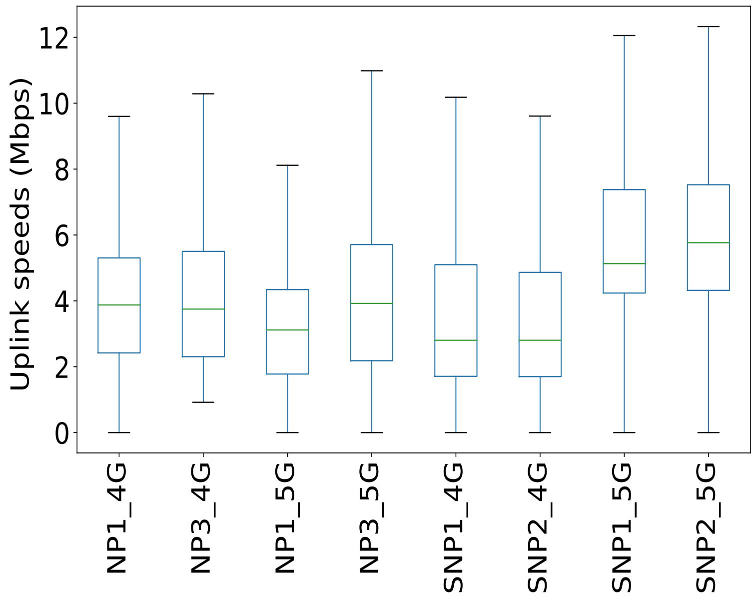


Fig. 4.24 Uplink analysis: Neither 5G nor 4G readings are meeting the 50Mbps benchmark but overall, the share of 5G readings achieving the 50Mbps benchmark is 19 pp higher than the share of 4G achieving same.

7. PLT: boosted by speedier downlink

Figure 4.25 shows that overall, 52% of readings achieved the 500ms 5G PLT benchmark, with 55% of the 5G devices and 49% of the 4G devices achieving it. For the 4G PLT benchmark of 1000ms, 68% of all measurements achieved it; 74% on 5G devices and 61% on 4G devices. Overall, ***the faster speeds on 5G has contributed to faster PLT on the 5G devices than on the 4G devices/network***. We note that, at 80%, the Korean 5G networks significantly outperform the UK websites (32%) in achieving the 500ms PLT benchmark. This is not surprising considering that the Korean 5G networks had the fastest downlink speeds: the faster the speed, the speedier it is to download all the components of a webpage. This ought to be a positive for the QoE but given the results for packet loss and latency, it suggests that PLT is not sufficient to deliver good QoE, corroborating [56].

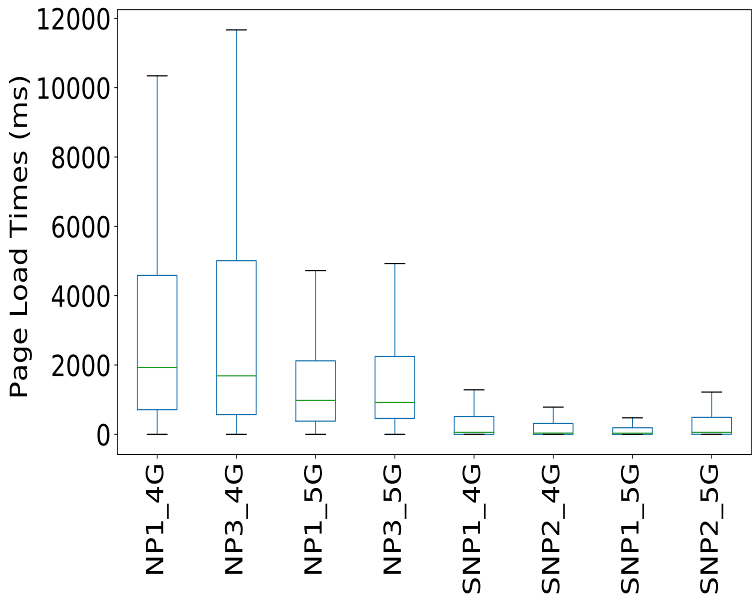


Fig. 4.25 PLT analysis: 5G does not meet benchmark but is better than 4G thanks to a speedier downlink

8. Packet Loss: local policy hurting Korea

The packet loss reflects the underlying data plane reliability along the communications path and we measure it by observing the losses recorded during our ping measurements. A lossy path will negatively impact application performance, hence undermining the overall reliability of 5G connectivity for enterprise services. From Figure 4.26, we see clearly that NP3's 5G network in the UK performs better than its 4G network while NP1's 5G and 4G networks have similar 86% of readings with 0% packet loss. The NP3 data ***suggests 5G was more reliable but this cannot be confirmatory given that the 5G network is currently not as loaded as the 4G network***.

For South Korea, Figure 4.26 shows that packet loss, whether for 4G or 5G were high, losing 100% of the packets in 65% of readings. Looking closer, we observe that losses were only happening for the Korean-focused websites whereas the more internationally-focused websites (i.e. Google, Facebook, Youtube) all had 0% losses. There were some Korean services (like the foreigner friendly KBS and coupang) with 0% losses too, implying that these sites were more web-standards compliant than other websites. Our investigation suggests that the loss of packets in many Korean services is linked to the use of ActiveX control for web security, as part of the SEED public key infrastructure (PKI) introduced in Korea in 1999[153]. Its use has led to a monoculture in web security, overly restricted use of non-Internet Explorer browsers, mandated stringent ID verification and generally made the web browsing experience poor in Korea[154]. Given that Android webview used in our app is based on Chrome, we observe that Korean-centric websites that have been optimised for Internet Explorer and ActiveX had poor performances compared to internationally-focused websites that rely on HTML5 technologies[9].

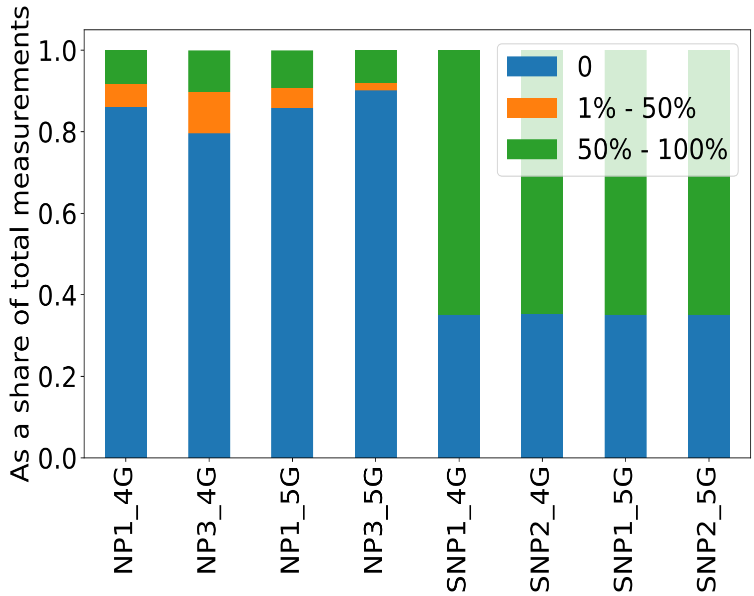


Fig. 4.26 Packet Loss analysis: Poor performance in South Korea is linked to the use of ActiveX in Korean-focused websites

9. Video Playback: better on 4G

To get further insights on application performance, we downloaded and played back a 30 seconds YouTube video, every three minutes. Figure 4.27 summarises the playback time of the videos, with an overall total number of playbacks of 6007. We note that 5G devices had a median playback time of 32.5 seconds while 4G devices was 30 seconds, suggesting that the 4G networks delivered better application QoE. ***This insight of superior performance on 4G was consistent for both South Korea and UK and corroborates [57] which showed a 19% inferior performance for the playback time on 5G***.

[9]The government and many website owners are aware of this and trying to tackle it. But as always, the ‘sunk cost’ is great and the transition has been slow, especially for non-mobile pages

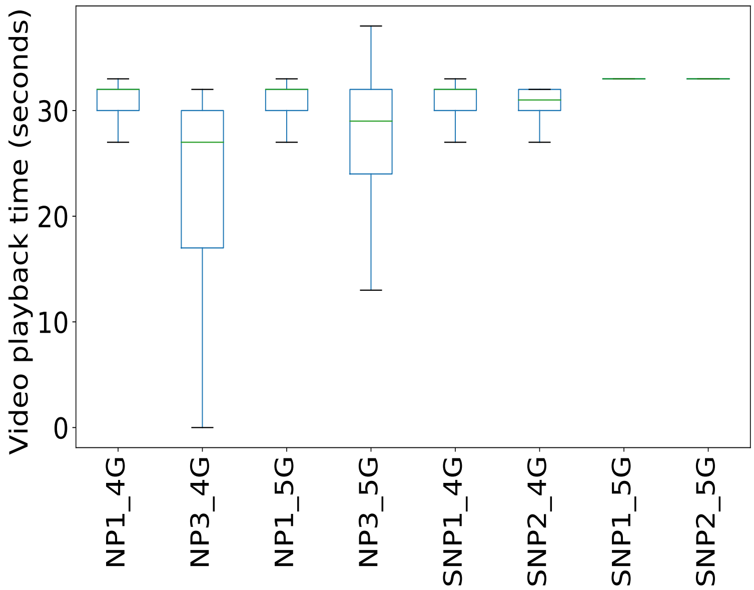


Fig. 4.27 Video playback time in South Korea & UK

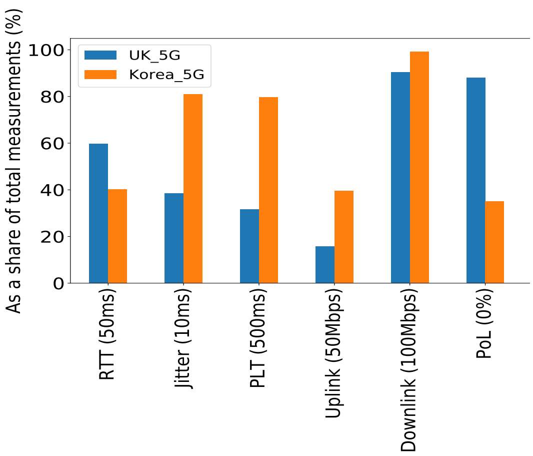
4.4.4 Discussion

1. Experiment 3 validates Experiment 2: 5G is not much better than 4G

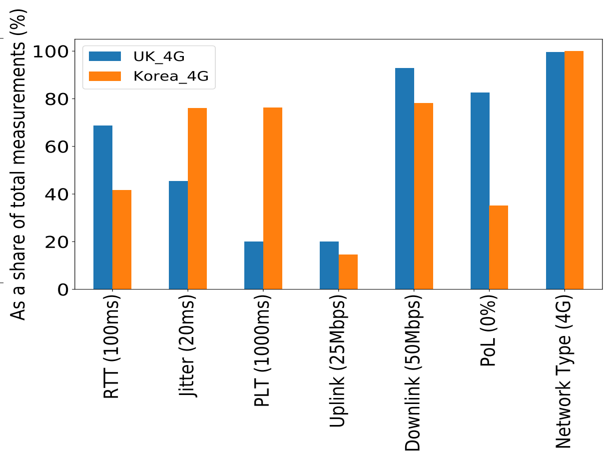
A major reason for our multi-country experiment was to investigate if the results in Experiment 2 were sufficiently representative of what 5G delivers in a stable, operational scenario. From the results presented, it is clear that Experiment 2 was not an anomaly and is actually in sync with the results in Experiment 3. We show that for latency and video playback, the more heavily loaded 4G networks are actually performing better. 5G performs better mostly for downlink speed, again, reflecting the perverse incentive which can encourage operators to optimise for speeds as that is the parameter that defines 5G in the public discourse. As such, the discussions in Section 3.4.4 on whether 5G networks are suitable to provide reliability guarantees to SCS, are equally valid in this section.

2. 5G Comparative analysis - South Korea vs UK

An advantage of a multi-country study is that we are able to compare the setup of the public 5G networks in different countries to better understand how they are able to support ECS users. In Figures 4.28a and 4.28b, we provide a summary of the comparison for 5G and 4G networks in South Korea and the UK showing the percentages of readings that met the benchmark in Table 6.2. On 5G networks, South Korea had better outcomes for Uplink and downlink speeds, RTT Variance, plt, and 100% record of 5G connectivity during our measurements. This is in line with market reports about South Korea's 5G leadership in downlink speeds[155]. In contrast, UK had better outcomes on 5G for latency, packet loss and video playback. On 4G networks, the UK additionally had better outcomes for uplink and downlink speeds. Regardless, apart from downlink speeds, neither the public 5G networks in South Korea nor the UK are delivering the performances required for many of the 5G use cases described in [2].



(a) On 5G networks



(b) On 4G networks

Fig. 4.28 Comparison of the performance on 4G and 5G networks in South Korea and UK

4.4.5 Conclusion and next step

I conclude from this analysis that NSA 5G networks does not deliver a compelling performance advantage over mature 4G networks. This is now evidenced in early 5G networks in the UK and the more mature 5G networks in South Korea. For example 55% of readings on the 4G device had RTT below 50ms whereas only 50% of readings on the 5G device achieved same. It thus suggests that the benefits of the early launch of 5G, which has caused some disruption in the usual process of introducing a new mobile technology, has been difficult to demonstrate. This corroborates [55, 150, 149, 57].

Accordingly, the conclusion from Experiment 3 is similar to the conclusion from Experiments 1 and 2 and the ***answer to RQ 1.2 is similar to the answer for RQ 1.1: 3G, 4G and NSA 5G networks cannot consistently provide reliability/QoS assurance for SCS.***

If all existing mobile networks cannot deliver reliability guarantees for SCS, then it follows that both operators and users should seek for remedies that can mitigate the performance shortfalls. In this work, I propose that D-QoS provides a robust set of options.

4.5 Chapter Summary

This chapter has reported the results from three extensive experiments to understand the nature of mobile network performance in normal/typical times, in abnormal/atypical times, at busy locations and across two countries. Experiment 1 was a four year longitudinal measurement study on 3G/4G in the London UK during the New Year fireworks celebration from 2016/17 to 2019/20. Experiment 2 was focused on 4G/5G measurements at the five busiest train stations in the UK in January 2020. Experiment 3 was on 4G/5G measurements at eight busy locations in UK and South Korea.

From the overall results, I highlight the latency figures as latency is a major selling point for the use of 5G for SCS. For experiment 1 in London UK between 2016 - 2020, we show that only 58% of RTT measurements on 4G would achieve a 50ms latency target. For experiment 2 in London UK in 2020, this was 27% on the 4G device but only 28% on the 5G device. For experiment 3 in London UK and Seoul South Korea in 2021, 55% of readings on the 4G device had RTT below 50ms whereas only 50% of readings on the 5G device achieved same.

Specifically, the insights from Experiments 1 - 3 provide strong evidence to answer RQ 1.1 and 1.2. I conclude that the ***answer to RQ 1.1 is similar to the answer for RQ 1.2: 3G, 4G and NSA 5G networks cannot consistently provide reliability/QoS assurance for SCS.***

I acknowledge that many of these results are obvious and can be intuitively inferred by a relatively knowledgeable person. This is a limitation in my work. However, this does not detract from the necessity to empirically confirm the results and, crucially, quantify the reality. Without such empirical confirmation and quantification, it will be difficult to convincingly make the case for D-QoS nor establish a comparative benchmark to compare a D-QoS vs a non D-QoS setup for QoS guarantees for SCS.

In the next chapter, I focus on addressing ***Research Question 2 on how D-QoS mechanisms can be used to provide QoS guarantees for SCS systems.*** I will explore the history of D-QoS to understand the motivations for their introduction and the limitations they have faced in the market. This will provide a basis to introduce a conceptual framework to codify and clarify D-QoS mechanisms and then, provide guidance on how to use them.

Chapter 5

Introducing Differentiated QoS Connectivity

Mkpịsị aka niile ahaghị otu (Igbo Proverb)

(All fingers are not equal)

5.1 Introduction

Since the emergence of mass market data communications networks in the 1970s, the networking community continues to develop a lot of standards and mechanisms to improve the 'tailorability' of data networks[25]. I refer to these approaches as Differentiated QoS (D-QoS) connectivity mechanisms/approaches and these can be used for commercial, operational or regulatory reasons. The logic for D-QoS is a classical economics approach - in a communications channel with finite capacity, each user or data packet is competing in a zero sum game for network resources. Data networks mostly offer probabilistic or 'best effort' connectivity in contrast to the deterministic connectivity of telephones offering a guaranteed circuit between two end points. In his seminal work, S. Keshav notes that "the Holy Grail of computer networking is to design a network that has the flexibility and low cost of the internet, yet offers the end-to-end quality-of-services guarantees of the telephone network" [87]. This conceptual chapter offers a historical review and critical analysis of how D-QoS has been used in the past to deliver superior QoS on a finite network channel. As we observed in [25], the absence of such a historical and critical analysis in the technical literature has not helped the industry to understand why D-QoS has struggled for acceptance and for commercial success for decades.

5.1.1 Key questions answered

In this chapter, I set out to explore if and how D-QoS mechanisms can be used to provide QoS assurance to SCS. This task applies to both fixed and mobile networks but is more pertinent for the latter because of the limited capacity of the electromagnetic spectrum. Accordingly, I focus on exploring the types of D-QoS mechanisms and the lessons from over 40 years of their use, with a particular emphasis on how it applies to mobile networks. The analysis in this chapter will provide answers to the research question:

Research Goal 2: *Explore the history of D-QoS approaches to understand if and how it can be used to provide QoS guarantees for SCS systems*

> **RQ 2.1:** Why and how can D-QoS mechanisms deliver better QoS guarantees for 5G era SCS?
>
> **RQ 2.2:** Given its relatively unsuccessful commercial history and current net neutrality paradigm, how should D-QoS be positioned for success?

5.1.2 Chapter layout

The rest of this chapter is subdivided into four sections. Section 5.2 provides a historical overview of D-QoS mechanisms to identify their origins, rationale, promoters, benefits and challenges. In Section 5.3, I identify five different ways of classifying D-QoS mechanisms that can be used for SCS, whether for commercial, operational or regulatory reasons. To the best of my knowledge, this is the first work to provide such a comprehensive view on D-QoS classification. Section 5.4 explores how to predict the success of D-QoS mechanisms, drawing lessons from the history of D-QoS mechanisms and on 15 key questions which can help to predict success. Section 5.5 provides two practical considerations for implementing D-QoS, mainly with the introduction of a clarifying taxonomy and the CLASP framework as a decision framework for securing official backing for any D-QoS mechanism.

5.1.3 Publications linked to chapter

There are three publications that inform the analysis in this chapter. Section 5.2, 5.3 and 5.4 are based on [25] which has been published in IEEE Communications Standard magazine in 2020. For Section 5.5, this was informed by [156] which was presented at the IEEE 5G World Forum in 2018 and [75] which was presented at the IEEE / ELMAR International Symposium in 2019.

5.2 D-QoS for Reliability Assurance for SCS

5.2.1 Introducing D-QoS mechanisms

The remarkable success of data communication networks relies on the numerous networking mechanisms, approaches and standards developed by the technical community in the past 50 years. Generally, these have steadily improved the quality of 'best effort' connectivity, providing the requisite capacity and reliability for an increasingly sophisticated array of services. Occasionally, the networking community creates mechanisms, whether explicitly or implicitly, that offer the possibility to tailor the quality of the best effort connectivity for different users or use cases. In this work, I refer to this subset of mechanisms as D-QoS mechanisms. Tables 5.1 provides a chronological summary of notable D-QoS mechanisms from the 1970s till date.

D-QoS mechanisms have a clear purpose in the networking community as they propose to grant different levels of QoS to different services for commercial, regulatory or operational reasons. Hence why this PhD argues that they are vital for the society critical services that are increasingly reliant on connectivity. However, their history has been inexorably shaped by attempts to use D-QoS commercially to provide a premium service to users. This is because offering *different* QoS levels to *different* services is linked with concerns about network neutrality - the possibility of selling different Classes of Service (CoS) and the likelihood that non-premium users will receive a poorer experience in contravention of Net Neutrality expectations (where they exist). Accordingly, D-QoS often pitches Internet Service Providers (ISPs) against customers, content providers (CPs) and regulators.

In this work, I focus only on D-QoS mechanisms that have been productised by an ISP for users, whether by using shared public infrastructure or exclusive private infrastructure, and regardless of whether for commercial or regulatory reasons. This means that this work will not cover D-QoS mechanisms used by ISPs on their network for operational reasons or networks that are under a single or closed group administrative control. This includes Deterministic Networks (DetNet) and Time Sensitive Networking (TSN)[157], X-Ethernet, and the use of Multi Protocol Label Switching (MPLS) for Mobile Backhaul (as defined in BBF TR-221)[158] etc.

Table 5.1 Selected history of differentiated QoS Standards 1970s till date

Name	Description	Typical OSI layer & Connection type	Standards authority & year	Current market status & value	Main benefit	Main constraint
1. Leased Lines	Provides a virtual, constant connection between two end points	Layer 2; Connection-oriented, limited market	ITU; 1974	Moderate usage; moderate revenues (Ofcom UK)	Provides guaranteed connectivity with clear SLAs	Too expensive for most users
2. X.25	A global packet switched network with rigorous error correction mechanisms and based on the PSTN paradigm	Layer 2; Connection-oriented, limited market	ITU; 1976	Little evidence of current usage	Provides QoS and error free delivery over links of high error-rate	Slow and inefficient due to resource allocation mechanism and error correction
3. IP TOS[1]	The first conceptual provision for QoS on IP networks. Uses an 8-bit field to indicate delay, throughput and reliability	Layer 3; Connection-less, mass market	IETF; 1981	Little evidence of current usage	Conceptualised QoS for IP networks	No real need in early IP networks because of unreliable connections

4. ATM[2]	An approach to recreate PSTN guarantees on a data network by providing ’virtual’ circuits. Supports resource reservation and admission control with decent performance guarantees	Layer 2; Connection-oriented, limited market	ITU; 1988	Decreasingly used; negligible revenues	Provides QoS guarantees in addition to allowing resource sharing for both high throughput traffic and real time traffic	High, and increasingly irrelevant complexity leading to high cost
5. Frame Relay	Transmits data in variable-length ’frames’ while maintaining a virtual circuit between end nodes	Layer 2; Connection-oriented, limited market	ITU; 1990	Decreasingly used; negligible revenues	Provided QoS guarantees without the cost of dedicated/leased lines and without the complexity of X.25	Slower than ATM with a similar cost burden
6. IntServ /RSVP[3]	The first real attempt to implement QoS on IP networks. Reserves the private circuits akin to the telephone using RSVP	Layer 3; Connection-oriented, limited market	IETF; 1994	Little evidence of current usage	Proof of concept for QoS	Too complex and not scalable
7. Diffserv[4]	Proposed to overcome the scalability constraint of IntServ and builds on IP TOS	Layer 3; Connection-less, mass market	IETF; 1998	Little evidence of current usage	Good for real time traffic (e.g. VoIP)	Needs every node on the path to support

8. MPLS[5]	Provides a mechanism for setting up a connection-oriented path over different network types	Layer 2.5; Connection-oriented, limited market	IETF; 2001	Widely used; $46.3 billion by 2020 (Grandview Research)	Provides QoS guarantees over an IP network by setting up paths prior to sending packets through the network	Costly for many enterprises and the improvement in internet quality questions the relevance of MPLS' QoS guarantees
9. Carrier Ethernet	An extension of the LAN into a WAN. It provides a mechanism to circumvent bandwidth bottlenecks that can occur when a large number of small networks are connected to a single larger network	Layer 2; Connection-oriented, limited market	MEF; 2001	Widely used; $22.5 billion by 2020 (Ovum)	Integrates neatly into the LAN as an all-Ethernet network	No integral support for differential QoS. Limited scale compared to MPLS. As such, mostly used for connecting data centres
10. ABC[6]	Mobile operators to connect netbooks to the best of either mobile or Wi-Fi	Layer 1; Connection-oriented, mass market	3GPP; 2002	Abandoned after being overtaken by a Wi-Fi first user behavior	Management by mobile operators to provide seamless switching	High cost of 2G mobile data discouraged a mobile-first approach
11. QCI[7]	Based on the DiffServ model and is the first QoS model for a fully packet-based cellular network	Layer 3; Connection-less, mass market	3GPP; 2008	VoLTE use worth $34.8 billion by 2022 (AM Research)	Adapts the DiffServ QoS paradigm to cellular services	Needs every ISP on the path to support. Hence commercially unattractive

12. ACB[8]	Access Class Baring mechanism to grant priority to different SIM cards in a cellular network based on pre-arranged policies	Layer 2; Connection-oriented, mass market	3GPP; 2008	Widely used	Provides mechanism to grant access to only emergency services during an emergency	Binary decision to accept or reject a connection; designed primarily for voice
13. SD-WAN[9]	Moves the control of WAN networks into the cloud whilst providing high QoS at a low cost over the public internet	Layer 3; Connection-less, limited market	2014	Growing usage; $1.3 billion by 2021 (Gartner)	Provides the QoS guarantees of MPLS at a fraction of the cost	Unlikely to match the QoS guarantees of MPLS for some business critical services
14. Multi-operator Connectivity	MVNOs offer best connectivity using National Roaming	Layer 1; Connection-oriented, mass market	None; 2015	Growing interest, especially for IoT	Offers seamless switching to the best connectivity	Industry inertia as operators still compete on network quality
15. ACDC[10]	Builds on ACB and seeks to enable preferential access to different apps/services	Layer 2; Connection-oriented, mass market	3GPP; 2016	Not yet launched	Overcomes ACB's binary limitation by allowing prioritisation at services/app level	Commercially unproven
16. 5QI[11]	Builds on QCI and will be the de-facto QoS standard to support ultra reliable and low latency 5G services	Layer 3; Connection-less, mass market	3GPP; 2017	Not yet launched	Designed to support ultra reliable and low latency services	Commercially unproven

17. Network Slicing	Mechanism to create multiple logical/virtual networks on top of a single physical network	Layer 2; Connection-oriented, limited market	3GPP; 2017	Not yet launched	Designed to provide connection-oriented QoS guarantees over cellular data networks	Commercially unproven
18. Non-public 4G/5G Networks	Deploying a private 4G/5G infrastructure to be used exclusively by a customer	Layer 1; Connection-oriented, Limited	3GPP; 2017	Strong interest for enterprise 5G connections	Requires strong technical support	Over 290 deployments globally (ABI Research) August 2021.
19. Backup Broadband	4G/5G, satellite or fixed wireless backup to fixed broadband internet	Layer 1; Connection-oriented, mass market	None; 2020	Growing interest, esp since Covid-19 and working from home	Provides reliability in the event of a connectivity blackout	Still early days

1. IP ToS = Internet Protocol Type of Service
2. ATM = Asynchronous Transfer Mode
3. IntServ / RSVP = Integrated Services / ReSerVation Protocol
4. DiffServ = Differentiated Services
5. MPLS = Multi-Protocol Label Switching
6. ABC = Always Best Connected
7. QCI = QoS Class Identifier
8. ACB = Access Class Baring
9. SD-WAN = Software-defined Wide Area Network
10. ACDC = Application specific Congestion control for Data Communication
11. 5QI = 5G QoS Identifier

5.2.2 History of D-QoS mechanisms

The history of D-QoS mechanisms on data networks can be said to have begun in the 1970s when exclusive infrastructure began to be used to connect data communication terminals. Leased lines were used to connect mainframe computers with terminals and remote sites, offering better reliable connectivity than was obtainable on the telephone network while separate phone lines (including Integrated Services Digital Network (ISDN) lines) were used for fax machines and computing terminals. By 1976, X.25 standard was introduced to provide a global packet switched network by using software to create "virtual calls" that establish reliable circuits in a shared network. The success of these data communication products created the need for more scalable solutions. ATM and Frame Relay networks were introduced in the 1980s and their modern variants - MPLS and Carrier Ethernet - were introduced in the early 2000s. These were all commercially offered to the enterprise market to serve a limited number of customers, while contributing sizeable revenues to ISPs. For the mass market, IntServ and RSVP were introduced in the late 1990s to provide a reserved circuit for some services on the network. 5G Network Slicing, expected to debut in the market with the introduction of 5G SA, continues with the concept of a reserved circuit and will provide both isolation and prioritisation of selected traffic flows[159].

While the above examples used exclusive/special infrastructure (whether physical or virtual), an alternative approach is to focus on giving priority levels to different data packets. This took off with the standardisation of IP Type of Service (IP ToS), as part of TCP/IP, in 1981, and used different labels on the headers of data packets to determine the priority to be assigned to each packet. Although IP ToS was hardly implemented, it informed the development of Differentiated Services (DiffServ) in 1998 which in turn, led to QoS Class Identifier (QCI) on 4G networks in 2008 and 5G QoS Indicator (5QI) in 3GPP Release 15[159]. Beginning from 2014, SD-WAN, a new network architecture that brings flexibility of deployment is gaining grounds as a replacement for MPLS in enterprise networks.

5.2.3 Rationale for D-QoS mechanisms

From a technical perspective, the underlying rationale for D-QoS is to provide a ‘guaranteed’ quality, to some services or users on a network that has been designed to operate on a best effort principle [160]. Regulators can also ask for D-QoS for services or users that are considered important or critical for society (e.g. voice, public service broadcasting or emergency services) [156]. The commercial rationale for D-QoS is to be able to charge a premium for users who want a superior service. There have been accusations of primordial drivers for this, with some suggestions that it is partly borne of the irritation that Content

Providers (CP) are unduly profiting from an infrastructure that they have not paid for [12]. However, there is also a logical desire to add variability to the ISP's product as part of a pricing strategy to charge some customers a premium, increase the monetisation of the network or bring in new sets of customers [161]. The authors of RFC 2475 introducing Differentiated Services (DiffServ) in 1998 saw it this way: "Service differentiation is desired to accommodate heterogeneous application requirements and user expectations, and to permit differentiated pricing of Internet service."

5.3 Classifying D-QoS Mechanisms

D-QoS mechanisms vary widely in their design intention, technical design and implementation making it almost impossible in the technical literature to compare them. For example, ATM and Frame Relay were not designed with the intention of becoming D-QoS. Yet, the nature of their design and implementation means they are D-QoS enabled. Likewise, there is very little technical similarity between use of ATM and QCI. Likewise, 5G Network Slicing with its end-to-end logical separation concept is difficult to compare to most other D-QoS mechanisms. However, despite their technical dissimilarity, all D-QoS mechanisms are eerily similar in seeking to provide different QoS levels for uses whether for commercial, regulatory or operational reasons. Consequently, in this work, the focus is on the *'goals similarity'* of D-QoS mechanisms rather than their *'technical similarity'*. This approach blends the technical differences and commercial reality into a framework that is useful and can provide insights to both technical and commercial stakeholders. In this work, I recognize five possible ways of classifying D-QoS mechanisms:

5.3.1 Purpose of D-QoS mechanism

D-QoS mechanisms can be Category 1 (used for operational reasons or within exclusive infrastructure) and Category 2 (designed to be productised and commercialised to deliver superior quality or premium service in a shared infrastructure). For example, in 4G mobile networks, the Allocation and Retention Policy (ARP) mechanism is an example of Category 1 while QCI is an example of Category 2. In WiFi networks, several examples of Category 1 D-QoS mechanisms are described in [162] while Wireless Multimedia Extension (WME) is an example of a Category 2 standard that have been productised by several vendors (e.g. Aruba), companies (e.g. iPass) and venues (e.g. airports, hotels). In this work, only Category 2 D-QoS mechanisms are considered.

[12] http://www.washingtonpost.com/wp-dyn/content/article/2005/11/03/-AR2005110302211.html

5.3.2 Infrastructure used for D-QoS

Category 2 D-QoS mechanisms can be classified based on the type of infrastructure where they are used on, as summarised in Figure 5.1. This could be on a Single Shared Infrastructure (SSI), Multiple Shared Infrastructure (MSI) or Single Exclusive Infrastructure (SEI). For SSI, this requires that prioritisation is used to differentiate between services in other to deliver a superior outcome. Many of the common D-QoS mechanisms - e.g. ATM, Frame Relay, IP ToS, 4G QCI etc - are under this type. MSI exploits the engineering advantages of redundancy to improve outcomes. Examples of D-QoS mechanisms using MSI includes 'Always Best Connected' (ABC) which uses both Wi-Fi and mobile connectivity; Multi-Operator Connectivity (MoC) which is typically offered by MVNOs using universal SIM cards or dual/multi SIM cards (e.g. Google Fi, Transtel, Twilio); and Backup Broadband (e.g. BT Hybrid Connect, Vodafone Broadband Pro) which use 4G/5G connection as a backup to a fixed broadband service. SEI provides a physically separate infrastructure (e.g. Private 4G/5G networks, Leased Lines) to deliver superior outcomes or achieve security via isolation. All the three types of D-QoS mechanisms are considered in this work

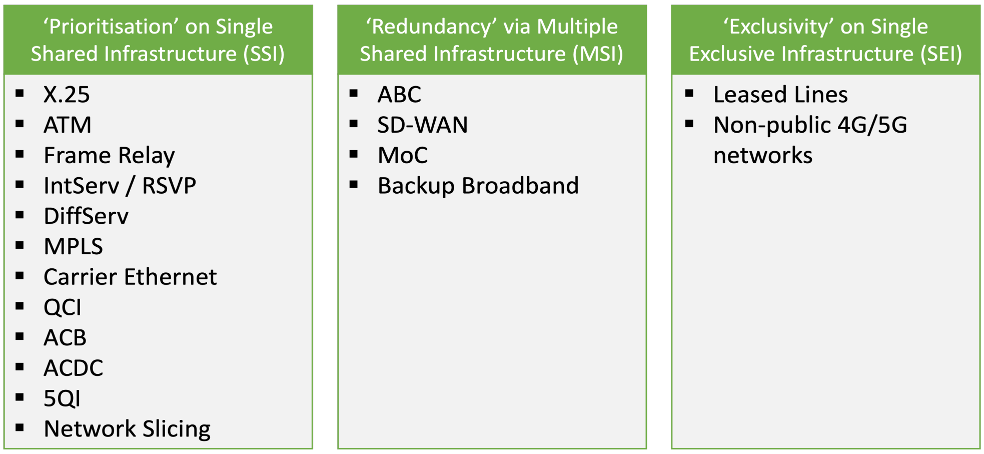


Fig. 5.1 Types of D-QoS mechanisms based on the infrastructure used. There have been many more historical examples of using prioritisation on Single Shared Infrastructure (SSI). However, increasingly, much of the industry activity is to implement D-QoS via redundancy using Multiple Shared Infrastructure (MSI) or Single Exclusive Infrastructure (SEI)

5.3.3 Packet handling by D-QoS mechanism

Regardless of whether the infrastructure is single vs multiple, or shared vs exclusive, D-QoS mechanisms can be grouped on how they handle data packets. This is done using prioritisation to treat some packet flows (Connection-oriented) or the individual packets (Connectionless) differently. Connection-oriented D-QoS mechanisms includes the use of leased lines and legacy efforts to create virtual circuits for ATM, Frame Relay, MPLS, the combined IntServ / RSVP, and 5G Network Slicing (because of its promise to provide different logical networks to different users based on a single 5G physical network). Connectionless D-QoS mechanisms build on the packet header concept introduced for IP ToS. It includes DiffServ, 4G QCI and 5G QoS Indicator (5QI). Both the first group and the second groups are considered in this work.

5.3.4 Applicable network layer for D-QoS mechanism

D-QoS mechanisms can be described based on which part of the network that they are targeted at. In [25], we use the Open System Interconnection (OSI) network model as a guide, and starting with the classification provided by X Xiao in [161], we categorise different D-QoS mechanisms into the layers where they are applicable. In the physical layer (Layer 1), a leased line provides an exclusive, physical connection between two end points. In the datalink layer (Layer 2), examples of D-QoS mechanisms that apply are X.25, ATM, Frame Relay, Carrier Ethernet and 5G Network Slicing. For the network layer (Layer 3), D-QoS mechanism examples include IP TOS, IntServ/RSVP, DiffServ, Software Defined Wide Area Networks (SD-WAN), QCI and 5QI. Some D-QoS mechanisms also apply at intermediate levels - e.g. MPLS, especially with the traffic engineering components (MPLS TE) is often regarded as an intermediate Layer 2.5 solution. For the transport layer (Layer 4), while several improvements over TCP and UDP - e.g. Stream Control Transmission Protocol (SCTP), QUIC, Recursive InterNetwork Architecture (RINO) - have been implemented, they are out of scope of this report as they have not been explicitly productised and commercialised as a premium D-QoS service over the ’best effort’ connectivity.

5.3.5 Target market for D-QoS mechanism

The choice of target market can provide a simple way to distinguish D-QoS mechanisms. Broadly, D-QoS mechanisms can be targeted at either the mass market (mostly consumers) or to a limited market (mostly enterprises). This distinction makes it possible to explore whether the different target markets will embrace the offering and how the D-QoS mechanism will perform commercially. Examples of D-QoS mechanisms that were designed and offered to the mass market include IntServ/RSVP, DiffServ, ABC, 4G QCI, 5G 5QI and Multi-Connectivity MVNOs (e.g Google Fi). Examples of D-QoS mechanisms for the limited enterprise market include ATM, Frame Relay, Carrier Ethernet, MPLS, SD-WAN and 5G Network Slicing.

5.4 Predicting the success of a D-QoS Mechanism for SCS

5.4.1 Lessons from the history of D-QoS mechanisms

The basic rationale for all D-QoS mechanisms is to provide superior internet experience to some users or to some traffic. Yet, despite this simplistic goal, there have been very few sustainable D-QoS success stories (e.g. ATM in the enterprise space) and many more failed attempts (e.g. IntServ). Some reasons for this have been documented in academic texts, industry articles and even political commentaries. [161] provides an academic exploration of the technical, commercial and regulatory challenges to D-QoS. From industry articles, writing in [163], Benjamin Teitelbaum and Stanislav Shalunov argued that premium IP service has not being commercially deployed and probably never will.

The political commentary against D-QoS is also insightful. For example, in his 7 February 2006 testimony to the US Senate Committee on Commerce, Science and Transportation hearing on Net Neutrality, Gary Bachula, the Vice President of Internet2 said: "For a number of years, we seriously explored various "quality of service" schemes, including having our engineers convene a Quality of Service Working Group. As it developed, though, all of our research and practical experience supported the conclusion that it was far more cost effective to simply provide more bandwidth. With enough bandwidth in the network, there is no congestion and video bits do not need preferential treatment."

While there is merit in these views, we argued in [25] that their observation applies mostly to D-QoS mechanisms that seeks to charge a premium price to a mass market audience. This is akin to the **'soft assurance'** described in [161] where an ISP charges a premium price for the the D-QoS mechanism on top of the price for its regular connectivity service. The alternative **'hard assurance'** D-QoS mechanisms [161] where the ISP provides predictable QoS *all the time* has proven to be more successful commercially. For example, Voice over LTE (VoLTE), which enjoys special status thanks to the regulation of voice on mobile networks, is on track to become a commercial success for the mass market. Likewise, D-QoS mechanisms that are targeted at a limited enterprise market - e.g. leased lines, ATM, carrier ethernet and MPLS - have been relatively successful commercially. These 'hard assurance' mechanisms also showcase how QoS can be sold as a 'feature' of a product and not as the product itself - in line with the recommendation of [161] to price QoS into the services whilst avoiding to sell QoS explicitly.

5.4.2 15 key questions on the success of a D-QoS mechanism for SCS

From the challenges identified in the literature, we identify and cluster 15 key questions that determine the relative success or failure of D-QoS mechanisms. These are questions that will need to be addressed before D-QoS can be applied to SCS. Using these questions, and based on an extensive review of D-QoS mechanisms and their history, the frameworks in Figure 5.2 provide a tool to predict the relative commercial success or failure of any new D-QoS mechanism.

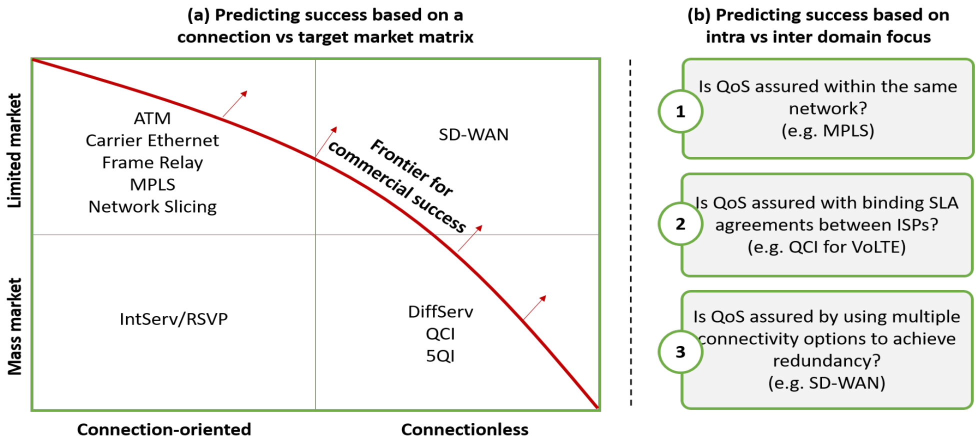


Fig. 5.2 The history of D-QoS mechanisms highlights how to predict success:
(a) Connection-oriented solutions that are targeted at the mass market are likely to fail
(b) Solutions that are delivered across ISP domains, and without a binding SLA, are likely to fail.

1. Are the technical standards ready?

This is the primary requirement and is necessary to avoid fragmentation in effort. It explains why standards bodies and industry groups are actively involved in creating technical standards for D-QoS mechanisms. While this is mostly a straightforward process, there are examples of squabbles in the standards-making process (e.g. disagreements and redrafting of the different DiffServ standards between 1999 and 2002 led to RFC2598, RFC3248 and RFC3246 [163]).

2. How simple is the implementation and operations?

Post standardisation, ISPs have to implement and operationalise any D-QoS mechanism that they intend to offer to customers. The actual implementation and its day-to-day operations can make or mar the commercial success of any D-QoS mechanism. Sometimes there are issues that emanate, but not yet addressed by the standards. Other times, it can be operational considerations. Examples of both include suitability of equipment (e.g. routers), error correction mechanisms, policing of ingress and egress, accounting of different traffic types, etc. X.25 offers an example of how operational realities can impact a D-QoS mechanism. Despite being one of the first D-QoS mechanisms to be launched, X.25 was eventually replaced because it was slow and inefficient due to its resource allocation mechanism and error correction procedures.

3. Does it align with the prevailing ethos of networking?

As the internet and TCP/IP have become the de-facto global standards for data networking, every D-QoS mechanism will be judged on how it conforms to the philosophical ethos of 'best effort' connectivity and its attributes of openness, robustness, flexibility, scalability and survivability. This ethos were imbibed in the early days of the internet and as [164] notes, the fundamental goal of the DARPA project that created the internet was to find an effective way to integrate existing networks, in a best effort way, such that there was survivability in the face of failure. [163] warns that if D-QoS mechanisms used for premium services were to become common, the best effort traffic will become degraded, radically overhauling the end-to-end design ethos for the internet. BEREC's guidelines for Net Neutrality in the European Union contain the 'necessity' and 'capacity' requirements, forcing ISPs to prove that any D-QoS mechanism is both necessary and will not degrade the quality of the best effort services.

4. Are Service Providers free to select users?

The question of how to select customers for D-QoS mechanisms is at the crux of most of the debate over Net Neutrality. As private companies, ISPs want a free hand to vary their service offering as they deem fit. But access to the internet is increasingly seen as a 'Public Good' and the quality of this service should be assured for all. As Figure 5.2 shows, there are two possible distinctions about this freedom to select users. Firstly, services which are offered to a limited user base (e.g. ATM or MPLS for enterprise customers) are generally permissible and successful compared to services which are aimed at the mass market (e.g. DiffServ)(see fig 5.2a). Secondly, in cases where regulators believe that it is beneficial to society to give preferential treatment to a service (e.g. voice), a mass market D-QoS mechanism (e.g. VoLTE on QCI Level 1) has a high probability of commercial success (see fig 5.2b).

5. Are Service Providers willing to offer service assurance?

While there are clear SLAs for D-QoS mechanisms designed for enterprise users, ISPs are generally unwilling to provide similar SLAs to the mass market user. As [161] observes, it is highly improbably to offer SLAs to the mass market because no one can guarantee such an SLA given the inevitability of network brownouts or catastrophic blackouts. As such, without a clear promise of the benefit, mass market users are reluctant to pay a premium for D-QoS.

6. Can customers verify the mechanism?

Humans generally like to test a product/service before buying to verify its conformance to their expectations. However, this is quite difficult for customers to do with most D-QoS mechanisms. [163] used the analogy of jiggling the locks on a door - to show that it works correctly - to compare with the absence of verification for D-QoS mechanisms. But service verification is not an insurmountable obstacle. For example, many people buy insurance services without testing and verifying because of price, peace of mind, convenience or it was legally mandatory. These are also similar reasons why enterprise users regularly buy D-QoS mechanisms such as ATM and MPLS.

7. Is the mechanism simple to sell?

Selling D-QoS mechanisms as a premium service to a mass market audience can pose a big dilemma for ISPs in two ways. First, it doubles the sales effort because this becomes an example of ‘double selling’ where ISPs have to first sell the basic network connectivity service and then sell the premium option on top. Second, from an ISP’s perspective, actively promoting its D-QoS offering is a tacit admittance that its regular service offering is poor. In a competitive market where ISPs compete on ‘network quality’, this can have consequences. ISPs circumvent this by exploiting a ruse where they can externalise the poor service level to someone else, or to a faceless entity such as the weather or the Internet. This can be seen in the marketing materials for MPLS, Carrier Ethernet or leased lines.

8. Does the mechanism depend on other Service Providers?

Given that the internet was designed to interconnect diverse, independently-managed networks, it will be a challenge to assure end-to-end premium QoS when the service relies on another ISP. In other words, an ISP should not promise specific QoS to its customers, and then expect a competing ISP to help it in fulfilling that promise without any compensation.

The approach to interconnection amongst ISPs is different for data and voice. For data, the typical interconnection agreements (e.g. a ‘Peering’ contract shown in [161]) shows that there is no binding QoS settlement in them. In contrast, for voice, the interconnection regime is underpinned by clear SLAs and a termination fee. This brings up the inter vs intra-domain criteria as a major reason why some D-QoS mechanisms are commercially successful. It holds that the probability of success is much higher when end-to-end QoS can be delivered within a single ISP’s domain (e.g. ATM, MPLS) or there is a binding interconnection agreement with enforceable QoS SLAs (e.g. QCI for VoLTE) (see fig 5.2b).

9. Can the mechanism be scaled easily?

The phenomenal growth of the internet has been largely supported by its core architectural feature where intelligence resides at the end nodes at the edge of the network while the equipment (e.g. routers) in the middle does only basic functions of forwarding packets. Any D-QoS mechanism that requires a change in this architecture to include intelligence on all the equipment in the core of the network, so that end-to-end QoS guarantees can be achieved, is unlikely to succeed. This is because it means that an ISP will need to invest *apriori* to upgrade its network even before it wins its first premium QoS customer. While this can be feasible for a limited market such as enterprises (e.g. MPLS), it cannot scale for a mass market (e.g. IntServ/RSVP) [165] (see fig 5.2a).

10. How simple is the business model?

In selling D-QoS mechanisms, ISPs need simple business models and clarity on who should pay for the premium service. If the business model is complex, ISPs will struggle to explain it to customers, and also struggle with additional complexity in their billing systems. Simplicity is also an issue between what is technically feasible and what is offered to customers. [163] observes that businesses that are built around service assurance do not strive to deliver 100% service reliability. Instead, they strive to offer an inferior ‘advertised service’ compared to the ‘engineered service’ so that they have some flexibility/buffer in their service delivery without breaching levels that trigger penalty. Likewise, deciding who will pay have often caused alarm and triggered Net Neutrality push-backs by businesses who assume that an ISP will ask them to pay. However, as the business model of Content Distribution Networks (e.g. AWS, Akamai) shows, it is possible to get businesses to pay for a D-QoS service offered to end users by making it a feature rather than the essence of the service.

11. Do customers understand the proposition?

It is generally cumbersome for ISPs to sell D-QoS mechanisms and for customers to understand them - cf. attempts to sell 'Always Best Connected' broadband in the early 2000s. This is in contrast to the clear historical evidence that customers prefer simplicity when buying communications service as is evident from the post 1830s Postal Service and post 1970s telephone service [166], post 2010 fixed broadband service and post 2015 mobile broadband service [13]. Generally, [166] notes that pricing for a service is based on either the Service Provider's desire to maximise utilisation and profits or a user's desire for value and simplicity. Over a long period of time, and across a number of industries, the trend is inexorably towards simplicity. In my engagements with industry stakeholders, a view that was often expressed for the low-use of the different QCI levels, apart from Level 1 for VoLTE, is that there is no customer appetite for them because customers do not understand the complexities of the differences.

12. Is the mechanism superior to its substitutes?

Selling D-QoS mechanisms as a premium service is hard because the D-QoS mechanism is competing against a progressively improving best effort connection. The reality of this can be seen in the mass market where customer preference has decidedly been in favour of the best effort connection. Even in enterprise markets where there is a stronger case for D-QoS, the trend is towards use of multiple redundant best effort connections. This can be seen in the growing preference of SD-WAN, which uses physical or virtual devices to create dedicated circuits on top of the best effort internet , compared to MPLS and leased lines by enterprise customers [14]. It also explains the emergence of universal MVNOs (e.g. Transtel, Twilio) which promise connectivity to multiple networks for connected cars.

13. Is the mechanism aimed at the mass market or a limited market?

The choice of addressable market for a D-QoS mechanism will largely influence its commercial success in the field/market. D-QoS services that are aimed at the mass market have a much higher threshold to meet to ensure widespread acceptance and adoption. Conversely, D-QoS mechanisms that are targeted at a limited, niche enterprise market have fared better in the market. For example, IntServ/RSVP and DiffServ were aimed at the mass market but failed woefully. In contrast, ATM, Frame Relay, Carrier Ethernet and SD-WAN which are aimed at the enterprise market have done relatively well commercially.

[13] https://www.linkedin.com/pulse/finnish-like-unlimited-mobile-data-model-now-europe-pal-zarandy/
[14] http://techblog.comsoc.org/2017/11/19/gartner-group-sd-wan-early-findings-yield-surprises/

14. Is the mechanism targeted to a tightly defined category?

While D-QoS mechanisms can be aimed at a mass market or a limited enterprise market, it is helpful to go a step further down to identify clear service categories where the D-QoS mechanism should apply. This requirement is well understood by the technical bodies, hence why 3GPP has identified twenty one 4G QCI and 5G 5QI values each which can be used to allocate appropriate quality of service levels to different services[159]. The expectation is that D-QoS mechanisms can then be targeted at specific service categories instead of the entire market. For example, QCI Level 1 is targeted at 'conversational voice' and hence have been applied to VoLTE on 4G networks.

15. Does the mechanism have regulatory support?

Given how the concerns about net neutrality have impacted D-QoS mechanisms, it should be expected that an endorsement from the government or regulator can be a game changer for the commercial prospects of the service. Most of the mass market D-QoS mechanisms that have struggled in the market (e.g. 4G QCI for services other than VoLTE) did not or do not have such official support, and have therefore struggled to gain wide acceptance. In contrast, the example of 4G QCI for VoLTE and its default acceptance by every mobile operator in the world illustrates the importance of official backing. This PhD proposes that many SCS services, especially CIS, ought to be given official recognition and backing to ensure their wide deployment and adoption across society.

5.5 Practical Considerations for the use of D-QoS for SCS in 5G era networks

5.5.1 Taxonomy for identifying SCS use cases for D-QoS

Key question 14 has identified the need for a tightly defined service category to improve the chances of success for D-QoS. In the 3GPP TS 23.501 reference document[159], 3GPP proposed 21 5QI values to define all the service categories that should be accommodated in the mobile network. While these technical provisions are a good start, I argue that they are not enough because 5QI and its QCI predecessor do not provide a holistic approach that will encompass all SCS services. Therefore, in this subsection, I introduce a holistic taxonomy for the 5G era - which we presented in [156] - and show how it augments the existing 5QI. The taxonomy incorporates technical and non-technical considerations to provide a structure for definitions, designs, policymaking and implementations for D-QoS for SCS. While the taxonomy is mostly for services to be prioritised, it also provides guidance for services that require multi-connectivity or exclusive connectivity.

The taxonomy recognizes seven possible service categories that could be differentiated in the 5G era (as summarised in Table 5.2) and assumes the following:

- Preferential treatment of SCS is expected to deliver improved QoS. However, D-QoS can also be used to deliver inferior QoS outcomes for services where their traffic flow does not need the default or neutral QoS (similar to the rationale behind IETF protocols such as LEDBAT (RFC 6817 [167])).
- Preferential treatment by an ISP can be incentive-driven (e.g. to earn more revenues, to improve capacity utilisation, to manage congestion etc), obligated (e.g. to provide QoS assurance) or even altruistic (e.g. to support corporate social responsibility).
- Policymakers can mandate preferential treatment for SCS similar to how they mandate preferential treatment for other non-telecoms services (e.g. Public Service Broadcasting such as BBC).
- Preferential treatment for SCS that is sanctioned by policymakers will more likely apply industry-wide and across many countries unlike those that are special arrangements with selected ISPs.

1. Statutory Services

These are the core services of the mobile industry and these are statutorily prioritised to comply with regulatory requirements. Currently, this mostly applies to communication services such as VoLTE. 3GPP recognises the need to prioritise these using 5QI Value 1 or QCI Level 1, with a guaranteed bit rate (GBR), Packet Delay Budget (PDB) of 100ms and Packet Error Loss Rate (PELR) of 10^{-2}. If it eventually becomes a mainstream service, Video over LTE (ViLTE) will also likely be prioritised although it is still uncertain if this will be based on GBR 5QI Value 2 / QCI Level 2 or non-GBR 5QI 8 / QCI 8. As digital services become integral to society, it is necessary for other services to be recognised as statutory services too. For example, services that come under non-BIAS (US) and specialised services (Europe) classes ought to qualify as statutory services that deserve prioritisation. Already, some of these have either being defined by 3GPP (e.g. Vehicle-2-X on 5QI/QCI 75/79) or being proposed by researchers (e.g. to support e-Health services [168]). Sadly, despite these efforts, there is little evidence of such prioritisation in practice.

Table 5.2 Seven Differentiated service classes for mobile networks. It shows that generally, SCS are currently not adequately captured under the QCI or 5QI framework because they are ill defined.

Service Class	**Description**	**Policy Position**	**5QI Value**	**Examples**
Statutory Services	Mostly voice and messaging services over IP, perhaps some Ultra Low Latency Low Reliable services	Promoted	1, 2, 5, 7, 50, 55	VoLTE, ViLTE, Public Service live TV
Society Critical Services	Data services that are critical to digital society	-	-	112/911/999 equivalent data services
Best Effort Services	Default option for most internet content	Accepted	8	Routine web browsing, email
Commercially-Preferred Services	Services that are allowed to be prioritised for commercial reasons	Discouraged	1, 2, 3, 4, 6, 7, 8, 9, 70, 75, 79, B, C, D, G	Live IPTV
Discounted Services	Services that are offered at a reduced price or quality	Scrutinized	-	Zero rated content and truncated streaming services
Delayed Services	Non-urgent services that can be delayed to periods of low network utilisation. Compare with WiFi API	-	-	IoT traffic, Software downloads / updates, catch-up TV, cloud backup
Blocked Services	Services that barred for security or regulatory reasons, or selected by customers to block	Unclarified	-	DDoS attacks Parental Consent, illegal bitcoin mining

2. Society Critical Services

These services are the focus of this PhD as I have demonstrated that they require special treatment to assure reliability. However, generally, SCS are currently not adequately captured under the QCI or 5QI framework because they are ill defined. When SCS is recognised as a specific class of service that requires special treatment to achieve reliability assurance, this special treatment can be delivered via prioritisation, multi-connectivity or exclusive connectivity. For example, on prioritisation, 3GPP has taken some steps to define several QCI levels: e.g. QCI 75 and 79. 5QI includes many more provisions for remote control, intelligent systems, mission critical data, and low latency data services. But these are not enough because beyond clear cases such as driverless cars, it may be in the best interest of society to give priority to several other services that are integral to the smooth functioning of the digital society (e.g. live arrival / departure schedules for train stations and airports, or electronic voting).

3. 'Best effort' Services

This is the default service class today for most internet services and these services are treated in an undifferentiated approach. On mobile networks, these services are typically based on QCI Level 9 (with non-GBR, PDB of 300ms and PELR of 10^{-6}) and by supposition, it is the only class that Net Neutrality seeks to maintain. The corresponding value for 5QI will be Value 8. The true meaning of best effort is hardly ever stated though. Although ISPs increasingly advertise internet speeds in Mbps, it is rare to see any ISP define what best effort actually means to customers in terms of both speed (i.e. throughput/ bandwidth) and delay (i.e. latency). 'Best effort' class of services will remain the default service class in the 5G era with no guarantees for service quality.

4. Commercially-Preferred Services

Where and when Net Neutrality principles permit, there exist services which mobile operators would want to prioritize for commercial reasons (e.g. IPTV, video services or 5G network slices [169]). However, there are no clear rules or expectations on what services can be added to this class and so operators are reluctant to explore this option for fear of falling foul of Net Neutrality. This service class is a call for mobile operators and regulators to clearly, and jointly, identify the services where net neutrality would not apply. Unless this is done, I argue that commercially-driven D-QoS will struggle to gain public acceptance because of the assumption that it violates net neutrality. Given that nearly all of the QCI and 5QI levels have been defined for services that would come under this category, perhaps this is why much of the QCI framework is unused and raises the prospect that 5QI will also go unused.

5. Discounted Services

Discounted services can be treated differently on price, quality or both and is managed from a policy/charging perspective. Examples of discounted services are Zero rating, a form of SDP that became popular with Facebook's Internet.org initiative [170], and T-Mobile's BingeOn which offered customers a zero rated service, albeit at a truncated speed, to selected streaming services [171]. Expectedly, Zero rating has attracted a lot of press scrutiny and regulatory attention over concerns that it violates Net Neutrality. So far, it seems to have survived [170]. In the 5G era, we will expect operators to continue experimenting with zero rating as they seek to manage demand on their networks.

6. Delayed Services

There are several non-urgent services that can be deprioritised or deferred to a later time to reduce congestion on the network. Examples include non-urgent software updates and cloud backups. On software updates, while the mobile industry failed to act on this, the device manufacturers stepped in to offer such categorisation between services that rely on Wi-Fi and those that rely on mobile connectivity. Currently, most iOS/Android apps are set to automatically update when the user is on Wi-Fi, a feature that is quite common in developed markets where Wi-Fi access is taken for granted. Yet, there is no commensurate mechanism to provide this capability on the mobile network, especially for customers without access to Wi-Fi in developing markets. On cloud backups, there may be a need to delay many of the automatic cloud backup and syncing during periods of resource constraints (e.g. in crowded events). For example, [67] observed that the ratio of downlink to uplink traffic during the Superbowl changed from the typical 9:1 to 1:1 due to background syncing of pictures to the cloud - this syncing could have been delayed with minimal impact. Notably, neither QCI nor 5QI has made any provision for delayed services, hence all such services are treated similarly as best effort (5QI 8/9 or QCI 9). In practice, operators will need to decide on a QCI level for software updates and cloud backups, and then develop an appropriate handshake mechanism that can be used by such services. The Network Information API from W3C on how apps can use Wi-Fi is an example [102].

7. Blocked Services

There are occasions when it may be necessary to lawfully block some services for commercial, regulatory or customer-choice reasons. These could be content that is malicious (e.g. AdGuard reported in Oct 2017 that over 220 websites are using the computers of their over 500 million visitors to illegally mine crypto-currencies[15]), causes nuisance (e.g. using Cookies) or imposes a security/privacy threat (e.g. age-related services blocked under parental consent) or as demanded by regulators based on their own criteria. Such blocking will likely get more frequent given the growing risk of hacking, security breaches and politicisation of the internet. Besides, given growing concerns about social media addiction [172], the ability for customers to request to block access to selective social media content may become part of the toolkit of treatment therapies. While most mobile operators have channels through which customers can report malicious websites and content, these can be broadened so that customers can decide which services to selectively block. Examples of such customer choice abound in practice (e.g. use of ad-blocking software, spam-reporting for emails, decision on tracking cookies etc).

5.5.2 CLASP: framework for deciding 'official' SCS use cases

Key question 15 has recognized the need for official backing for D-QoS to be successful. For SCS use cases that require either prioritisation on single shared infrastructure or redundancy with multiple shared infrastructure, it is necessary to establish a criteria for determining which services require official backing for D-QoS. Accordingly, in [75], we proposed the CLASP (Critical, Localized, Authorized, Specific, Perishable) framework as a guide on how to determine whether to give official backing to any SCS service or not to be supported by D-QoS. Table 5.3 summarises the elements of the CLASP framework.

Our observation is that current official decision making is often non-existent, ambiguous, opaque or each request is dealt within a bespoke manner. In cases where regulators have discussed D-QoS, they have mostly focused on commercially-exploitable mechanisms. For example, in both the US and in the European Union, regulators may have inadvertently authorised 'fast lanes' for some commercial services, with no clarity or procedure on how the prioritisation should be implemented, and often, with only an *ex post* regulatory review by authorities to prove that no harm was done.

CLASP acknowledges that there are SCS services from commercial providers which are also critical for the optimal functioning of a digital society. However, in the spirit of Net Neutrality, and to avoid the possibility of 'regulatory capture' (i.e. when private entities use

[15] https://blog.adguard.com/en/crypto-mining-fever/

Table 5.3 CLASP Framework: how to decide on ‘official’ backing for DQoS on SCS

CLASP	**Description**	**Example 1 (www.tfl.gov.uk)**	**Example 2 (www.google.co.uk)**
Critical	Generally accepted to be critical to the proper functioning of the digital society of the future	Yes - for the >8.5 million London residents who rely on it daily	Yes - most visited website in the UK and commands 87% of the UK search market
Localised	Only prioritised within locality where relevant, with the consent of the people of the region	Yes - applies only in London and environs	Yes - applies to all of the UK market
Authorised	Considered to be important, and authorized by a mandated government agency (e.g. Ofcom)	Yes - managed by a government agency and provides information that is already prioritised on radio and TV	No - a commercial service with no official mandate. Can be replaced by several competing commercial alternatives
Specific	Only specific functionalities of the service can be prioritised and not all services from a particular provider	Yes - can be restricted only to the "TFL Journey Planner" section	Yes - can be restricted to only the "search" function
Perishable	Offers timely information which expires in short durations	Yes - provides timely information for all modes of transport in London	Yes - provides timely information for users on their particular search query
-	Prioritised?	Yes	No

their overwhelming resources or clout to bend regulation to their whims), CLASP draws a distinction between services that are authorized as critical versus those that have evolved to become critical (e.g. based on market share).

5.6 Chapter Summary

This chapter has provided an extensive qualitative insight into the nature of D-QoS and how it can be used to improve QoS assurance for SCS use cases. I first started with a historical overview of D-QoS mechanisms, to identify their origins, rationale, promoters, benefits and challenges. Secondly, I identified five ways of classifying D-QoS mechanisms, becoming the first work (to the best of my knowledge) to provide a comprehensive view on how D-QoS mechanisms can be classified. Thirdly, I focused on how to predict success for a D-QoS mechanism, drawing lessons from the history of D-QoS mechanisms and on 15 key questions which can help to predict the success of any D-QoS mechanism. Finally, I shifted to the practical considerations for the use of D-QoS for SCS in the 5G era, introducing a taxonomy to help categorise different D-QoS classes of service and the CLASP framework for helping to determine which D-QoS mechanisms deserve official backing.

Specifically, the analysis presented in this Chapter provide strong insights to RQ 2. I conclude that the ***answer to RQ 2.1 is that D-QoS mechanisms can be used to provide QoS guarantees for SCS in scenarios where the SCS use case is clearly identified, aimed at a limited - perhaps enterprise - target market and there is official backing for D-QoS to be used.*** This response also address RQ 2.2 because the lessons from the history of D-QoS corroborates this.

Having established the implication of mobile network performance brownouts in Chapter 4 and how D-QoS mechanisms can be used to ameliorate this in Chapter 5, I will focus in the next Chapter (Chapter 6) on how to empirically evaluate a type of D-QoS mechanism to determine its usability for an SCS service. This evaluation will be a system implementation of a multi-operator redundant connectivity option for a Connected Car (CC) on live UK mobile networks.

Chapter 6

Evaluating D-QoS for Connected Cars in the Field

Nḍụ mmiri, nḍụ azụ; mmiri atala ma azụ anwụla (Igbo Proverb)

(Life to the water, life to the fish; may the waters not dry and may fishes not die)

6.1 Introduction

The theoretical conceptualisation of D-QoS is pretty well established and a lot of effort has gone into developing several D-QoS mechanisms in the industry[25]. However, given their relative commercial under-performance, it is pertinent to ask if there is clarity as to the commercial benefits, the challenges in their implementation and the impact on the market status quo when a D-QoS mechanism is widely implemented. In this chapter, I set out to evaluate the use of multi-operator connectivity (MoC) for Connected Cars (CC). This is an example of an MSI as multiple CCs are using the same shared public infrastructure for network connectivity. CCs are a good example of SCS as their vehicle-to-everything (V2X) use cases require stringent reliability for safety and non-safety uses. Analyses Mason notes that 164 million passenger cars were connected by the end of 2018 (about 16.5% of the total) and forecast this will rise to 831 million by 2027 (about 56% of total)[173] as CCs use connectivity to communicate with other vehicles, the transport infrastructure, pedestrians or with the internet. This chapter covers a full system implementation of an MoC setup and a road driving test on >800 kilometers of major and minor roads in South East England. In particular, it explores the comparative advantages and differences of managing the MoC at the user device by the user - i.e. demand side managed (DSM-MoC) - vs managed at the network by the network provider - i.e. supply side managed (SSM-MoC).

6.1.1 Key questions answered

In this chapter, I report on our work to empirically evaluate an MSI D-QoS mechanism to understand how it will perform in practice. The work is focused on understanding how redundancy, provided by a multi-operator connectivity approach, can be used to provide QoS assurance for CCs. The chapter will provide answers to the research question:

Research Goal 3: *Evaluate the reality of D-QoS approaches to understand their relative advantages in providing QoS guarantees for SCS*

In particular, the work will address:

> **RQ 3.1:** If D-QoS is a feasible approach, what are options and implementation approach to using D-QoS to provide QoS guarantees to SCS?
>
> **RQ 3.2:** What are the quantifiable benefits of using any of the D-QoS approaches on public mobile networks for selected use cases?

6.1.2 Chapter layout

In this chapter, I will use the 'scientific writing format' to structure and subdivide the work into seven sections. Section 6.2 introduces the aim, motivation and background for the research while Section 6.3 develops the system model for the analysis. In Section 6.4, I detail the methodology for the research. This covers the software and hardware implementation, plus the routes, dates/times and logistical approach for the research. Section 6.5 presents the results from the work, Section 6.6 discusses the implications of the findings from the research and Section 6.7 summarises the key contributions of the research. Finally, Section 6.8 draws conclusions from the work to support the thesis on the use of D-QoS for SCS.

6.1.3 Publications linked to chapter

There are two publications that inform the analysis in this chapter. The first is [174] which quantified the hypothetical benefits of MoC via extensive field measurements on live networks in the UK. This was presented at the 2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS). The second is [113] which builds on the earlier result to investigate actual benefits of MoC on live UK networks and was published in the September 2022 edition of IEEE Transactions on Network and Services Management.

6.2 Setting the Context

6.2.1 Aim

The aim of this work is to investigate a suitable means of improving the reliability for CC connectivity. CCs are increasingly mainstream across society and are clear examples of an SCS use case as they increasingly rely on connectivity for important safety actions or for better management of the transportation network. CCs use connectivity to communicate with other vehicles, the transport infrastructure, pedestrians or to access content from the internet. In this work, CC covers all variations of connectivity for vehicles, regardless of whether they are for safety or non-safety reasons (e.g. infotainment, navigation and control), and encompassing 'driverless / autonomous' road vehicles and basic infotainment in road vehicles. In general, there are four groups of CC V2X use cases: Vehicle-to-Vehicle (V2V), Vehicle-to-Person (V2P), Vehicle-to-Infrastructure (V2I) and Vehicle-to-Network (V2N)[121]. Broadly, there are two competing CC connectivity technologies: the IEEE802.11p based V2X version - also known as Dedicated Short Range Communication (DSRC) - and Cellular V2X (cV2X). However, academic and commercial commentary suggests that the ecosystem is coalescing on cV2X [175], deepening a technology trend that has been happening with 2G, 3G, 4G and now 5G[2].

6.2.2 Motivation

The primary motivation for this work is to investigate a suitable means of improving the reliability for CC connectivity. Our work identified five approaches to doing this for cV2X: (i) the use of Sidelink, introduced in 3GPP Release 16, for near field communications[128], (ii) improvements in the underlying reliability of the network to support V2N use cases (e.g. moving from 4G to 5G)[13], (iii) adoption of prioritisation for different service classes (e.g. via 5G Network Slicing)[73], (iv) use of bespoke/private networks on the roads[129] and (v) via multi-operator connectivity or national roaming agreements on public cellular networks[131]. However, the focus of [113] is on (v) and compares three options for providing multi-operator CC connectivity (Figure 6.1).

Option 1 (Figure 6.1a) is the default where CCs are connected via either an embedded or an aftermarket module, to a single public mobile network. Crucially too, there is an exclusive contractual relationship between the network provider and the car manufacturer or user. Option 2 (Figure 6.1b) exists so that users who demand better reliability, or where regulation compels it (e.g. e-Call in Europe [120]), can fall back on another network. Option 2 is a supply-side managed multi-operator connectivity (SSM-MoC) option where the CC

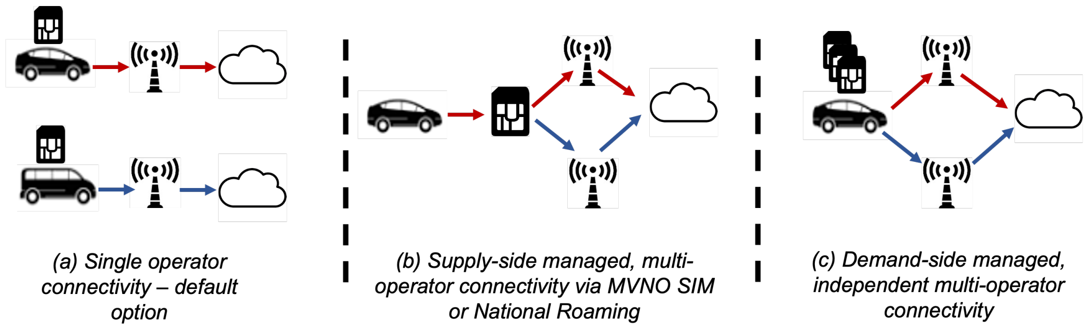


Fig. 6.1 Vehicle-to-Network (V2N) connectivity options for Connected Cars (CCs). While most existing CC contractual relationships with cellular network operators are based on Option (a), some MVNOs are stepping in to offer Option (b). But would Option (b) provide better reliability and QoS assurance than the hardly used Option (c)?

connects to a single service provider who then manages the actual network connectivity in the backend. In practice, the service provider in Option 2 is either a network operator who relies on 'national roaming' (i.e permitted to use other networks[176]) or a mobile virtual network operator (MVNO) who relies on wholesale deals with multiple operators. Regardless, the decision on which network to use in Option 2 is still made far away from the user, by an entity whose priorities may not always align with that of the user. This is the justification for Option 3 (Figure 6.1c), where the connectivity decision is repatriated to be managed *in situ* as a demand-side managed multi-operator connectivity (DSM-MoC) setup.

Our work is about 'reliability arbitrage' in a system-implementation on a CC, compared to other works on multi-connectivity (e.g. Google Fi, iCellular) which are mostly aimed at 'price & quality arbitrage' on smartphones[177, 24, 112]. By using a system-level implementation with onboard computers (i.e Raspberry Pi), networking equipment and all the inherent unknown factors in a practical, real-life CC scenario, this approach demonstrates the advantage of MoC in a real-life and real-time scenario.

6.2.3 Background

The background and literature review for our CC experiment is presented in Section 2.6. In summary, the opportunity to use redundant MoC is informed by softwarisation of the mobile network value chain and the unique characteristics of CCs. Increasing softwarisation of the cellular infrastructure is opening up previously impossible usage scenarios as it enables the integration of intelligence at the device, edge and core into a composite framework for improved operations & management of networks. Likewise, CCs have unique characteristics that make them different from smartphones or other connected devices. Unlike smartphones, and given their mobility, CCs endure significant geographical variation in connectivity performance as they move from one base station to the other. [126] notes that this reality means that, compared to smartphones, there is only a small window to deliver large volumes of data, making it imperative to assure reliability in that small window.

In practice, redundant MoC can be managed at the supply side (SSM-MoC) or the demand side (DSM-MoC). SSM-MoC is a continuation of the 'intelligent core, dumb node' philosophy in the operations of cellular networks, while DSM-MoC assumes that each network node is imbued with intelligent decision-making capabilities to optimize network reliability. Our work builds on this softwarised and intelligent edge/device philosophy to explore how CCs, with huge computational power, can optimise their connectivity, either via SSM or DSM.

6.3 System Model

In this section, I report on the system model that provides the theoretical basis for the work on MoC in [113]. Given a CC connectivity system, with several connectivity options, we explore which multi-operator connectivity option will experimentally deliver superior overall system reliability. This problem statement can be investigated theoretically from two dimensions: first, the acceptable threshold required for each reliability metric; second, the number of MoC options needed to optimise system reliability.

6.3.1 Threshold for reliability parameters

For any given system, there is an instantaneous Network Reliability R_{t} and an instantaneous Performance-constrained Reliability function Q_{t}. R_{t} is based on the system's availability, while Q_{t} is the reliability when the overall system performance meets a pre-determined threshold. In other words, R_{t} represents the reliability when the system is available while Q_{t} represents the reliability when the system is available and meets the required performance threshold. Therefore,

$$Q_{\mathrm{t}} = W_{\mathrm{t}} R_{\mathrm{t}} \tag{6.1}$$

and W_{t} is an integer binary function [0, 1].

There are both endogenous and exogenous levers to optimize R_{t} for a mobile connectivity system. For the former, upgrading from 3G to 4G to 5G is the most critical endogenous lever to improve intrinsic reliability. Beyond that, exogenous levers become important to manage customer demand, and mitigate sub-optimal network outcomes. Here, we focus on the two main exogenous levers of probability of network congestion (Y_{t}) and the probability of network connectivity availability (Z_{t}.)

If the instantaneous cumulative failure rate α_{t} of the system is

$$\alpha_t = \int_0^\infty (Y_{\mathrm{t}} + Z_{\mathrm{t}})dt \tag{6.2}$$

then the instantaneous reliability $\mathrm{R_t}$ is

$$R_{\mathrm{t}} = 1 - \alpha_t = 1 - \int_0^\infty (Y_{\mathrm{t}} + Z_{\mathrm{t}})dt \tag{6.3}$$

Y_{t} and Z_{t} can be explored using mathematical modeling approaches. The probability of network congestion Y_{t}, can be modelled using the Hazard Rate model as a form of Survival Analysis to approximate the probability that a particular connection is congested. The probability density function of the Hazard Rate model is:

$$Y_{\mathrm{t}} = t\phi e^{\text{-p}} \tag{6.4}$$

where t is the instantaneous time during operations, ϕ represents the >0 hazard rate per inherent failure of the system, and p (-1 < p < 1) is the changing rate of the network traffic density.

The probability of network connectivity availability Z_{t} can be modelled using the Weibull distribution to approximate the probability that a particular network connection is available. The probability density function of the Weibull distribution is given by:

$$Z_t = \frac{\beta}{\eta}\left(\frac{t}{\beta}\right)^{\beta-1} e^{-(\frac{t}{\eta})^\beta} \tag{6.5}$$

where β and η are the shape and scale parameters respectively and are > 0.

The Hazard Rate and Weibull Distribution probability density functions provide a theoretical framework to investigate the failure rate (α_t) and reliability (R_{t}) of each individual network connection for CC.

6.3.2 Number of multi-operator connectivity options needed

In a CC system with n parallel redundant networks, the overall system reliability can be approximated as a composite function of the Hazard Rate and Weibull Distribution of each individual network. If each of the n parallel redundant networks has a uniform failure rate α_{t}, then:

$$R_t = 1 - \left(1 - e^{-\alpha_t}\right)^n \tag{6.6}$$

and the Mean Time To Fail (MTTF) μ is

$$MTTF = \mu = \int_0^\infty R_t dt \tag{6.7}$$

In the UK with four mobile networks, each with same failure rate α_t, the instantaneous R_t are:

$$\begin{aligned} n_1 &= e^{-\alpha} \\ n_2 &= 2e^{-\alpha} - e^{-2\alpha} \\ n_3 &= 3e^{-\alpha} - 3e^{-2\alpha} + e^{-3\alpha} \\ n_4 &= 4e^{-\alpha} - 6e^{-2\alpha} + 4e^{-3\alpha} - + e^{-4\alpha} \end{aligned} \tag{6.8}$$

and the MTTF μ_t are

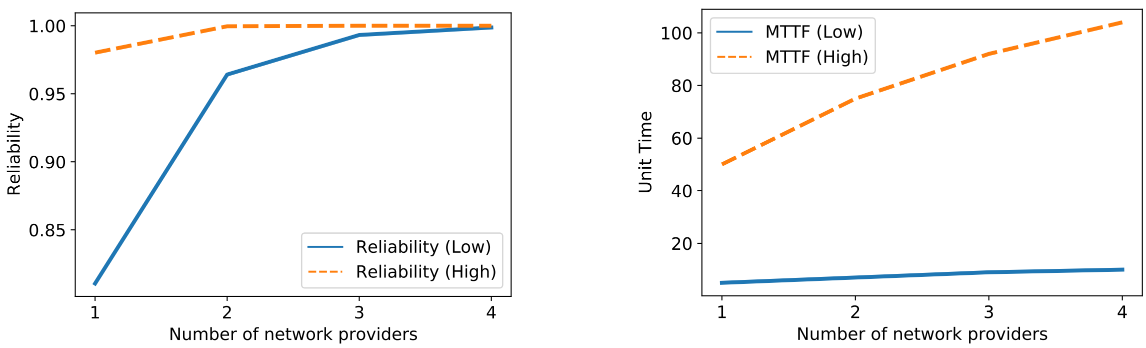
$$\begin{aligned} n_1 &= \frac{1}{\alpha} \\ n_2 &= \frac{3}{2\alpha} \\ n_3 &= \frac{11}{6\alpha} \\ n_4 &= \frac{25}{12\alpha} \end{aligned} \tag{6.9}$$

Next, we use real life data for UK mobile networks to map into the theoretical analysis results. In its Connected Nations 2020 report[139], UK telecoms regulator, Ofcom, reports that UK operators provide outdoor coverage to 98%-99% of premises and their networks' coverage of the UK landmass ranges from around 79% to around 85%. We use 98% as the upper limit and 79% as the lower limit for the expected availability (i.e unavailability / failure rate: 2% - 21%). Figure 6.2 summarises the aggregate R_t and μ_t expected for a road test across South East England. The field measurement is designed to confirm and validate these.

6.4 Experiment Design, Parameters & Setup

6.4.1 Experiment goals & design

We start off with a hypothesis that a redundant connectivity setup will deliver higher values for R_t and Q_t. Overall, we had four key experiments & design goals to proof the hypothesis:



(a) Reliability

(b) MTTF

Fig. 6.2 Theoretical evaluation of expected Reliability and Mean Time to Fail based on 4G penetration levels from [139]

1. Experiment 1: Designed to quantify the value of the hypothetical advantage of using a redundant connectivity setup. This investigates the hypothesis in Chapter 4 that individual 3G, 4G and early 5G networks are unable to meet the stringent performance requirements for CCs. For this, we measured 4G performance for all four UK providers on a CC, at the same time, location and using the same type of device.

2. Experiment 2: Designed to determine the best setup for redundant connectivity options (i.e. SSM-MoC vs DSM-MoC). This explores the practicalities of designing redundant systems for a given engineering problem. For this, we measured 4G performance on a global Universal SIM provider in the UK and compared it with the performance of the four UK providers from Experiment 1.

3. Experiment 3: Set up an actual DMS-MoC implementation to validate the assumption that DSM-MoC is hypothetically better. This explores the implementability of redundant systems and investigates the impact of network congestion and failures on DSM-MoC. For this, we built a system implementation of DSM-MoC as shown in Figure 6.3 (the schematic diagram) and Figure 6.4 (the actual picture).

4. Experiment 4: Designed to investigate suitability of different types of applications for DSM-MoC and compare performance in stationary and high-speed scenarios. Given the differences between smartphones and CCs, this explores to what extent the cellular connectivity can serve CCs in different usage scenarios. For this, we used our system implementation to explore behavior of TCP & UDP for DSM-MoC for different TCP/UDP packet sizes.

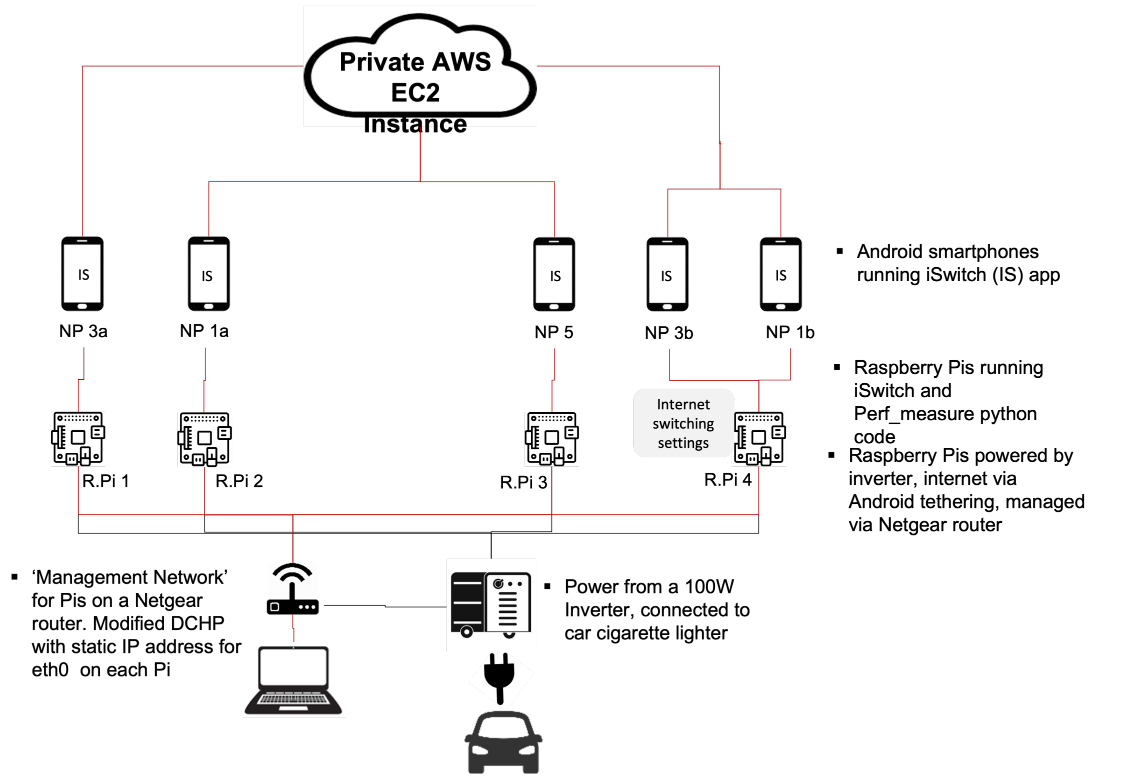


Fig. 6.3 Schematic layout of equipment setup for field measurements on Day 2

6.4.2 Experiment route and dates

Experiments 1-4 were conducted on major and minor roads, and at Stevenage, a suburban location in South East England. This was done on four dates that were selected based on convenience - each of the experiments was self contained and there was no need for cross-day comparisons. Our road drive was on a stretch of roads in South East UK totaling 237 miles (380km) twice on 15 November 2020 (Day 1) and 14 June 2021 (Day 2), and then 28 miles (44.8km) on 15 November 2021 (Day 3). The stationary measurements were at Stevenage, on 16 November 2021 (Day 4). Overall, and including other exploratory drives for testing, we drove for over 1,000k for this work. The entire route in shown in Figure 6.5.

The drives on Day 1 & 2 included most of the length of the M25, the 120 miles (192km) ring road around London - the busiest motorway in the UK - and another 117 miles (188km) drive through rural roads in Hertfordshire and Bedfordshire. Day 3 was on the A1(M), another busy motorway into London. The justification for measuring on both major (motorways) and minor (rural) roads is to establish the connectivity experience for a CC on all road types. This is necessary because any expectation that the future will be about pervasive connectivity for CCs will need to consider major roads (assumed to be adequately covered) and minor rural roads with low traffic density. This is summarised in Table 6.1.

Fig. 6.4 Equipment setup for field measurements on Day 2: top view layout of equipment - laptop, inverter, five Android devices and a box with four Raspberry Pis and two netgear routers

Table 6.1 Multi Operator Connectivity for Connected Cars - Summary of Experiments

Experiment	**Date**	**Route**
Experiment 1	Day 1 (15 Nov 2020)	Done along 237 miles of major and minor roads in South East England
Experiment 2	Day 1 (15 Nov 2020)	Same route as Experiment 1
Experiment 3	Day 2 (14 Jun 2021)	Same route as Experiment 1 & 2
Experiment 4	Day 3 (15 Nov 2021) & Day 4 (16 Nov 2021)	Done along 28 miles of A1(M) motorway on Day 3 and at a stationary location in Stevenage on Day 4

Fig. 6.5 Drive route during the measurement for Day 1 & 2 as mapped by the GPS tracker on the Android application

6.4.3 Hardware setup

1. Android devices

We used five Xiaomi M1 4i devices (released April 2015). The CPU of the devices is the Octa-core (4x1.7 GHz Cortex-A53 & 4x1.0 GHz Cortex-A53), running on the Qualcomm MSM8939 Snapdragon 615 (28 nm) chipset. The devices support Android SDK version of up to API level 23. For Day 1, these were each connected to SIM cards for Operators NP 1 - 4 and SSM-MoC (NP 5). For Day 2, 3 & 4, there was no NP 2 and NP 4. Instead, there were two NP 1 and NP 3 connections. While NP 1 - 4 SIM cards were picked up locally and the cost of data usage on them was approximately £1 / GB, NP 5 SIM card had to be specially ordered and its aggregate cost of data was £95 / GB.

2. Raspberry Pis

For Day 2, 3 & 4, we used four Raspberry Pi 4 devices (released June 2019). Each of the devices has the Broadcom BCM2711 SoC processor with a 1.5 GHz 64-bit quad-core ARM Cortex-A72 processor, and 1 MB shared L2 cache; had 8GB of RAM and ran the Raspbian 10 Buster operating system. They were housed in our specially-constructed glass-top box and are connected to the internet via USB tethering from the Android devices. Three of the Pis were connected to NP 1, 3 and 5. The fourth was connected to both NP 1 and 3 with the ability to switch between them.

3. Management network

For Day 2, 3 & 4, to access the Pis without individual screens, keyboards and mice, we setup a management network made up of a HP Pavilion laptop (Windows 10, AMD E2-1800 processor and 6GB RAM) and two Netgear DG834G routers with four ports, port speeds of 10/100 and maximum LAN speeds of 100Mbps. The routers are housed in the same box as the Pis and we used two routers because we needed five ethernet ports (four Pis and the laptop) but each router had only four ports. We modified DCHP with static IP address for eth0 on each Pi and then from the laptop, we used SSH and Microsoft Remote Desktop to access the Pis. Microsoft Remote Desktop was used for the initial setup and during the actual field measurement while SSH was faster for copying out results of our measurements from the Pis.

6.4.4 Software setup

1. Measurement software - Netperf-KCL

For Day 1, we used our specially designed *Netperf-KCL* Android measurement app, which was used for the measurements reported in Chapter 4. Netperf-KCL is based on an adaptation of Multiping-for-Android app[133] and enables us to conduct 24 different measurements for 14 websites every 5 minutes. These include RTT on both TCP port 80 and TCP echo port 7, PLT, packet loss, RTT Variance, uplink speed, downlink speed, network type (e.g. LTE, HSPA+ etc), base station ID, base station location, device CPU parameters, device RAM usage. The choice of websites was based on their Alexa Ranking as at December 2019 and on perceived importance to the UK digital society. In total, we recorded 728 readings per device in 5 hours 20 minutes, a total of 3,640 measurement cycles and 87,360 individual readings.

2. Measurement & Internet switching software - iSwitch

For Day 2, 3 & 4, we used iSwitch.py to switch between two internet connections based on which one had better performance at a given time. The iSwitch tool performed three roles. Firstly, it did a ping to our own AWS server every 3 seconds via the tethered phones to establish the RTT. Secondly, based on an average of all RTT measurements in every 10 seconds, the code switched between the available networks. Thirdly, the code measured page load times every 5 seconds by downloading a 1.3MB webpage. In total, 7.2K ping packets/phone were sent in 5 hours 37 minutes and the location and webpage download performance were monitored every 5 seconds.

3. TCP/UDP observation software: tcpServer & udpServer

For Day 3 & 4, we used separate client and server versions of tcpServer.py and udpServer.py to measure the performance of TCP and UDP for single network operations and for a DSM-MoC. We ran eight experiments concurrently each time (4 TCP and 4 UDP) and setup eight AWS EC2 Instances for each of the experiments. TCP experiments used port 4545 while UDP used 4646. We ran the experiments with different packet sizes in bytes (200, 1024, 2048 and 4096). For each iteration, we sent 10 packets, every minute, from the client (Raspberry Pi) to the server (AWS EC2 instance) and measured the time it took to receive all the packets at the server. The session timed out in 60 seconds if there was no confirmation from the server.

6.4.5 Benchmarks for analysis

Table 6.2 summarises the benchmarks that were used for evaluating the coarse-grained results from our measurements. These benchmarks are partly based on what is already obtainable on operational 4G networks globally (as reported by Opensignal)[135], plus the connectivity expectations for V2X use cases as stated in [73].

Table 6.2 Performance thresholds for Day 1 measurements

Metric	**Benchmark**
Uplink speed	25 Mbps
Downlink speed	50 Mbps
Packet Loss	0%
Round Trip Times (RTT)	100 ms
RTT Variance	20 ms
Page Load Times (PLT)	1000 ms
Availability	RTT < 10,000 ms

6.4.6 Comparison with drive-through tests

There is a long and well-established history of drive through tests to measure mobile network performance in society. These are used by regulators, network providers, equipment vendors and other agencies to map, assess and improve network performance[178] or to determine performance violations and penalties. However, a common methodological challenge for these is that most drive test systems are architecturally diverse, with each drive using incompatible interfaces and data formats[179]. Such tests are usually done only for limited number of times, in tightly controlled test environments, and with equipment that is carefully positioned and calibrated. In real life, a CC that needs to make decisions on connectivity options on a regular basis will have to rely on continuous measurement while using less-ideal equipment. Our methodology is designed to provide insights for this latter scenario.

6.5 Results

The results presented here are structured into three subsections. In the first subsection, we summarise the course-grained results from Day 1 and use it as the baseline for comparing Day 2 and Day 3/4 results. In the second subsection, the focus shifts to the fine-grained results from Day 2 which is based on actual switching of network connectivity in real-time. In the final subsection, we analyse the relative impact on TCP and UDP performances based on Day 3/4 results.

6.5.1 Day 1 - Establishing the hypothetical benefit of DSM-MoC

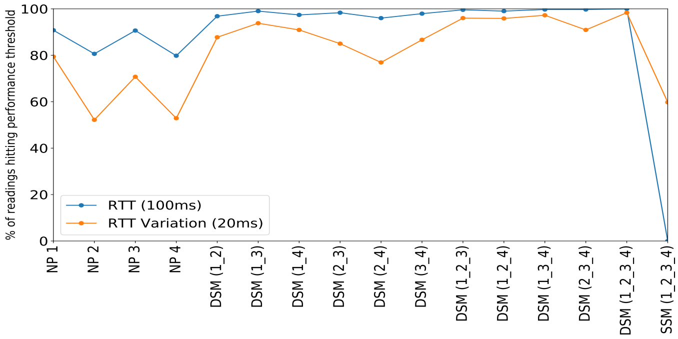
1. DSM-MoC offers best hypothetical reliability

Our first result is to show that a comparison of the actual performance of NP 1 - 4, NP 5 and 11 hypothetical combinations of DSM-MoC will show that DSM-MoC offers the best Q_s outcome (Figure 6.6). The Q_s result for NP 1 - 4 and NP 5 (SSM-MoC) is based on the field measurement on *each network* while the Q_s for DSM-MoC options is based on meeting the threshold on *any network*. Figure 6.6 provides the overall picture from our Day 1 driving measurements. From it, we demonstrate that there are significant gaps in performance across all the four NPs and SSM-MoC on all benchmarks. The implication is that any CC that decides to rely on any single NP is unlikely to have consistent reliability nor assured QoS across time and in all locations. Out of the 3,640 measurements, 68% achieved RTT of 100ms, 54% for PLT of 1000ms, 60% for uplink speeds of at least 25Mbps, 57% for downlink speeds of at least 50Mbps and 80% were connected to LTE. The cumulative performance data masks the diversity in performance on each of the NPs too; e.g. percentage of NP 2 PLT measurements <1000ms was only 19% vs 75% for NP 1 and 81% for NP 3/4.

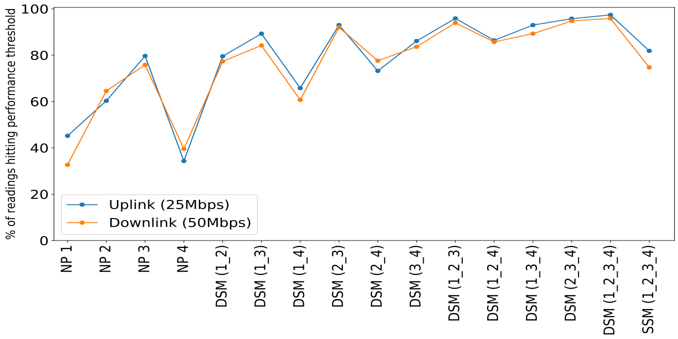
Comparing the different options in our result, it is evident from Figure 6.6 that, hypothetically, a DSM-MoC setup that is able to switch between NPs will always provide a much higher reliability than any single-operator option or the SSM-MoC option. This shows that if a four operator, DSM-MoC setup (i.e DSM-MoC 1/2/3/4) is implemented, overall system Q_s for RTT will improve to 100%. For RTT Variance, Q_s improves to 98%, for uplink speed 97%, for downlink speed 96%, for PLT 94%, and for packet loss 86%. The delta between DSM-MoC 1/2/3/4 versus the best single-operator NP 3 is as high as 28 percentage points (pp) for RTT Variance, 20pp for downlink speed, 18pp for uplink speed and 9pp for RTT.

2. SSM-MoC delivers inferior app performance

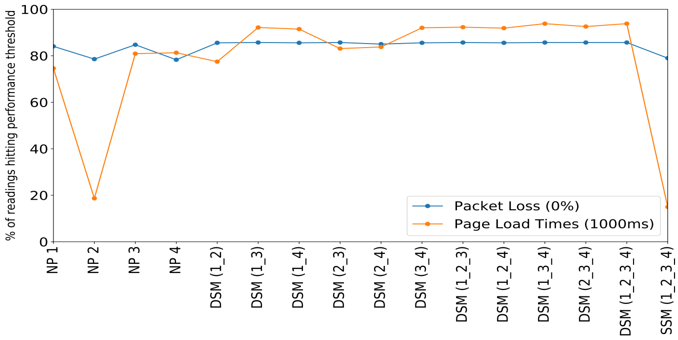
SSM is the default choice today for all CC connectivity solutions. From our results, we show that although NP 5 (SSM-MoC) achieved the fastest median speeds of 25.4Mbps for uplink and 50.7Mbps for downlink compared to NP 1 - 4, its performance on packet loss, latency, RTT Variance, packet loss and page load times were inferior. This negates any benefits of the faster speed it achieves. For downlink/uplink speeds, *getLinkDownstreamBandwidthKbps()* and *getLinkUpstreamBandwidthKbps()* return the speeds in kbps. For latency (RTT recorded



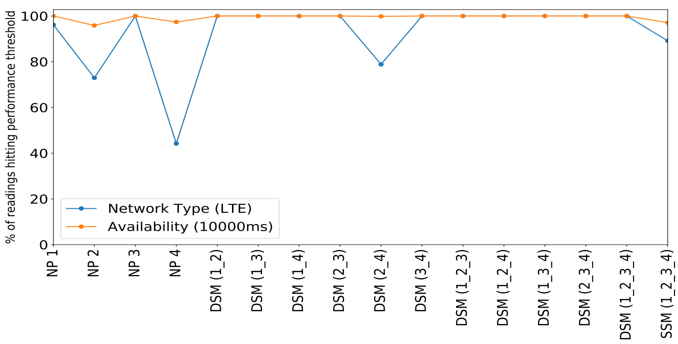
(a) RTT and RTT Variation



(b) Uplink and Downlink Speeds



(c) Packet Loss and Page Load Times



(d) Network Type and Availability

Fig. 6.6 Day 1 - Performance comparison for 8 parameters: actual performance from 4 network providers (NP 1- 4) & Supply-Side Managed provider (SSM); projected performance for 11 combinations of Demand-Side Managed (DSM) setups. The clear out-performance of DSM shows that a multi-operator setup will provide better reliability and QoS assurance.

in accessing a TCP port 80 or TCP echo port 7 for particular websites from our app), the median RTT for NP 5 was 152ms, four times worse than NP 3's 38ms. For RTT Variance (taking the average of five RTT readings), NP 5 was 15ms, same as for NP 3 and higher than NP 1's 11ms. For packet loss (observing packet losses recorded during our ping measurements), using NP 5 was the lossiest path, losing 100% of the packets in 18.3% of readings. For page load times (using Android's webview), NP 5 again had the worst outcome with only 15% of its readings meeting the 1000ms benchmark. Given that data costs on NP 5 costs >95 times more, the observations point to a profound conclusion: NP 5's performance is grossly inferior and does not warrant the cost.

3. Availability benchmark is insufficient for CCs

A simplistic approach to reliability assumes that reliability based on availability (i.e. R_s) equates to performance-constrained reliability (Q_s), ignoring any path dependency considerations. Yet, Figure 6.7 shows that for measurements on all five NPs, R_s is significantly better than Q_s. This provides confirmation that any assessment of reliability for CC connectivity that is based on availability will always show a better outcome than an assessment based on a performance constrained benchmark (i.e $R_s >> Q_s$). In our work, we use latency for analysis and assess availability (R_s) as an RTT of 10,000ms (10 seconds). This value is the Internet Control Message Protocol (ICMP) sessions default timeout and a delay after which user experience, even for web browsing on smartphones, becomes unbearable[180]. For Q_s, we use 100ms which is required for competitive gaming[180], and is close to double the overall measured median RTT on NP 1 - 4. We conclude from this that any official KPI or SLA that is benchmarked against availability R_s will suggest a better performance than a KPI based on Q_s.

4. SSM-MoC performs poorly vs Ofcom data

We sought to map the Day 1 result with Ofcom's coverage data - another independent empirical data - to explore for further insights. For this, we use the Day 1 results to compare the hypothetical reliability for the best combinations of DSM-MoC and SSM-MoC with the expected lower and upper reliability limits from Ofcom's coverage data. For a redundancy system with four network operators, Figure 6.8 shows that SSM-MoC performed significantly poor compared to the original expectation that such an MoC will be better than a single NP. Using Availability (RTT = 10000ms), SSM-MoC achieves only 97% even when NP 3 achieved 100%. For RTT = 100ms, SSM-MoC was 0% while NP 3 was 91%. For the rest

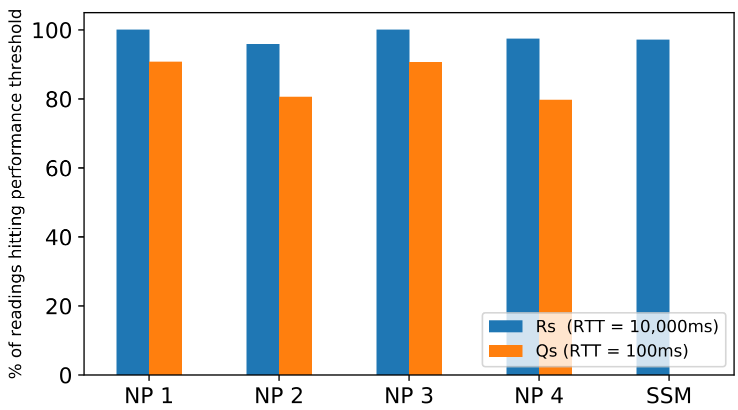


Fig. 6.7 Day 1 - Availability: Percentage measurements under 10,000ms vs 100ms. Benchmarks based on availability *R(s)* differ from performance constrained reliability *Q(s)*.

of the configurations, Figure 6.8 shows that the achieved reliability fits neatly between the lower and upper reliability limits from Ofcom data.

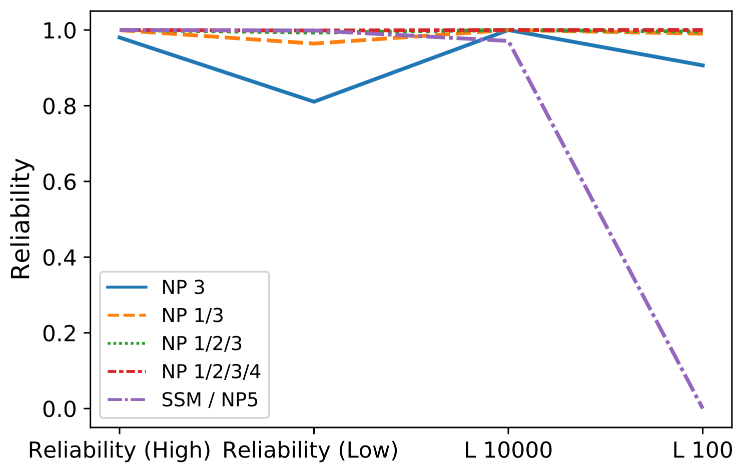


Fig. 6.8 Day 1 - Comparison with predicted reliability based on Ofcom data. It is clear that a 4-network SSM-MoC is inadequate, whether for availability (RTT = 10000ms; $R(s) = 97\%$) or specific threshold (RTT = 100ms; $Q(s) = 0\%$).

6.5.2 Day 2: Implementing DSM-MoC & validating its benefit

1. DSM-MoC achieves better performance overall

Unlike the SSM result on Day 1 which did not conform to expectation, the NP 1/3 DSM-MoC on Day 2 delivered better overall performance than NP 5 (SSM-MoC) or using either of NP 1 or NP 3. We measured RTT every 3 seconds on each of the Raspberry Pis with iSwitch and took the average of the readings every 10 seconds to determine best network to switch to. For Pi 1- 3 with only a single network connection, there is no switching to do. For Pi 4, we switched between NP 1/3. We measure actual application performance by measuring every 5 seconds on the Pis, the PLT for a 1.3MB webpage hosted in AWS. Figure 6.9 shows that the DSM-MoC of NP 1/3 achieved a median page load times of 103 milliseconds, compared to 496 for SSM-MoC, 115 for NP 3 and 1475 for NP 1; that is, DSM-MoC was 12% better than NP 3, 382% better than SSM/NP 5 and 1332% better than NP 1. In doing so, we validate that DSM-MoC provides better QoS assurance whether in a hypothetical scenario (Day 1) or in reality (Day 2).

It is necessary to put an emphasis on the 'overall' performance instead of an instantaneous performance because DSM-MoC did not consistently deliver the better outcome always. In fact, Pi 3 / NP 3 achieved the lowest maximum and lowest minimum page load times. This can be seen in Figure 6.9 too as the NP 3 began to consistently outperform the combined NP 1/3 option beyond the median mark when NP 1/3 began to track NP 1 more closely. This observation highlights the less-than-perfect reality of implementing DSM-MoC due to technical and operational challenges where switching delays or unexpected hardware/software behaviours force a deviation from the anticipated performance.

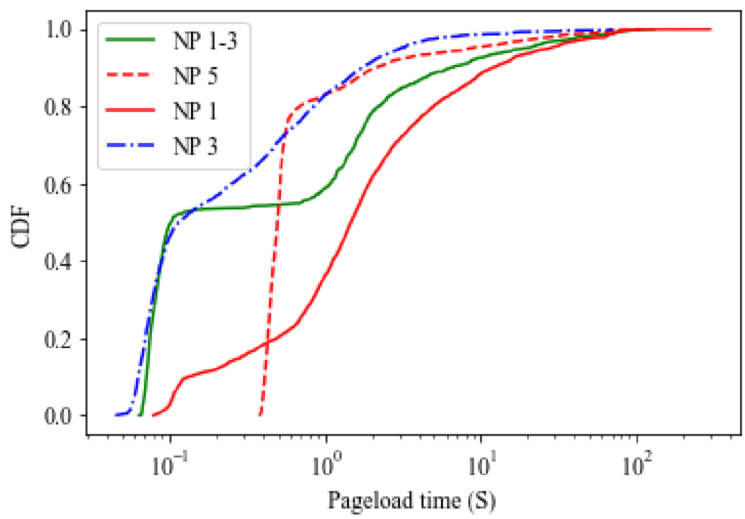


Fig. 6.9 Day 2 - Distribution of page load latency to a webpage of size 1.3MB hosted in AWS. CDF shows that, overall, the combined NP 1/3 DSM-MoC achieved a better page load times outcome than the SSM-MoC / NP 5 or either NP 1 or NP 3

2. Predictive algorithms better for DSM-MoC

Our DSM-MoC implementation is based on RTT readings every 3 seconds, and switching based on the average of readings every 10 seconds. However other implementation options exist as our extensive evaluation of the latency readings in Table 6.3 show that there is a trade-off in terms of performance improvements vs number of switching required. In this case, switching based on a 10 seconds window will deliver 47.71% improvement vs NP 1 and 5.95% improvement vs NP 3 alone. But this requires 182 switches, creating more room for error and switching delays in any real life implementation. This observation highlights the challenge to meeting our QoS assurance targets with SSM-MoC. For time critical scenarios, the delays in getting to the cloud-based switching centre and the inherent switching delays all contribute to degrading the realised performance outcomes.

One other implementation option for DSM-MoC is to use predictive tools to determine when to switch rather than using static switching windows. This gives the opportunity to make the switching decision with a view to optimising either for performance or for number of switches. In our analysis in Table 6.3, we see that a 60 second window achieves a not so dissimilar outcome from the 10 second window but with, significantly, only 54 required switches. In this case, our hypothetical Oracle achieves the optimal performance improvements for NP 1 (58.13%) and NP 3 (5.17%), and also the fewest number of switches (52).

3. Geographical variation in DSM-MoC performance

For a CC driving along disparate road scenarios, the challenges of implementing DSM-MoC can also be seen in the spatial variation of the latency readings along our 380 km route. In Figure 6.10, we show the latency readings along a North-East-South-West coordinate map, highlighting which of the four Raspberry Pis delivered the best latency readings in our measurements. For the combined NP 1/3 Pi, we compared the latency that would have been recorded if our Oracle was taking the two latency readings from NP 1 and 3. We show that despite the spatial variations along the route, the combined setup is well positioned to realise the lowest latency readings.

6.5.3 Day 3/4: Determining impact of DSM-MoC on TCP/UDP

1. DSM-MoC benefits exist across different TCP/UDP scenarios

Having confirmed from Day 2 measurements that DSM-MoC achieves better performance, we designed Day 3/4 measurements to deepen our understanding the impact of DSM-MoC

Table 6.3 Day 2: Latency readings per ‘t’ seconds used for switching vs performance improvements & no of switches. A DSM-MoC oracle will deliver a more optimal outcome

t (s)	NP 1 (%)	NP 3 (%)	#switches
10	47.71	5.95	182
20	47.62	6.18	132
30	47.17	5.01	110
40	45.64	2.32	92
50	46.21	3.07	64
60	47.08	5.17	54
Oracle	58.13	22.49	52

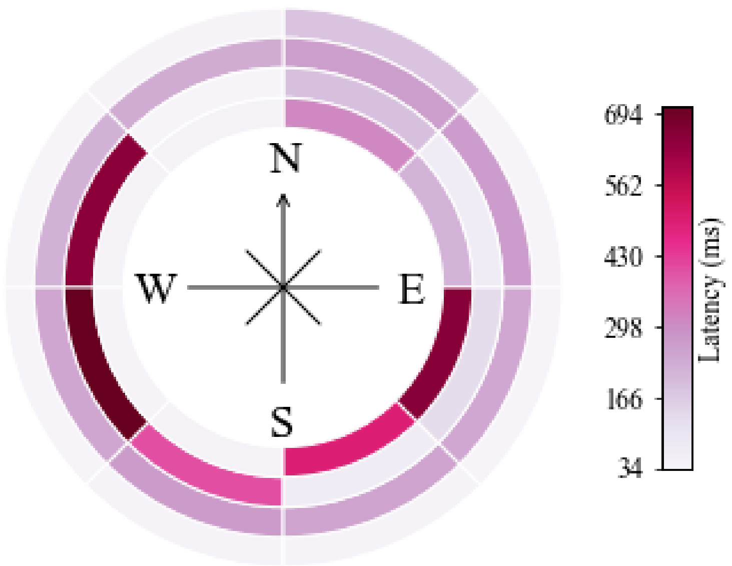


Fig. 6.10 Day 2: Latency comparison across different directions on our route for the four Raspberry Pis: Oracle (NP 1/3), NP 5, NP 3, NP 1 (from outer to inner)

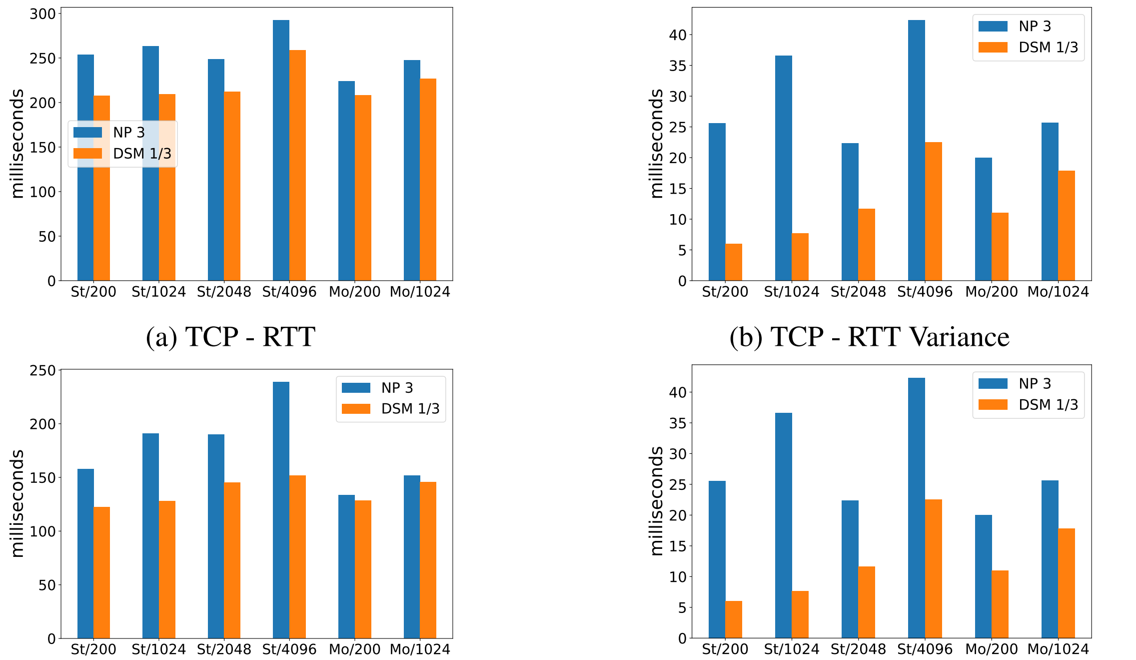
switching on the key internet transport protocols of TCP and UDP. We did this for different TCP/UDP packet sizes and for a stationary vs mobile environment. Figure 6.11 shows that, compared to a single operator, DSM-MoC achieves better performance for different TCP/UDP packet sizes and in both stationary and mobile environments. For simplicity, we compared with only NP 3 which has already proven to have the best network performance on the same route from Day 1 results. From the 2165 readings recorded, we observe wide variation in the minimum and maximum RTT and RTT Variance values for different TCP/UDP packet sizes. Overall, the median improvements were 13% for RTT on TCP, 23% for RTT on UDP, 47% for RTT Variance on TCP & 67% for RTT Variance on UDP. Clearly, UDP benefited more from sharing, suggesting that for applications that rely more on UDP, a DSM-MoC approach is advantageous.

2. DSM-MoC benefits were higher for stationary vs mobile scenario

Regardless of whether we are using DSM-MoC or a single network operator only, our assumption for Day 3/4 experiments in a suburban stationary location with few or no changes in base station connections is that there would be less impact on TCP/UDP performance as there would be fewer cases of abruptly terminated TCP/UDP connections. Figure 6.11 confirms that this assumption is correct. The opposite is also true - in a mobile test scenario, TCP/UDP performance deteriorates more than in a stationary test scenario. In fact, [116] reports that the impact of switching TCP sessions can be as large as 10s. We recognize that this is applicable to DSM-MoC and the impact of switching networks, in addition to switching cell sites, will impact the relative performance advantages expected. Nonetheless, even for the mobile environment with cell handovers and inter-operator changes, Figure 6.11 is still clear that DSM-MoC provides a superior performance than a single operator. We observe that, for all packet sizes, RTT for data transfer on both TCP and UDP on DSM-MoC achieve >10% improvements over NP 3 in a stationary environment and single digit improvements in a mobile environment. For RTT Variance, the improvements are bigger but again, with better improvements from switching in a stationary scenario.

3. Smaller packet sizes did not achieve best improvement

Another assumption we had for Day 3/4 experiments was that smaller packet sizes will be less susceptible to disruption as the TCP/UDP packets are small enough to be transmitted in the short time windows. However, our results show that we were wrong as the benefits of switching do not favour the lowest packet size of 200 bytes. Instead, Figure 6.12 shows that measurements with 1024 bytes packet sizes had the best performance for both TCP and UDP



(a) TCP - RTT

(b) TCP - RTT Variance

(c) UDP - RTT

(d) UDP - RTT Variance

Fig. 6.11 Day 3/4: Median Latency (RTT) and RTT Variance achieved by NP 3 vs DSM-MoC 1/3 for different TCP & UDP packet sizes in stationary (st) and mobile (mo) scenarios. For simplicity, we compared only with NP 3 because we have already established from Day 1 that NP 3 had the best network performance on the route. Three clear insights emerge. First, a DSM-MoC implementation consistently achieved better RTT and RTT Variance for all packet sizes and in stationary and mobile environments. Second, UDP improvement was much higher, suggesting that a DSM-MoC approach is advantageous in use cases with UDP-reliant applications. Third, RTT measured on the Raspberry Pi is significantly worse than RTT measured on smartphones on Day 1, further highlighting the difference between hypothetical scenarios and system-level implementation.

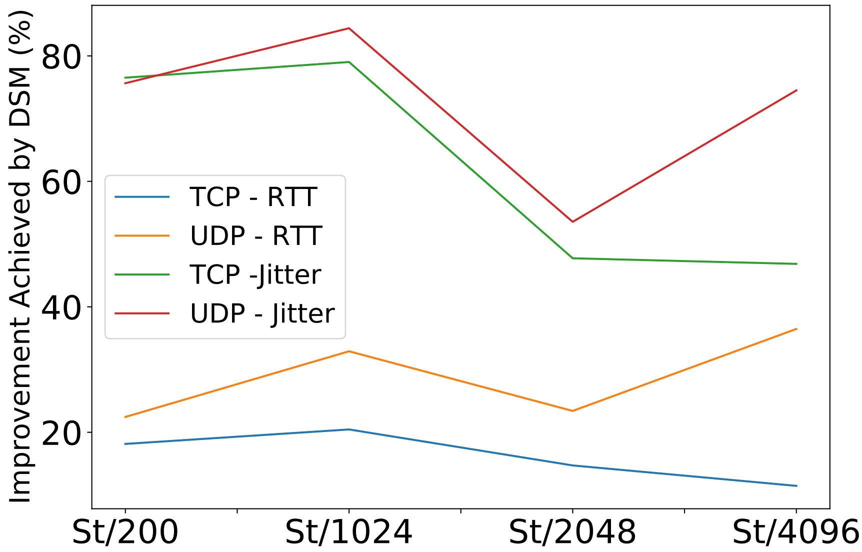


Fig. 6.12 Day 4: Stationary DSM-MoC benefits was highest for TCP/UDP packet sizes of 1024 bytes (except outlier for 4096 bytes on UDP) contrary to expectations that 200 byte packets will witness the best improvements

for RTT and RTT Variance. Our hypothesis is that as 1024 bytes is the size of the file system so the buffering is optimised for 1024 in order to ensure that file transfers are more efficient.

4. System implementation differs from smartphone implementation

An interesting insight from Figure 6.11 is that the RTT and RTT Variance measured on the Pi in the CC on Day 3/4 is significantly worse than the RTT and RTT Variance measured on smartphones on Day 1 on the same route. This insight highlights a major difference between our work and previous works which have focused on multi-connectivity on smartphones. In our system implementation, the smartphone acts as the radio unit while the Raspberry Pis and the Netgear router are the main computing units running the application. The additional footprint and distance means that, compared to a smartphone-only setup, the data traffic traverses more network nodes and frontiers which would explain the longer RTT and RTT Variance.

6.6 Discussions, Implications & Recommendations

6.6.1 DSM-MoC delivers superior performance

Whether from Day 1 hypothetical scenario or Day 2 actual measurement data, it is clear that DSM-MoC delivers superior performance. For a 4-network redundancy setup, Figure 6.13 shows that across all parameters, a hypothetical DSM-MoC setup will deliver a superior performance (based on Day 1 measurements). In fact, based on Day 2/3/4 results, even the 2-network DSM-MoC, as long as it includes NP 3, will deliver superior performance compared to the 4-network SSM-MoC. Yet, digging deeper, we see from Figure 6.14 for Day 1 that the worst 2-network DSM-MoC would have under-performed the best performing single network (NP 3) for uplink/downlink speeds and network type. In contrast, the worst 3-network fallback option will consistently outperform NP 3.

While this observation can not be generalised, it suggests that using only a 2-network DSM-MoC cannot always provide assurances of superior performance, especially if those two networks are the worst performing. As such, rather than rely on the two worst networks for QoS assurance, a CC manufacturer is better off striking an exclusive deal with the best-quality network provider. In practice, and before factoring in network sharing, most countries have at least three network providers (UK has four), making it feasible to deliver DSM-MoC.

6.6.2 SSM-MoC's network architecture creates challenges

In our setup, we note specifically that the SSM-MoC operator promotes its service as being able to optimise on coverage, performance and price. In contrast, our data suggests that it has, instead, optimised mostly for speed to the detriment of other performance indicators. The implication is that the performance we have seen for SSM-MoC contradicts the expectation of using such setup (whether via an MVNO or national roaming) and the justification for its hefty fees.

Our investigation suggests that the relative under-performance of SSM-MoC is an intrinsic outcome of its architectural design. The SSM-MoC provider makes it clear that it routes cellular traffic through its own cloud-based mobile core network, presumably with its own packet gateway (PGW). We confirm this longer route in Figure 6.15 by noting the number of traceroute hops to reach our AWS server. By design, we would expect this setup to negatively impact latency-related measurements (e.g. RTT, RTT Variance and PLT) as the route from our measurement device to the application server is elongated. However, on the purely network-dependent parameters (e.g. downlink/uplink speeds, network type), our expectation was that SSM-MoC should be outperforming NP 1 - 4 as it is capable of switching host

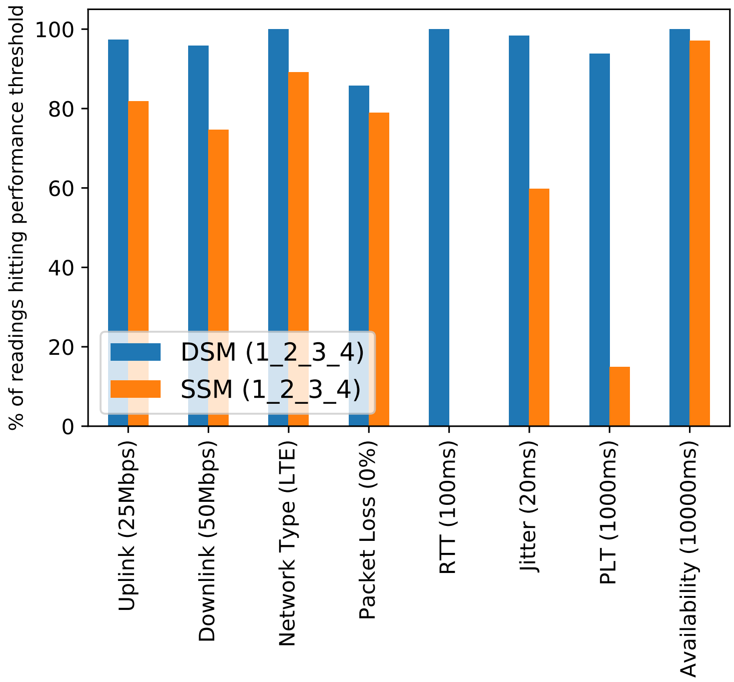


Fig. 6.13 Day 1: Comparison of SSM vs DSM performance showing that SSM significantly under-performs a similar DSM

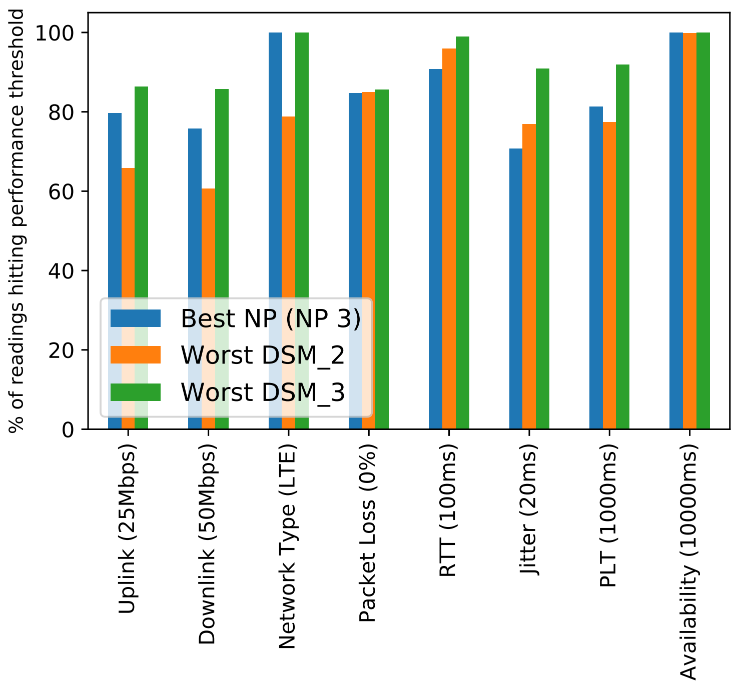


Fig. 6.14 Day 1: Comparison of best NP vs the worst 2-network & 3-network DSM to understand optimal fallback options

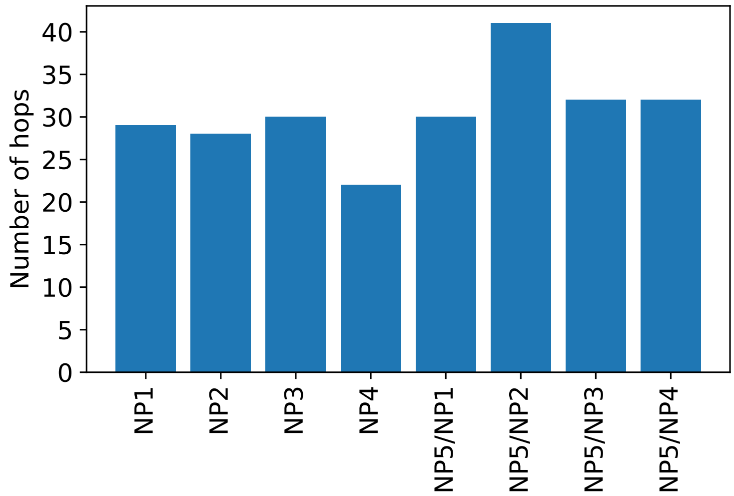


Fig. 6.15 No of traceroute hops to 3.8.114.122, showing that SSM-MoC takes a much longer route

NPs to achieve better performance. This was only the case for uplink/downlink speed where SSM-MoC's 25.5Mbps and 50.7Mbps median was the highest. Unfortunately, our analysis of the root cause of this is limited given that SSM-MoC had the same base station ID with NP 2 for only 7% of readings (and none with NP 1/3/4).

6.6.3 Operational complexities for DSM-MoC are surmountable

In our Day 2/3/4 experiments, we have successfully demonstrated that DSM-MoC can be implemented in a test environment, albeit with teething problems that can be resolved to enable its deployment in a production-grade, commercial environment. For example, we note that there are commercial offerings, using an AI engine in the cloud on the DSM-MoC principle, for smart ambulances (e.g Juniper's Contrail SD-WAN solution demoed at Mobile World Congress 2019[1]).

6.6.4 Multi-radio connectivity as default for CCs

A major hindrance to the use of DSM-MoC is on how many radios to have in a CC and how many should be switched on simultaneously. This question determines the power efficiency of the connectivity unit and, ultimately, the overall cost of the system. Despite these concerns, and while a full examination of the energy efficiency are out of scope

[1]https://blogs.juniper.net/en-us/industry-solutions-and-trends/smart-ambulance-demo-at-mwc-showcases-critical-5g-sd-wan-use-case

in our work, our expectation is that CCs should have at least two radios switched on at the same time in keeping with the expectation for redundancy as contained in ISO 26262 (“Road vehicles - Functional Safety") and its master guide, IEC61508 (“Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems")[29]. We argue that this should be a prerequisite for any vehicle operating on a fully autonomous basis. A multi-radio connectivity provides a fall-back, redundant option and also makes it possible to split the data packets across several radios or to send all the data packets simultaneously through all the radios.

6.6.5 CCs have computational power to manage DSM-MoC

In a multi-radio setup, the question of who makes the decision on which radio to use is at the heart of whether the system is SSM or DSM. Figure 6.16 illustrates this dilemma on whether to repatriate the decision making to the nodes/CCs or not. In our work, we argue that cars, as ‘datacentres on wheels’, have the requisite computing power and AI capability to automate such decisions. This explains why driverless cars are increasingly able to make their decisions onboard instead of relying on a central coordinator[125], an approach in contrast to the industry default where the design philosophy of the mobile network has encouraged dumb nodes and intelligent core. This philosophy explains SSM-MoC, National Roaming and new initiatives such as Juniper’s Contrail SD-WAN solution. Under this approach, if a user decides not to hand over decision making to a centralised authority, then the user with multi-SIM devices have to manage the multi-connectivity themselves, including resetting/restarting their devices where necessary. With our manual setup, we learnt how difficult this can be from our Day 2/3/4 measurements. Despite months of setup and test runs, the system still misbehaved on the day and required multiple manual restarts to ensure seamless operations.

6.6.6 Cost constraints discourage DSM-MoC

Based on current industry structure, DSM-MoC is likely unaffordable in its current setup requiring multiple radios. In a car industry with razor-thin margins, the cost of supporting multiple connectivity options is almost a non-starter. This probably explains why most CC deals are exclusive with a single network provider. McKinsey forecasts that connectivity will present a revenue potential of between $130 to $210 for basic connectivity and $400 to $610 for advanced connectivity; annual cost savings are in the range of $100 to $170 and $120 to $210 per vehicle, respectively[125]. Putting them together, the total value from connectivity per car will range from $230 to $820, giving an average of $530 per annum or $44.2 per month, a value that is similar to the average revenue per user for a typical mobile phone user

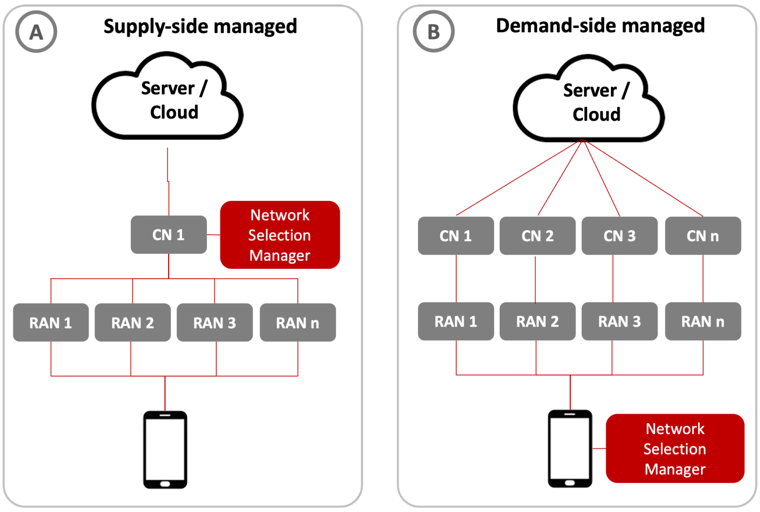


Fig. 6.16 Where is redundancy managed? [A] SSM-MoC - Primary network operator manages redundancy at the core network (CN) and makes decisions on which RAN to use. [B] DSM-MoC - User manages redundancy at the local level and selects RAN to optimise performance requirements. Modern cars have enough computing power for DSM-MoC

in the US. Our industry view is that IoT monthly ARPU is significantly less than smartphone ARPU. However, even if IoT can achieve the same $44 per month, it is still a huge challenge to profitably divide the $44 across multiple network providers.

6.6.7 Policy guidance is imperative

In a scenario where safety is the goal, ultimately, the DSM-MoC vs SSM-MoC trade-off will come down to policy guidance. We note that our measurements, on mature 4G networks in a country with over 10 years of 4G availability, highlight the inadequacy of a single cellular connectivity to meet stringent performance requirements. While 5G will bring improvements to reliability, we are not convinced that it will suddenly make any single network adequate. Performance studies on early 5G networks in China[56], UK[57] and USA[55] do not yet suggest any significant performance improvement above 4G nor would 5G coverage become extensive any time soon to cover all roads. Given that the industry expectation is for 5G & 4G to coexist well into the 2030s[2], we argue that our proposals for DSM-MoC are applicable now and deserve attention from policymakers, commercial executives and technical planners today.

We recognize existing steps towards improved reliability for CCs. Already, there are rules in place to support fallback and roaming for selected services (e.g. eCall in Europe[120]). 3GPP has also made a start by reserving QCI/5QI 3 and 79 for V2X messages or by proposing special network slices for V2X[73]. While these mechanisms can improve the reliability, they are unlikely to provide as much geographic and performance reliability as using redundant connectivity options, especially for safety-critical V2X use cases (e.g platooning). Besides, as argued in [156], technical specification without an enabling commercial and policy framework is unlikely to be actualised.

In the spirit of technology neutral policy making, a concerted push by regulators and policymakers for stringent connectivity KPIs for all cars will nudge the industry towards developing commercial & technical models for DSM-MoC.

6.7 Key contribution

From the results of Day 1 (published in [174]), we made two main contributions:

1. We quantified, via extensive field measurements on live networks in the UK, that using a hypothetical DSM-MoC redundancy setup can deliver superior performance advantages over the best-performing single-operator option. This can be up to 28 percentage points for latency.

2. We showed that a hypothetical edge-based DSM-MoC can offer significantly better outcomes than cloud-based SSM-MoC which only achieves up to 4.8x longer page load times. The emphasis on DSM is contrary to common practice today where, under the concept of 'national roaming'[176], it is assumed to be the responsibility of the network provider to provide SSM redundancy.

For Day 2, 3 & 4 measurements, we built & deployed a system implementation of DSM-MoC on a CC on live UK networks. Based on the Day 2/3/4 measurements (published in [113]), we make additional three contributions:

1. We demonstrate that despite the challenges in switching networks in real time in a deployment scenario, DSM-MoC still outperforms both SSM-MoC and single network options. Our results show that the median page load times for DSM-MoC 1/3 of 103 milliseconds was 12% better than for NP 3, 382% better than SSM/NP 5 and 1332% better than NP 1. This validates the hypothetical results from Day 1 and provides a compelling need for commercial and regulatory action to encourage use of DSM-MoC for safety-critical systems.

2. We show that DSM-MoC delivers greater improvements to UDP compared to TCP. For RTT, this was 23% for UDP compared to TCP's 13%. For RTT Variance, this was 67% for UDP compared to 47% for TCP. This has practical implications for safety

critical services in CCs that rely on UDP protocol such as voice/video communications, computer gaming, and the emerging trend to create 3D simulation and virtual worlds for the metaverse.

3. We demonstrate that DSM-MoC benefits are applicable at the typical packet sizes used for data communication because of the limit of the Ethernet system (i.e. 1500 bytes) with measurements using 1024 bytes achieving the best outcomes compared to 200 bytes or 2048 bytes. This observation is helpful because it means that, as the overall network infrastructure favours these sizes, there is no need to re-architect the internet infrastructure and plumbing to support DSM-MoC.

6.8 Chapter Summary

This Chapter has provided an extensive empirical evaluation of the reality, benefits and challenges of a multi-operator connectivity option for connected cars. This is important because CCs are becoming increasingly common on the roads and as a bona fide SCS, the expectation is that they will be served by ultra reliable connectivity solutions to assure safety of drivers, passengers and passers-by. The work investigated how to improve connectivity reliability and provide QoS assurance for CCs using an MoC setup. Unlike other works that are mostly for price/quality arbitrage on smartphones, the novelty of our work is that we focus on assuring reliability for safety (i.e. safety arbitrage) on a system-level implementation in real-time and real-life on a CC.

Our work quantifies the benefits and shortcomings of MoC and shows that the DSM-MoC option delivers superior performance than reliance on a single network operator. We show that the hypothetical delta between DSM-MoC of four operators versus the best single-operator is as high as 28 percentage points (pp) for RTT Variance, 20pp for downlink speed, 18pp for uplink speed and 9pp for RTT. For the actual system implementation, we show that a two operator DSM-MoC achieved a 12% better performance than the best single-operator in downloading a 1.3MB webpage hosted in AWS. This provides insights to ***answer RQ 3.2 on what are the quantifiable benefits of using any of the D-QoS approaches on public mobile networks for selected use cases.***

I have now established that D-QoS is necessary because of the shortcomings in live 3G/4G/5G networks (Chapter 4), that it can be used to address those shortcomings despite its challenging history (Chapter 5) and shown, via empirical measurements on live 4G networks, that a D-QoS implementation can deliver its benefits in a real live deployment (Chapter 6). In Chapter 7, I will introduce a decision framework to guide stakeholders on how to use D-QoS to provide reliability assurance for SCS.

Chapter 7

D-QoS Guidance Framework

A kọruoro ngwere, ya ekwee n'isi (Igbo Proverb)

(A lizard nods its head in approval after hearing the detailed story)

7.1 Introduction

The history of D-QoS shows that it is not a new phenomenon. Yet the reality is that D-QoS adoption and market traction has been poor (as explored in Chapter 5 and [25]). Much of the academic and industry literature on D-QoS have focused on the technicalities of each D-QoS implementation or their incorporation into the standards (e.g. 3GPP). While these have been okay to elucidate the underlying technical details of each D-QoS mechanism, there has been lack of clarity on how to categorise different D-QoS mechanisms and as such, little or no comparative analysis of them. The result is that there is no industry-level guidance to commercial and policy stakeholders on when to ask for D-QoS and which ones to use. In this Chapter, I strive to close this gap by conceptually distilling the technicalities of D-QoS into practicalities that can be understood and adopted by the broader ecosystem. Accordingly, I introduce the D-QoS Guidance Framework (DGF) (Figure 7.1) to provide a step-by-step guide to stakeholders on selecting the D-QoS to use or mandate on mobile networks for SCS. In particular, I develop the DGF based on the type of infrastructure where they are used on - Single Shared Infrastructure (SSI), Multiple Shared Infrastructure (MSI) or Single Exclusive Infrastructure (SEI). My underlying rationale is that as all mobile networks experience occasional periods of under-performance (as empirically demonstrated in Chapter 4), the DGF should focus on how to achieve QoS assurances for SCS on unreliable mobile networks. Accordingly, once the need for mobile connectivity is established for SCS, the DGF drives for clear guidance on when private mobile networks (examples of SEI), prioritisation (examples of SSI) or redundant multi-connectivity (an example of MSI) should be used.

7.1.1 Key questions answered

In this chapter, I bring together all the insights, analysis and empirical evaluations in my research into a single framework that can be used to inform commercial and policy action. I then use the example of a 999 (or 119/911) app to illustrate how the framework can guide all stakeholders to understand how to treat the app. The DGF will continue to provide answers to the research question:

Research Goal 3: *Evaluate the reality of D-QoS approaches to understand their relative advantages in providing QoS guarantees for SCS*

In particular, the framework will address:

> **RQ 3.3:** How can key industry stakeholders (including policy makers) support the application of D-QoS?

7.1.2 Chapter layout

In this chapter, I will introduce the "D-QoS Guidance Framework" (DGF) to summarise the decision points on the use of D-QoS for SCS. Section 7.2 explores the DGF using five sequential questions. It will then focus on the type of infrastructure used for D-QoS (Section 5.3.2 and Figure 5.1) to make recommendations on when to use each D-QoS option. In Section 7.3, the framework is applied to the 999 app example to illustrate how DGF can be used in practice. Finally, Section 7.4 provides conclusions from the work in this Chapter.

7.1.3 Publications linked to Chapter

There are no specific publications that inform this chapter. Instead, the chapter draws inspiration from all the publications listed in Section 1.7.

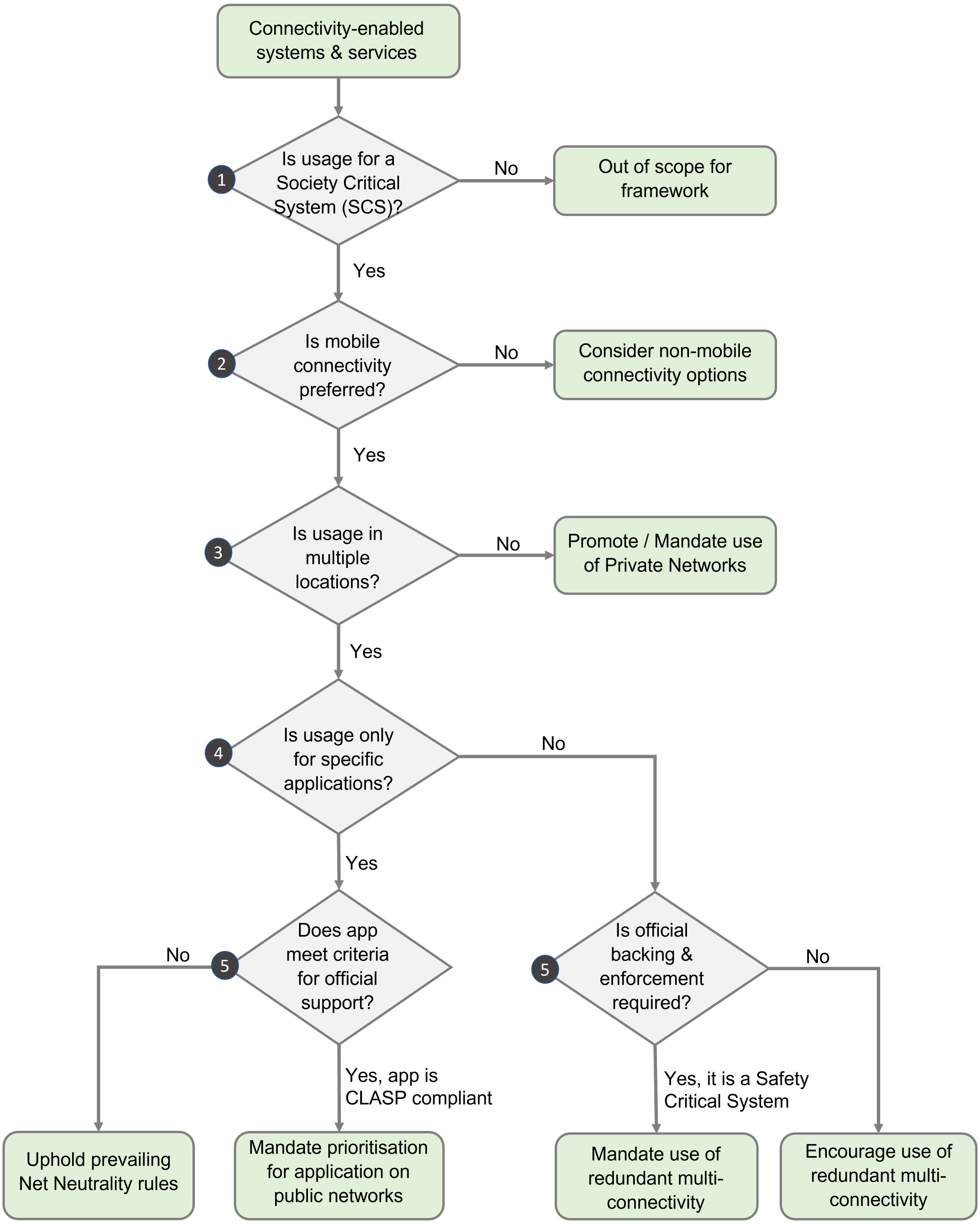


Fig. 7.1 D-QoS Guidance Framework (DQF): a guidance framework for deciding how to apply D-QoS for different Society Critical Services (SCS) that require wireless connectivity. Framework offers choices for using a Single Exclusive Infrastructure (SEI), Single Shared Infrastructure (SSI) or a Multi Shared Infrastructure (MSI)

7.2 Exploring the D-QoS Guidance Framework

7.2.1 Scope

The DGF offers a set of five questions to help commercial and policy stakeholders to determine whether D-QoS is needed and what type of D-QoS is needed. However, the DGF is not a stand-alone decision framework and does not provide an effective assessment of the feasibility nor viability of the use of D-QoS.

7.2.2 Question 1: Is it an SCS use case?

Question 1 determines whether the usage scenario falls within the scope of DGF or not. In Chapter 2, I introduced five types of Society Critical Systems: Safety Critical, Mission Critical, Business Critical, Critical Information and Security Critical. If the usage scenario falls into any of these SCS use cases, then it is necessary to consider the use of D-QoS to improve reliability and provide QoS assurance. The answer to Question 1 will determine if D-QoS is needed or not.

7.2.3 Question 2: Is mobile connectivity preferred?

Question 2 calls for a frank assessment of the connectivity options for the SCS use case. In many cases, a wired connectivity option will be more reliable and secure. Occasionally, a WiFi connection can suffice, especially for use cases where connectivity is only required sporadically. Likewise, other proprietary connectivity solutions (e.g. SigFox, LoRa and TETRA/Project 25[181] and LTE for Public Safety[20]). Figure 7.2 summarises the comparative advantages of 5G in supporting SCS. The answer to Question 2 will determine if mobile D-QoS is needed or not.

7.2.4 Question 3: Is usage in multiple locations?

Question 3 is the first decision point on the type of D-QoS to use. The intention is to determine if the usage scenario is relatively fixed in a single location. If so, and the answer to Question 2 recommends mobile connectivity, then it becomes necessary to suggest that a localised mobile network may be needed on site/premise. As such, the answer to Question 3 will determine if the use of a private mobile network (i.e. an example of SEI) is needed or not. While the use of private mobile network will mostly be non-compulsory, commercial and policy stakeholders may decide to mandate its use for security or Business Continuity purposes.

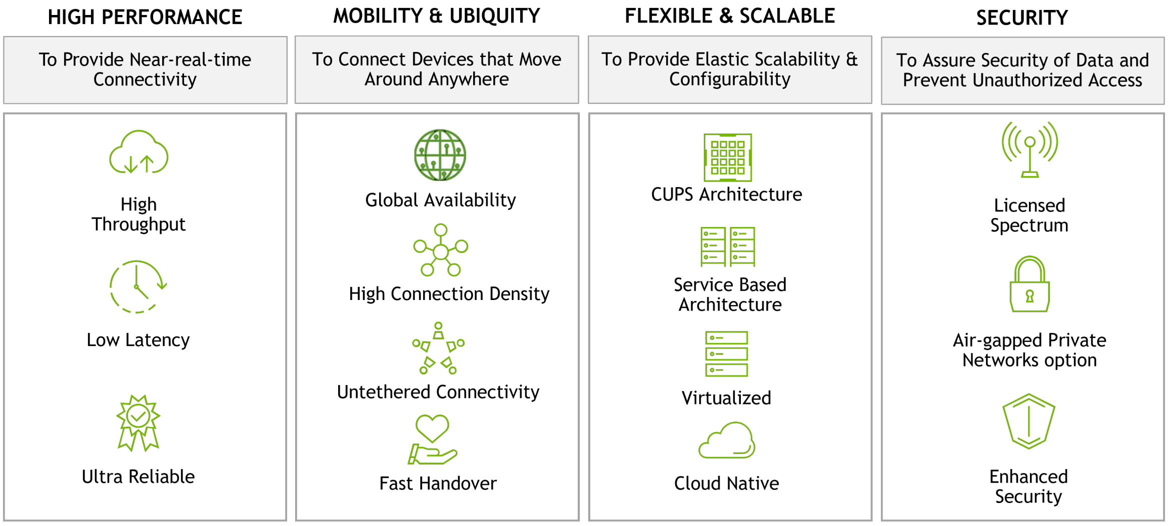


Fig. 7.2 Advantages of 5G: Qualities of 5G that provide a comparative advantage for SCS

7.2.5 Question 4: Is usage only for specific applications?

Question 4 drills down to the actual service that is used, to understand whether QoS assurance is required for only one service, specific services, or for all services. For example, is the usage only an app on a phone (or a specific webpage) or is it a multi-system usage in a CC. If it is the former, then it becomes appropriate to provide a targeted solution to provide QoS assurance to that service. If the latter, then it is appropriate to seek ways to provide QoS assurance to the entire system. The answer to Question 4 will determine if to use a prioritisation mechanism (i.e. an example of SSI) or redundant multi-connectivity (i.e. an example of MSI).

7.2.6 Question 5: Does usage meet criteria for official support?

Question 5 is a recognition that not all SCS use cases can benefit from official backing and so policymakers must be selective in how they intervene to mandate the use of D-QoS. For the use of prioritisation, we introduced CLASP (Critical, Localised, Authorised, Specific and Perishable) in Chapter 5 and in [75] as a set of conditions that need to be met for policymakers to provide official backing for any service to be prioritised on public mobile networks. A 999 app is an example of a service that is CLASP-compliant whereas traffic updates on Google Map would not be CLASP-compliant unless it is officially ‘Authorised’. For redundant multi-connectivity, I propose that Safety Critical Systems which are already under tight scrutiny by the Health & Safety Authorities, should be mandated to use a multi-connectivity option. A Connected Car is an example of such a system. Overall, the answer to Question 5 will determine if the use of D-QoS is mandatory and enforced or advisory only.

7.3 Implementability of DGF

7.3.1 Policy and Regulatory Considerations

DGF is anchored on the ability of key stakeholders to adhere to clear rules and guidelines. Accordingly, the DGF will more likely be adopted by the industry if it is backed by regulatory fiat and mandated to be adopted by the appropriate regulators with full political and legal backing. Policy and regulation considerations permeate the five questions that are central to the DGF, especially Question 1 and 5.

With Question 1, the concept of an SCS is intrinsically a policy and regulation affair. As described in Chapter 1, the failure or malfunction of an SCS has severe consequences which society tries to minimise or avoid. So far, there is no existing policy frameworks globally to guide regulators on how to legislate for the use of mobile connectivity for SCS. This means that rules are based on the existing Health and Safety policies which are generally inadequate to prevent harm from any failure or malfunction due to unreliable mobile connectivity.

For Question 5, policymakers will need to make a decision as to whether the SCS merits official backing. In doing so, they will weigh the pros and cons to society, and consider their fiduciary duty to protect the safety of lives and businesses.

7.3.2 Technical Considerations

DGF is a quasi-technical framework and seeks to provide guidance to ensure that the failings of previous D-QoS mechanisms are avoided. As such, it is not strictly prescriptive on which technical mechanism to achieve its aims. Rather, it sets out the goals that need to be achieved while leaving it to the stakeholders to work out which specific D-QoS mechanism to use. In this way, the choice of technical solution can be localised to the specific requirement of the SCS use case.

Technical evaluation is especially key for Questions 2 and 3 for exploring the DGF. On Question 2, an SCS provider will need to evaluate the merits and demerits of using mobile connectivity in the solution. In some cases (e.g. Automated Guided Vehicles in a warehouse), this will be a fairly straightforward decision as other wired and wireless options may proof inadequate. On Question 3, the SCS provider will then need to establish which of its D-QoS solution is preferred based on whether the service is static or mobile. Examples such as

Connected cars are straightforward scenarios where Multiple Shared Infrastructure D-QoS solution will be optimal.

7.3.3 Commercial Considerations

By its very nature, the DGF is designed for SCS which means that the commercial advantage from using DGF is more about minimizing cost and human suffering rather than maximizing commercial gain. This is the compelling commercial driver for adopting and implementing DGF as the cost to society or to the brand of businesses, in the event of an SCS failure due to unreliable mobile connectivity, can be huge. There is already wide ranging moral imperative across society to prioritise safety above commercial gain and so the DGF fits into that scenario.

For commercial stakeholders, the task is to deliver satisfactory reliability to SCS at optimal costs. Doing this for all services and users may be prohibitively costly. Hence the DGF focus on only those services and users who fall into a tightly-defined SCS category. Mobile network providers already use PWS, a form of D-QoS, to deliver emergency alerts. During the Covid-19 pandemic, they equally used SMS to distribute several government messages[1]. The DGF calls for these to be extended and formalised so that they are applicable to all SCS and at all determined times.

7.3.4 Time Considerations

The time to implement the DGF is now. Across society and in many industrial segments, the promise of the 5G era is to use low latency and ultra reliable connectivity to drive transformation and create value. Yet, these opportunities are not going to materialize without an assurance of the reliability of the underlying connectivity. It is therefore imperative for action to happen now to ensure that the right framework, policies, incentives and legalities are in place to drive progress.

A key part of the timeliness regards energy use and environmental sustainability. Much of the ‘smartness’ in connected systems such as Smart Cities, Smart Transportation, Smart Manufacturing and Smart Retail can only be realized if the underlying connectivity is assured. It is therefore incumbent on all stakeholders to ensure that the principles of the DGF are widely disseminated and adopted across society.

[1]https://www.gov.uk/government/news/coronavirus-sms-messages

7.4 DGF Evaluation for a 999 CIS app

7.4.1 Introducing the 999 app

In [8], we introduced the concept of the 999 app, the first example in the literature - to the best of our knowledge - detailing such an app. The idea builds on the success of the 999 voice system which was set up in the UK in June 1937, as the first emergency services number in the world [2]. Since then, most countries have set up similar systems to support public safety initiatives (e.g. 112/911) and the 112/911/999 concept is now globally accepted.

Given the right technical solutions and a favourable policy framework, an app-based 999 Critical Information Systems (CIS) system would build on the architecture of existing 999 voice systems. Using the same architectural design (Figure 7.3 for the UK system), the 999 app will aggregate all relevant information and is given priority access in the network. The app aggregates all the relevant CIS content into a single portal, and can also integrate PWS systems. It will involve a content aggregator, operating at the local level, who is able to integrate relevant local information sources.

Unlike current Public Warning Systems (PWS), the 999 app could also include more multimedia content and with a lower threshold to accommodate both emergency alerts and a wider range of CIS messages. Current PWS systems can continue to be pushed-to-the-customer. But non-emergency CIS notices can be delivered to the app with no alert and the customer pulls the information whenever they open the app. Information delivery to the app can be via broadcast mechanism (similar to PWS/SMS) or they can be pulled via eXtensible Messaging and Presence Protocol (XMPP) or an Application Programming Interface (API).

Operationally, my design proposes to use Application specific Congestion control for Data Communication (ACDC) as the D-QoS mechanism for an officially-approved 999 app. In doing so, I consider three mechanisms to ensure that the 999 app can deliver CIS messages everywhere, and regardless of congestion.

1. Background updating: As it is originally specified to activate when a device is in ‘idle’ mode, ACDC will provide a mechanism to refresh the content of the app regularly during typical scenarios.

2. Priority during congestion: Using an enhanced ACDC which permits prioritisation for devices in ‘connected mode’ will ensure the app is given priority during congestion.

[2]http://home.bt.com/tech-gadgets/what-happens-when-you-call-999-the-secrets-of-the-emergency-services-number-11364191315763

3. Zero rating & data-off exempt: Similar to SMS, the 999 CIS app would be included as part of 3GPP 'Data Off Exempt Service' package [39] to ensure that it is operational even if users have switched off data or roaming internationally.

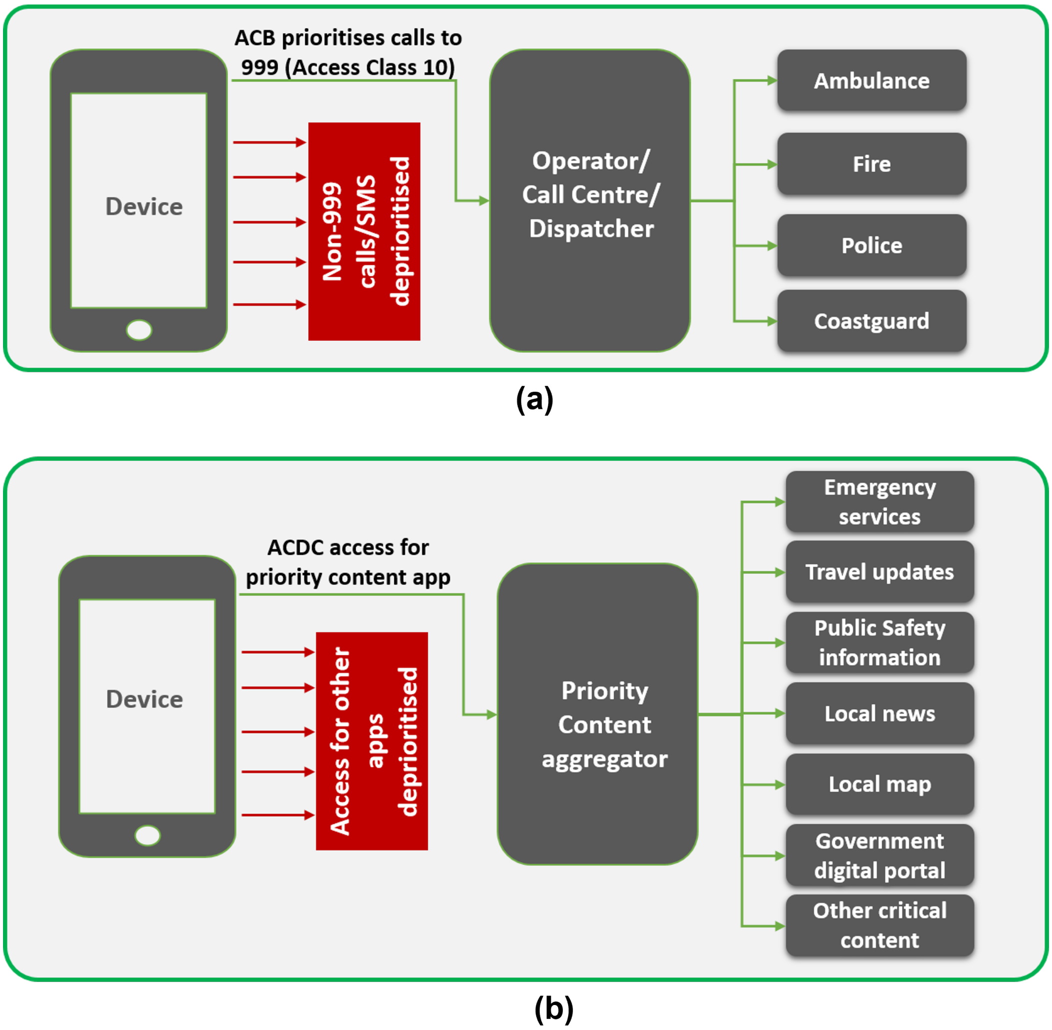


Fig. 7.3 Architectural comparison of a 999 app and the 999 call system. (a) Schematic design of the 999 call system in the UK. (b) Proposed design of a 999 CIS app, using Application specific Congestion control for Data Communication (ACDC) as the D-QoS mechanism.

7.4.2 Question 1: 999 app is a CIS and an SCS

Throughout history, every society has sought ways to communicate important and critical information to citizens. This is usually done using CIS which was described in Section 2.2.4. Increasingly, traditional CIS (i.e TV, radio, newspapers) are being supplanted or complemented by the internet, and mobile phones, either via SMS or through Public Warning Systems (PWS) (e.g. Cell Broadcast) [34]. This is because of better convenience or that the citizenry are increasingly disinterested in traditional information sources or are questioning the integrity of existing news providers. As such, society needs to evolve the distribution medium for critical/important information that is currently prioritised on TV, radio and newspaper (e.g. warning/emergency/safety information, traffic/weather news, governmental information). Recently, smartphones have emerged as an important channel for emergency and critical information because of their ubiquity and ability to support multimedia messages, and because they can be used for CIS at a lower threshold than that permissible for Public Warning Systems (PWS) - concerns about spamming users with irrelevant warnings means that the threshold to use PWS is set very high [36, 37]. [38] provides a general survey of smartphone systems for emergency management.

7.4.3 Question 2: 999 app needs mobile connectivity

By default, a 999 app is a mobile service and proposed on the assumption that it is a fully integrated mobile service. Accordingly, mobile connectivity, especially with high performance 4G/5G networks, ensures that the app can deliver an acceptable QoS to users.

7.4.4 Question 3: 999 app needs to work everywhere

Ubiquity of mobile communications ensures that a 999 app can be designed to benefit from a global, ubiquitous, and mostly uniform user experience across all countries. This means that the QoS assurance can be improved for the 999 app either via prioritising the app on public mobile networks or using multi-connectivity to at least two public mobile networks.

Global uniformity is a key design goal for a 999 app and I note in this PhD that there is the opportunity to bring together the wide variety of critical and emergency mobile and smartphone services and apps that are currently in the market into a single seamless service. In doing so, this will resolve the lack of a coordinated approach (whether at global, national or local levels) to validate and standardise a reliable smartphone system for CIS, and then educate users on what to expect (similar to how users are trained on what to expect when they dial 999/911/112 numbers). For example, there is no claim or expectation that the 112 Suomi, Alertable or 311 apps will always work in the event of congestion in a location. Similarly, there is no provision to zero rate the app to always make it accessible in 'Data off' circumstances [39]. Likewise, there is little awareness nor incentive for non-residents or visitors to download and use such app in these locations.

7.4.5 Question 4: 999 app plays a specific role

A 999 app is a specific, distinguishable service on smartphones and so, its data traffic can be prioritised on public mobile networks. This possibility has always been recognised in 3GPP, hence there are provisions in the QCI and 5QI schemas for Mission Critical Data. In Section 6.3.1 above, I already note that prioritisation for the 999 app can involve background updating of its data, granting priority to its data during congestion, and zero-rating its data so that it works during data-off and international roaming. Currently, the main Smartphone OS providers (Apple and Google) have made provisions to support emergency alerts on iOS and Android - a government approved 999 app will enjoy similar treatment on smartphones.

7.4.6 Question 5: 999 app should be officially enforced

Identifying and proposing for the 999 app to be prioritised does not automatically make it a reality. We chronicled this in [25] and in Chapter 5 of this PhD. Likewise, in my engagements with engineers managing live production networks, it is clear that application level access control for end user services is uncommon. This can be seen in the usage levels for QCI. Despite a choice of at least 15 QCI levels, most services on 4G networks use QCI Level 9 as default bearer; the only exception is Voice over LTE (VoLTE) which uses QCI Level 1. We are also unaware of any live production network using ACDC widely (perhaps because it was only finalised in 3GPP Release 13).

Therefore, this PhD is making a strong recommendation for the 999 app to be given official endorsement and enforcement in the same way that the 999 emergency voice service is. As outlined in Table 7.1, the 999 app meets all the CLASP criteria and as such, deserves to be officially backed by the authorities.

7.5 Chapter Summary

This Chapter starts with an introduction to the D-QoS Guidance Framework (DGF) to provide a step-by-step guide to stakeholders on how to select the D-QoS mechanism to use or mandate for SCS on public mobile networks. I then proposed the 999 app, as an example of a Critical Information System and used the DGF to show why the 999 app should be prioritised globally.

The Chapter provides a fitting finale to the discussions in this PhD work and practicalises the discussions in Chapters 2-6. I have now established that D-QoS is necessary because of the shortcomings in live 3G/4G/5G networks (Chapter 4). Following this, I have provided a critical examination that D-QoS can be used to address those shortcomings despite its

Table 7.1 CLASP Framework Analysis for 999 app

CLASP	**Description**	**Government-backed 999 app**
Critical	Generally accepted to be critical to the proper functioning of the digital society of the future	Yes - provides critical communications links similar to the already prioritised 999 voice call
Localised	Only prioritised within locality where relevant, with the consent of the people of the region	Yes - each country or region would have its own version of the 999 app
Authorised	Considered to be important, and authorized by a mandated government agency (e.g. Ofcom)	Yes - managed by a government agency and provides information that is already prioritised on radio and TV
Specific	Only specific functionalities of the service can be prioritised and not all services from a particular provider	Yes - can be restricted only to CIS services
Perishable	Offers timely information which expires in short durations	Yes - provides timely information for safe functioning of society
-	Prioritised?	Yes

challenging history (Chapter 5). Then, I have shown, via empirical measurements on live 4G networks, that a D-QoS implementation can deliver its benefits in a real live deployment (Chapter 6). And now, I have encapsulated this discussions into a guidance framework that can be adopted and implemented by commercial and policy stakeholders (Chapter 7). In Chapter 8, I will provide a summary of the work in this PhD and some recommendations on future research directions to further develop this work, on how to use D-QoS to provide reliability assurance for SCS.

Chapter 8

Summary, Limitations & Conclusion

E jewe ejewe, a gwa eze; a lawa alawa, agwakwa eze make adịghị ama ama (Igbo Proverb)

(At the start of a journey, inform the king; at the end, inform the king too just in case)

This report explores the rationale for, the empirical evidence for, the history of, and proposals on how to provide differentiated performance to different services in 5G era mobile networks. In the preceding seven chapters of this work, I have detailed the research and analysis conducted by me and and my colleagues/collaborators to tell the story of using D-QoS to improve connectivity QoS assurance for SCS. In this Chapter, I recap the originating thesis, research goals and research question for my work. I then provide a summary of each of the chapters, drawing out the takeaways and how the insights therein are addressing the research goals and questions. Finally, I provide some recommendations on future direction for this research and then share some final thoughts about my PhD experience.

8.1 PhD research in a nutshell

I summarise my PhD in five sequential points:

1. I identified that Society Critical Systems need QoS guarantees for their mobile network connectivity to effectively deliver to society's expectation.

2. While it is generally obvious that 3G/4G mobile networks can not provide such QoS guarantees, I conducted extensive field measurements in UK and South Korea to confirm and quantify this inadequacy on 3G, 4G, and crucially, 5G.

3. I acknowledge the role of D-QoS mechanisms since the 1970s as remedies for poor network performance for selected use cases. However, given their relative lack of

sustained commercial success, I conducted a critical review and historical analysis of D-QoS mechanisms to identify factors that can improve their market outcomes. Government backing emerged as the key factor and I conceptualised a set of criteria to guide governments on if and when to provide such backing.

4. Assuming that government backing is in place, I drove through over 800 kilometers of roads in UK to exemplify a D-QoS mechanism in action and empirically evaluate how the use of a redundant multi-operator connectivity option can improve QoS guarantees for connected cars in real-life and real-time.

5. Having empirically established the need for D-QoS for SCS, conceptually identified success criteria for D-QoS and empirically exemplified and evaluated a D-QoS mechanism in a real-life and real-time scenario, I proposed a conceptual guidance framework to support commercial and policy stakeholders in making decisions on if, when and how to use D-QoS for SCS. I illustrate the framework with a 999 app idea.

8.2 PhD report summary

8.2.1 Recap: Thesis, Research Goals & Research Questions

My overarching thesis in this PhD is that an officially-supported Differentiated QoS approach will provide improved QoS guarantees on public mobile networks for society critical systems in the 5G era.

I set three research goals for my research. First, investigate the adequacy of public mobile networks to provide QoS guarantees for society critical services. Second, explore the history of D-QoS approaches to understand if and how it can be used to provide QoS guarantees for SCS systems.Third, evaluate the reality of D-QoS approaches to understand their relative advantages in providing QoS guarantees for SCS.

To achieve the research goals, I identified key Research Questions for each goal:

1. RQ 1: Will capacity and design improvements to mobile networks be sufficient to deliver QoS assurance to SCS in the 5G era?

2. RQ 2: Given its relatively unsuccessful history and current net neutrality paradigm, could a D-QoS approach deliver a better QoS assurance for SCS?

3. RQ 3: If D-QoS is a feasible approach, what are the D-QoS options and implementation approach to ensure a more successful D-QoS reality?

During the course of my PhD, I interrogated my research goals and questions through series of field measurements, system implementation and a critical examination of D-QoS history.

8.2.2 Summary & Takeaways

1. Evaluating QoS assurance for SCS based on performance of mobile networks

The entire *raison d'être* of my PhD is premised on the inadequacy of the mobile network to provide QoS assurance for SCS use cases. In Chapter 4, I detailed three sets of extensive field measurements on 3G, 4G and 5G, to empirically proof the case. The measurements help to understand the nature of mobile network performance in normal/typical times, in abnormal/atypical times, at busy locations and across two countries (South Korea and UK). Experiment 1 was a four year longitudinal measurement study on 3G/4G in the London UK during the New Year fireworks celebration from 2016/17 to 2019/20. Experiment 2 was focused on 4G/5G measurements at the five busiest train stations in the UK in January 2020. Experiment 3 was on 4G/5G measurements at eight busy locations in UK and South Korea.

From the overall results, I highlight the latency figures as latency is a major selling point for the use of 5G for SCS. For experiment 1 in London UK between 2016 - 2020, we show that only 58% of RTT measurements on 4G would achieve a 50ms latency target. For experiment 2 in London UK in 2020, this changes to 27% on the 4G device but only 28% on the 5G device. For experiment 3 in London UK and Seoul South Korea in 2021, 55% of readings on the 4G device had RTT below 50ms whereas only 50% of readings on the 5G device achieved same.

The experimental findings provide strong evidence to answer RQ 1 that improvements in 3G, 4G and NSA 5G networks are inadequate to provide reliability/QoS assurance for SCS.

2. Delivering QoS assurance to SCS using D-QoS mechanisms

Given its less than stellar history, D-QoS mechanisms are an unconvincing solution to the problem of QoS assurance for SCS. This is partly because they have been treated mostly as a technical topic whereas their implementation, on public networks that work effectively as a zero sum game, is a classic economic/policy topic. Furthermore, there has being a distinct lack of clarity on how to classify D-QoS mechanisms, the criteria for their use and the conditions to guarantee their success.

In Chapter 5, I set out to fill these gaps in the industry's knowledge base. Through an extensive and critical examination of the history of D-QoS mechanisms, I classified them, and identified the conditions that need to exist for them to become commercial success stories. I then introduced a taxonomy to classify data traffic on mobile networks to codify which ones warrant different treatments. Finally, I introduced the CLASP framework to guide policy makers on how to determine which services deserve to be given official backing.

The analysis and frameworks in Chapter 5 provide good insights to address RQ 2 on how to use a D-QoS approach to deliver a better QoS assurance for SCS.

3. Field evaluation of a multi-connectivity D-QoS for CCs

Having proposed D-QoS mechanisms as an option for QoS assurance for SCS, it became imperative to text this in the field to validate the hypothesis. In Chapter 6, we selected a Connected Car as an example of an SCS that is becoming increasingly mainstream across society. We designed and built a system implementation of an onboard multi-operator connectivity module for a CC. Then we drove for over 800 kilometers across South East England, on both major and minor roads, to explore how connectivity varies along the route. Our goal was to establish if, and by how much, a multi-connectivity option, which is managed by the user on the demand side (DSM-MoC), will outperform a single connectivity option.

From our results, we show that the hypothetical delta between DSM-MoC of four operators versus the best single-operator is as high as 28 percentage points (pp) for RTT Variance, 20pp for downlink speed, 18pp for uplink speed and 9pp for RTT. For the actual system implementation, we show that a two operator DSM-MoC achieved a 12% better performance than the best single-operator in downloading a 1.3MB webpage hosted in AWS.

Our work shows that as edge nodes (e.g. cars) gain more computing power, it becomes attractive to repatriate decision making to the edge instead of relying on a central coordinator, an approach in contrast to the industry default where the design philosophy of the mobile network has encouraged dumb nodes and intelligent core.

The results in Chapter 6 provide empirical validation to partly address RQ 3 that a DSM-MoC for CC, as an example of a D-QoS for an SCS, will improve QoS assurance.

4. D-QoS Guidance Framework & enable official backing for 999 app

In the penultimate Chapter of this work, I strive to practicalise the insights from my work as a fitting finale to my PhD. I introduce the DGF framework for commercial and policy stakeholders who need to make decisions regarding D-QoS and SCS. As such, the DGF provides a step-by-step guide to such stakeholders on how to select the D-QoS mechanism to use or mandate for SCS on public mobile networks.

Bringing it to life, I then proposed the 999 app, as an example of a Critical Information System and used the DGF to show why the 999 app should be prioritised globally. To the best of my knowledge, our work on the 999 app is the first and only documented instance in the literature to make the case for a globally adopted, and officially backed 999 app for CIS to replicate the existing 999/119/911 emergency call system.

Together with the results in Chapter 6, the DGF and the 999 app provide a sufficiently complete answer to address RQ 3 on the options and implementation approach to ensure that D-QoS is able to deliver QoS assurance for SCS.

8.3 Limitations and future directions

This PhD breaks new grounds in arguing for a deeper consideration of political and economic factors, in addition to the standard technical factors, in determining how to support society critical systems. This is a prescient topic for society as the pace of digital transformation across vast swaths of society is inexorably increasing society's reliance on reliable connectivity. Researchers and practitioners will have fertile grounds to explore this research space to identify the approaches that work, unearth factors that are critical, and to design, implement and operationalise D-QoS mechanisms for SCS.

However, there are several limitations and unfinished business in this PhD plus new research directions which can be pursued by other researchers. I highlight five areas that need further work.

8.3.1 Insufficient unique experimental results

Although it is a good learning outcome to develop my research expertise in organising large scale and multi-country experiments, I still agonise over the relative non-uniqueness of many of the results from the measurements for mobile network performance. I continued to scale up the measurements in the forlorn hope that I will uncover new insights. Sadly, that did not materialise at scale. In hindsight, I should have been more cognisant of the reality that observational studies may not uncover much unique insight.

8.3.2 Incomplete evaluation of introduced concepts

Given a PhD's limited time frame, it has not been possible to comprehensively evaluate two of the concepts introduced in this PhD. The DGF and the 999 app idea deserve to be studied further and I believe they could become standalone PhD topics in their own right.

8.3.3 Gaps in understanding D-QoS

While this PhD takes pride in being the first comprehensive coverage in the literature of D-QoS mechanisms since the 1970s, it has barely touched our understanding of D-QoS. My expectation is that each of the mainstream D-QoS mechanisms ought to be studied deeply to unearth all the factors and nuances that influenced their performance.

8.3.4 SCS connectivity requirements

In this PhD, I have relied on generic guidance to determine the connectivity performance that SCS use cases require. This is an inadequate approach and I recommend that a much more detailed analysis is done for each SCS use case to determine its performance requirements. This will help to avoid unnecessary agitation for the use of D-QoS mechanisms in public networks.

8.3.5 Reusing data in analytical models

Despite the limitations in this work, it is also clear that the data that has been developed can be of great value in building analytical models for further research, especially for Intelligent Transport Systems. This will enable other researchers to understand the possibilities and limitations of existing cellular connectivity on UK road networks. Accordingly, I commit to make the data available to other researchers via the appropriate mechanism provided by King's College London.

8.3.6 Increased cross disciplinary approach

It was obvious to me during this PhD that a greater cross disciplinary approach, covering technical, economics, policy and social factors, is needed to deliver optimal outcomes for society for digital services. Despite having an above average commercial background than typical PhD students, I recognize that my PhD has barely scratched the surface. I encourage researchers to collaborate more with other disciplines to improve research outcomes.

8.4 Final Remarks

My PhD journey has offered me an intellectually stimulating and practically rewarding experience. It afforded me the opportunity to work with like minded professionals to make a contribution to society. While studying part time, along with my full time day job was challenging, it also paved the way for a cross fertilisation of ideas. This allowed me to bring practical insights from industry to academia and enabled me to take cutting edge academic insights back to industry. Such mutually symbiotic research experience enabled me to practicalise my research so that it is oven baked for adoption by commercial and policy stakeholders globally. In my academic and industry career, I will continue to champion these thoughts and promote a future where the telecommunications infrastructure can deliver reliable connectivity and QoS assurance for society critical systems.

1

References

[1] E. Obiodu and M. Giles, “The 5g era: age of boundless connectivity and intelligent automation,” *GSM Assocation*, 2017.

[2] E. Obiodu, “The 5G Guide: a reference for operators,” *GSM Assocation*, 2019.

[3] Ericsson, “5G for business: a 2030 market compass,” *Ericsson*, 2019.

[4] M. Hinchey and L. Coyle, “Evolving critical systems: A research agenda for computer-based systems,” in *2010 17th IEEE International Conference and Workshops on Engineering of Computer Based Systems*, pp. 430–435, IEEE, 2010.

[5] M. Rausand, *Reliability of safety-critical systems: theory and applications*. John Wiley & Sons, 2014.

[6] K. Fowler, “Mission-critical and safety-critical development,” *IEEE Instrumentation & Measurement Magazine*, vol. 7, no. 4, pp. 52–59, 2004.

[7] S. Torabi, H. R. Soufi, and N. Sahebjamnia, “A new framework for business impact analysis in business continuity management (with a case study),” *Safety Science*, vol. 68, pp. 309–323, 2014.

[8] E. Obiodu, N. Sastry, and A. Raman, “Is it time for a 999-like (or 112/911) system for critical information services*f*,” in *NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium*, pp. 1–6, IEEE, 2020.

[9] T. Novak and A. Gerstinger, “Safety-and security-critical services in building automation and control systems,” *IEEE Transactions on Industrial Electronics*, vol. 57, no. 11, pp. 3614–3621, 2009.

[10] A. O. Arewa, S. Theophilus, A. Ifelebuegu, and P. Farrell, “Analysis of penalties imposed on organisations for breaching safety and health regulations in the united kingdom,” *Safety and health at work*, vol. 9, no. 4, pp. 388–397, 2018.

[11] C. D. Bârcă, D. Sorin, C. M. Bârcă, *et al.*, “Radio data system applications,” *Journal of Information Systems & Operations Management*, vol. 5, no. 1, pp. 138–143, 2011.

[12] Ericsson, “Ericsson mobility report,” November 2021.

[13] Q. C. Li, H. Niu, A. T. Papathanassiou, and G. Wu, “5g network capacity: Key elements and technologies,” *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 71–78, 2014.

[14] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, “What will 5g be?,” *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[15] N. T. Le, M. A. Hossain, A. Islam, D.-y. Kim, Y.-J. Choi, and Y. M. Jang, “Survey of promising technologies for 5g networks,” *Mobile information systems*, vol. 2016, 2016.

[16] Rewheel, “Capacity utilization and fixed-to-mobile broadband substitution potential – a study of 64 european operators,” 2017.

[17] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, “Understanding traffic dynamics in cellular data networks,” in *2011 Proceedings IEEE INFOCOM*, pp. 882–890, IEEE, 2011.

[18] J. Erman and K. Ramakrishnan, “Understanding the super-sized traffic of the super bowl,” in *Proceedings of the 2013 conference on Internet measurement conference*, pp. 353–360, ACM, 2013.

[19] S. Sonntag, L. Schulte, and J. Manner, “Mobile network measurements-it’s not all about signal strength,” in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 4624–4629, IEEE, 2013.

[20] T. Doumi, M. F. Dolan, S. Tatesh, A. Casati, G. Tsirtsis, K. Anchan, and D. Flore, “Lte for public safety networks,” *IEEE Communications Magazine*, vol. 51, no. 2, pp. 106–112, 2013.

[21] I. SG05, “Draft new report itu-r m.[imt-2020. tech perf req]-minimum requirements related to technical performance for imt-2020 radio interface (s),” *ITU-R SG05 Contribution*, vol. 40, 2017.

[22] ITU-R, “ITU-R M.[IMT-2020.TECH PERF REQ] - Minimum requirements related to technical performance for IMT-2020 radio interface(s)/ITU-R SG05 Contribution 40,” tech. rep., Feb 2017.

[23] A. Ghosh, A. Maeder, M. Baker, and D. Chandramouli, “5G evolution: A view on 5G cellular technology beyond 3GPP release 15,” *IEEE Access*, vol. 7, pp. 127639–127651, 2019.

[24] D. C. Odikayor, I. Oghogho, S. T. Wara, and A.-A. Adebayo, “Dual-SIM Phones: A Disruptive Technology?,” in *Disruptive Technologies, Innovation and Global Redesign: Emerging Implications*, pp. 462–469, IGI Global, 2012.

[25] E. Obiodu and N. Sastry, “From atm to mpls and qci: The evolution of differentiated qos standards and implications for 5g network slicing,” *IEEE Communications Standards Magazine*, vol. 4, no. 2, pp. 14–21, 2020.

[26] A. Maurya and D. Kumar, “Reliability of safety-critical systems: A state-of-the-art review,” *Quality and Reliability Engineering International*, vol. 36, no. 7, pp. 2547–2568, 2020.

[27] R. Bell, “Safety critical systems – a brief history of the development of guidelines and standards (parts 1 - 3),” 2021.

[28] R. Bell, “Introduction to iec 61508,” in *Acm international conference proceeding series*, vol. 162, pp. 3–12, 2006.

[29] D. Negi, N. Bagri, and V. Agarwal, “Redundancy for safety-compliant automotive & other devices,” *EDN network*, 2014.

[30] IEC, “IEC 61508: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems,” 2010.

[31] British Continuity Institute, “A Guide to Global Good Practice in Business Continuity,” 2013.

[32] S. Tangen and D. Austin, “Business Continuity ISO 22301 when things go seriously wrong,” *ISO Focus*, june 2012.

[33] O. Standard, “Common alerting protocol version 1.2,” *Jul*, vol. 1, pp. 1–47, 2010.

[34] A. Aloudat and K. Michael, “The application of location based services in national emergency warning systems: Sms, cell broadcast services and beyond,” 2011.

[35] N. A. of Sciences Engineering and Medicine, “Emergency alert and warning systems: Current knowledge and future research directions,” 2018.

[36] M. M. Hossain, M. Sharmin, and S. Ahmed, “Understanding the feasibility of a location-aware mobile-based 911-like emergency service in bangladesh,” in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, vol. 1, pp. 861–868, IEEE, 2017.

[37] M. Jovanovic, I. Babic, M. Cabarkapa, J. Misic, S. Mijalkovic, V. Nikolic, and D. Randjelovic, “Soserbia: Android-based software platform for sending emergency messages,” *Complexity*, vol. 2018, 2018.

[38] H. Maryam, M. A. Shah, Q. Javaid, and M. Kamran, “A survey on smartphones systems for emergency management (spsem),” *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 6, pp. 301–311, 2016.

[39] 3rd Generation Partnership Project, “Study on 3GPP PS Data Off # 23.702 Release 14,” tech. rep., December 2016.

[40] L. Fraccascia, I. Giannoccaro, and V. Albino, “Resilience of complex systems: state of the art and directions for future research,” *Complexity*, vol. 2018, 2018.

[41] A. Clark-Ginsberg, “What’s the difference between reliability and resilience?,” *Department of Homeland Security. March*, 2016.

[42] G. Liu, Y. Huang, Z. Chen, L. Liu, Q. Wang, and N. Li, “5G Deployment: Standalone vs. Non-Standalone from the Operator Perspective,” *IEEE Communications Magazine*, vol. 58, no. 11, pp. 83–89, 2020.

[43] J. Lee, Y. Kim, Y. Kwak, J. Zhang, A. Papasakellariou, T. Novlan, C. Sun, and Y. Li, "LTE-advanced in 3GPP Rel-13/14: an evolution toward 5G," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 36–42, 2016.

[44] G. Barb and M. Otesteanu, "4G/5G: A comparative study and overview on what to expect from 5G," in *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, pp. 37–40, IEEE, 2020.

[45] B. Gopal and P. Kuppusamy, "A comparative study on 4G and 5G technology for wireless applications," *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)*, vol. 10, no. 6, pp. 67–72, 2015.

[46] GSMA 5G Status Indicator Taskforce, "LS Reply to 3GPP RAN on Requirements for NR carrier frequency mismatch in 5G status indication; RP-193053," tech. rep., 2019.

[47] M. S. Islam, M. Kamruzzaman, T. Jessy, M. S. Zahan, and M. S. Hassan, "Performance analysis of massive mimo for 5g wireless communication systems," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 1579–1583, IEEE, 2016.

[48] E. C. Strinati and S. Barbarossa, "6g networks: Beyond shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, 2021.

[49] C. Zhang, Y.-L. Ueng, C. Studer, and A. Burg, "Artificial intelligence for 5g and beyond 5g: Implementations, algorithms, and optimizations," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 10, no. 2, pp. 149–163, 2020.

[50] GSMA, "5G Spectrum: GSMA Public Policy Position ," *GSM Assocation*, March 2021.

[51] A. Mihovska, R. Prasad, *et al.*, "Overview of 5g new radio and carrier aggregation: 5g and beyond networks," in *2020 23rd International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pp. 1–6, IEEE, 2020.

[52] A. Research, "Mobile backhaul options: Spectrum analysis and recommendations," *GSM Assocation*, September 2018.

[53] S. Kavanagh, "5G vs 4G: No Contest," 2020. [Available online] https://5g.co.uk/guides/4g-versus-5g-what-will-the-next-generation-bring/; [Accessed] 11/12/2020.

[54] G. Soós, D. Ficzere, P. Varga, and Z. Szalay, "Practical 5G KPI measurement results on a non-standalone architecture," in *Noms 2020-2020 IEEE/IFIP network operations and management symposium*, pp. 1–5, IEEE, 2020.

[55] A. Narayanan, E. Ramadan, J. Carpenter, Q. Liu, Y. Liu, F. Qian, and Z.-L. Zhang, "A First Look at Commercial 5G Performance on Smartphones," in *Proceedings of The Web Conference 2020*, pp. 894–905, 2020.

[56] D. Xu, A. Zhou, X. Zhang, G. Wang, X. Liu, C. An, Y. Shi, L. Liu, and H. Ma, "Understanding operational 5g: A first measurement study on its coverage, performance and energy consumption," in *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, pp. 479–494, 2020.

[57] E. Obiodu, A. K. Abubakar, and N. Sastry, "Is it 5G or not? Investigating doubts about the 5G icon and network performance," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–6, IEEE, 2021.

[58] M. Dano, "Your 5G Icon Is Virtually Meaningless," 2019. [Available online] https://www.lightreading.com/mobile/5g/your-5g-icon-is-virtually-meaningless/a/d-id/751238; [Accessed] 24/12/2020.

[59] 3rd Generation Partnership Project, "Study on Small Cell enhancements for E-UTRA and E-UTRAN; Higher layer aspects # 36.842 v12.0.0," tech. rep., 2013.

[60] 3rd Generation Partnership Project, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification # 36.331 v15.3.0 Release 15 ," tech. rep., 2018.

[61] 3rd Generation Partnership Project, "Clarification of the UpperLayerIndication in SIB2 ," tech. rep., 2018.

[62] R. Silva, P. Carvalho, P. Sousa, and P. Neves, "Heterogeneous mobility in next generation devices: An android-based case study," in *International Conference on Mobile Multimedia Communications*, pp. 316–330, Springer, 2010.

[63] Google, "TelephonyManager," 2020. [Available online] https://developer.android.com/reference/android/telephony; [Accessed] 24/12/2020.

[64] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *2011 Proceedings IEEE INFOCOM*, pp. 882–890, April 2011.

[65] M. El-Sayed, A. Mukhopadhyay, C. Urrutia-Valdés, and Z. J. Zhao, "Mobile data explosion: Monetizing the opportunity through dynamic policies and qos pipes," *Bell Labs Technical Journal*, vol. 16, no. 2, pp. 79–99, 2011.

[66] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "Characterizing and optimizing cellular network performance during crowded events," *IEEE/ACM Transactions on Networking*, vol. 24, pp. 1308–1321, June 2016.

[67] J. Erman and K. Ramakrishnan, "Understanding the super-sized traffic of the super bowl," in *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, (New York, NY, USA), pp. 353–360, ACM, 2013.

[68] S. Sonntag, J. Manner, and L. Schulte, "Netradar-measuring the wireless world," in *2013 11th International Symposium and Workshops on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, pp. 29–34, IEEE, 2013.

[69] M.-R. Fida, A. Ocampo, and A. Elmokashfi, “Measuring and localising congestion in mobile broadband networks,” *IEEE Transactions on Network and Service Management*, 2021.

[70] A. Elnashar, “Practical aspects of lte network design and deployment,” *arXiv preprint arXiv:1810.02970*, 2018.

[71] D. Baltrunas, A. Elmokashfi, and A. Kvalbein, “Measuring the reliability of mobile broadband networks,” in *Proceedings of the 2014 internet measurement conference*, pp. 45–58, ACM, 2014.

[72] V. Petrov, M. A. Lema, M. Gapeyenko, K. Antonakoglou, D. Moltchanov, and et al, “Achieving end-to-end reliability of mission-critical traffic in softwarized 5g networks,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 485–501, 2018.

[73] C. Campolo, A. Molinaro, A. Iera, and F. Menichella, “5G network slicing for vehicle-to-everything services,” *IEEE Wireless Communications*, vol. 24, no. 6, pp. 38–45, 2017.

[74] I. Labriji, F. Meneghello, D. Cecchinato, and et al, “Mobility aware and dynamic migration of mec services for the internet of vehicles,” *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 570–584, 2021.

[75] E. Obiodu, N. Sastry, and A. Raman, “Clasp: a 999-style priority lanes framework for 5g-era critical data services,” in *2019 International Symposium ELMAR*, pp. 101–104, IEEE, 2019.

[76] Q. C. Li, H. Niu, A. T. Papathanassiou, and G. Wu, “5g network capacity: Key elements and technologies,” *IEEE Vehicular Technology Magazine*, vol. 9, pp. 71–78, March 2014.

[77] V. Stocker and J. Whalley, “Speed isn’t everything: a multi-criteria analysis of broadband access speeds in the UK,” 27th European Regional ITS Conference, Cambridge (UK) 2016 148707, International Telecommunications Society (ITS), 2016.

[78] M. Belshe, “More bandwidth doesn’t matter (much),” *Google Inc*, 2010.

[79] Ofcom, “Connected Nations 2017, UK report,” tech. rep., 2017.

[80] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, “Characterizing and optimizing cellular network performance during crowded events,” *IEEE/ACM Transactions on Networking (TON)*, vol. 24, no. 3, pp. 1308–1321, 2016.

[81] D. Wu, W. Li, R. K. Chang, and D. Gao, “Mopeye: Monitoring per-app network performance with zero measurement traffic,” *arXiv preprint arXiv:1610.01282*, 2016.

[82] J. Huang, C. Chen, Y. Pei, Z. Wang, Z. Qian, F. Qian, B. Tiwana, Q. Xu, Z. Mao, M. Zhang, *et al.*, “Mobiperf: Mobile network measurement system,” *Technical Report. University of Michigan and Microsoft Research*, 2011.

[83] N. Vallina-Rodriguez, A. Auçinas, M. Almeida, Y. Grunenberger, K. Papagiannaki, and J. Crowcroft, "Rilanalyzer: a comprehensive 3g monitor on your phone," in *Proceedings of the 2013 IMC conference*, pp. 257–264, 2013.

[84] U. Goel, M. P. Wittie, and K. C. Claffy, "Survey of end-to-end mobile network measurement testbeds, tools, and services," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 105–123, 2015.

[85] W. Li, R. K. Mok, D. Wu, and R. K. Chang, "On the accuracy of smartphone-based mobile network measurement," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 370–378, IEEE, 2015.

[86] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Smart data pricing (sdp): Economic solutions to network congestion," *Recent advances in networking*, vol. 1, pp. 221–274, 2013.

[87] S. Keshav, *An engineering approach to computer networking: ATM networks, the Internet, and the telephone network*, vol. 1. Addison-Wesley Reading, 1997.

[88] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Smart data pricing: Using economics to manage network congestion," *Commun. ACM*, vol. 58, pp. 86–93, Nov. 2015.

[89] M.-J. Sheng, C. Joe-Wong, S. Ha, F. M. F. Wong, and S. Sen, "Smart data pricing: Lessons from trial planning," in *INFOCOM, 2013 Proceedings IEEE*, pp. 3327–3332, IEEE, 2013.

[90] A. Sathiaseelan, G. Tyson, and S. Sen, "Exploring the role of smart data pricing in enabling affordable internet access," in *2015 IEEE Conference on Computer Communications Workshops*, pp. 558–563, April 2015.

[91] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "A survey of smart data pricing: Past proposals, current plans, and future trends," *ACM Comput. Surv.*, vol. 46, pp. 15:1–15:37, Nov. 2013.

[92] P. Maille, G. Simon, and B. Tuffin, "Toward a net neutrality debate that conforms to the 2010s," *IEEE Communications Magazine*, vol. 54, pp. 94–99, March 2016.

[93] M. Geddes, "2006-2016: A "net neutrality" retrospective," April.

[94] S. Jordan, "Traffic management and net neutrality in wireless networks," *IEEE Transactions on Network and Service Management*, vol. 8, pp. 297–309, December 2011.

[95] J. Crowcroft, "Net neutrality: The technical side of the debate: a white paper," *SIGCOMM Comput. Commun. Rev.*, vol. 37, pp. 49–56, Jan. 2007.

[96] C. Joe-Wong, S. Ha, and M. Chiang, "Sponsoring mobile data: An economic analysis of the impact on users and content providers," in *Computer Communications (INFOCOM), 2015 IEEE Conference on*, pp. 1499–1507, IEEE, 2015.

[97] European Union, "Regulation (EU) 2015/2120 of the European Parliament and of the Council," tech. rep., November 2015.

[98] BEREC, "BEREC Guidelines on the Implementation by National Regulators of European Net Neutrality Rules," tech. rep., August 2016.

[99] BEREC, "BEREC Opinion for the evaluation of the application of Regulation (EU) 2015/2120 and the BEREC Net Neutrality Guidelines," tech. rep., December 2018.

[100] S. Ely, "Rds-alert: a drive project to develop a proposed standard for the traffic message channel feature of the radio data system rds," in *Car and its Environment-What DRIVE and PROMETHEUS Have to Offer, IEE Colloquium on*, pp. 8–1, IET, 1990.

[101] Y. Yiakoumis, S. Katti, and N. McKeown, "Neutral net neutrality," in *Proceedings of the 2016 conference on ACM SIGCOMM 2016 Conference*, pp. 483–496, ACM, 2016.

[102] W3C, "Network Information API," tech. rep., April 2014.

[103] C. Pappas, K. Argyraki, S. Bechtold, and A. Perrig, "Transparency instead of neutrality," in *Proceedings of the 14th ACM Workshop on Hot Topics in Networks*, p. 22, ACM, 2015.

[104] J. P. Sterbenz, D. Hutchison, E. K. Çetinkaya, and et al, "Redundancy, diversity, and connectivity to achieve multilevel network resilience, survivability, and disruption tolerance invited paper," *Telecommunication Systems*, vol. 56, no. 1, pp. 17–31, 2014.

[105] J. Downer, *When failure is an option: Redundancy, reliability and regulation in complex technical systems*. No. DP 53, Centre for Analysis of Risk and Regulation, London School of Economics and ..., 2009.

[106] P. Popov, L. Strigini, J. May, and S. Kuball, "Estimating bounds on the reliability of diverse systems," *IEEE Transactions on Software Engineering*, vol. 29, no. 4, pp. 345–359, 2003.

[107] M. Randles, D. Lamb, E. Odat, and A. Taleb-Bendiab, "Distributed redundancy and robustness in complex systems," *Journal of Computer and System Sciences*, vol. 77, no. 2, pp. 293–304, 2011.

[108] C. Luciani, "From MPLS to SD-WAN: Opportunities, Limitations and Best Practices," 2019.

[109] J. L. Romeu, "Understanding series and parallel systems reliability," *Selected Topics in Assurance Related Technologies (START), Department of Defense Reliability Analysis Center (DoD RAC)*, vol. 11, no. 5, pp. 1–8, 2004.

[110] K. Aggarwal, Y. Chopra, and J. Bajwa, "Capacity consideration in reliability analysis of communication systems," *IEEE Transactions on Reliability*, vol. 31, no. 2, pp. 177–181, 1982.

[111] E. Gustafsson and A. Jonsson, "Always best connected," *IEEE Wireless communications*, vol. 10, no. 1, pp. 49–55, 2003.

[112] Y. Li, H. Deng, C. Peng, and et al, "icellular: Device-customized cellular network access on commodity smartphones," in *13th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 16)*, pp. 643–656, 2016.

[113] E. Obiodu, A. Raman, A. Abubakar, S. Mangiante, N. Sastry, and H. Aghvami, "Dsm-moc as baseline: Reliability assurance via redundant cellular connectivity in connected cars," *IEEE Transactions on Network and Service Management*, 2022.

[114] GSMA, "eSIM Whitepaper," *GSM Assocation*, 2018.

[115] T. Faisal, D. Di Francesco Maesa, N. Sastry, and S. Mangiante, "AJIT: Accountable Just-in-Time Network Resource Allocation with Smart Contracts," in *Proceedings of the ACM MobiArch 2020 The 15th Workshop on Mobility in the Evolving Internet Architecture*, pp. 48–53, 2020.

[116] Q. Xiao, K. Xu, D. Wang, L. Li, and et al, "TCP performance over mobile networks in high-speed mobility scenarios," in *2014 IEEE 22nd International Conference on Network Protocols*, IEEE, 2014.

[117] L. Li, K. Xu, T. Li, and et al, "A measurement study on multi-path tcp with multiple cellular carriers on high speed rails," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pp. 161–175, 2018.

[118] C. Xu, J. Wang, Z. Ma, and et al, "A first look at disconnection-centric TCP performance on high-speed railways," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 12, 2020.

[119] D. of Health UK, "Health technical memorandum 06-01: Electrical services supply and distribution," 2017.

[120] R. Oorni and A. Goulart, "In-vehicle emergency call services: ecall and beyond," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 159–165, 2017.

[121] G. I. Association, "5G Automotive Vision," *Oct 20th*, 2015.

[122] X. Wang, S. Mao, and M. X. Gong, "An overview of 3gpp cellular vehicle-to-everything standards," *GetMobile: Mobile Computing and Communications*, vol. 21, no. 3, pp. 19–25, 2017.

[123] R. Ramnath, N. Kinnear, S. Chowdhury, and T. Hyatt, "Interacting with android auto and apple carplay when driving: The effect on driver performance," *IAM RoadSmart Published Project Report PPR948*, 2020.

[124] M. Wisniewski, D. Bryla, and A. Koziowski, "Connected car: Challenges & opportunities for automotive industry," *Grape Up*, 2020.

[125] C. Bertoncello, C. Martens, T. Möller, and T. Schneiderbauer, "Unlocking the full life-cycle value from connected-car data," *McKinsey & Company*, 2021.

[126] C. E. Andrade, S. D. Byers, V. Gopalakrishnan, E. Halepovic, D. J. Poole, and et al, "Connected cars in cellular network: a measurement study," in *Proceedings of the 2017 Internet Measurement Conference*, pp. 235–241, ACM, 2017.

[127] C. Raiciu, C. Paasch, S. Barre, and et al, "How hard can it be? designing and implementing a deployable multipath TCP," in *9th USENIX symposium on networked systems design and implementation (NSDI 12)*, pp. 399–412, 2012.

[128] R. Molina-Masegosa and J. Gozalvez, “LTE-V for sidelink 5G V2X vehicular communications: A new 5G technology for short-range vehicle-to-everything communications,” *IEEE Vehicular Technology Magazine*, vol. 12, no. 4, pp. 30–39, 2017.

[129] T. Lecklider, “Autonomous mining equipment years ahead of car development,” *RN (editor)*, 2018.

[130] H. England, “Connecting the Country: Planning for the long term,” *UK Government Asset Publishing*, December 2017.

[131] S. S. Husain, A. Kunz, A. Prasad, and et al, “Ultra-high reliable 5G V2X communications,” *IEEE Communications Standards Magazine*, vol. 3, no. 2, pp. 46–52, 2019.

[132] C. Blackman and S. Forge, *Roaming: One Year After Implementation*. European Parliament, 2018.

[133] Softgearko, “Github/MultiPing-for-Android,” 2015. [Accessed] 11/12/2020; https://github.com/softgearko/MultiPing-for-Android.

[134] S. D. Strowes, “Passively measuring tcp round-trip times,” *Communications of the ACM*, vol. 56, no. 10, pp. 57–64, 2013.

[135] Boyland, Peter, “State of the mobile network experience,” tech. rep., 2019.

[136] B. Krenik, “4G wireless technology: When will it happen? What does it offer?,” in *2008 IEEE Asian Solid-State Circuits Conference*, pp. 141–144, IEEE, 2008.

[137] G. Intelligence, “Global Mobile Economy 2022,” *GSM Assocation*, 2022.

[138] G. H. Carvalho, I. Woungang, A. Anpalagan, and I. Traore, “When agile security meets 5g,” *IEEE Access*, vol. 8, pp. 166212–166225, 2020.

[139] Ofcom, “Connected Nations 2020, UK report,” tech. rep., 2020.

[140] P. Schmitt, M. Vigil, and E. Belding, “A study of mvno data paths and performance,” in *International Conference on Passive and Active Network Measurement*, pp. 83–94, Springer, 2016.

[141] A. Xiao, Y. Liu, Y. Li, F. Qian, Z. Li, S. Bai, Y. Liu, T. Xu, and X. Xin, “An in-depth study of commercial mvno: Measurement and optimization,” in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 457–468, 2019.

[142] GSA, “5G - 5G Networks December 2021,” 2022. [Available online] https://gsacom.com/paper/5g-market-snapshot-2021-end-of-year/; [Accessed] 04/05/2022.

[143] W. Webb, *The 5G Myth: When vision decoupled from reality*. Walter de Gruyter GmbH & Co KG, 2018.

[144] Office of Rail and Road, “Estimates of Station usage 2017-18 key facts,” 2018. [Available online] https://www.orr.gov.uk/media/16283; [Accessed] 11/12/2020.

[145] Qualcomm, “Global Mobile Industry Leaders Commit to Accelerate 5G NR for Large-scale Trials and Deployments,” February 2017. [Available online] https://www.qualcomm.com/news/releases/2017/02/25/global-mobile-industry-leaders-commit-accelerate-5g-nr-large-scale-trials; [Accessed] 24/12/2020.

[146] T. Bradley, “Not so fast: fake 4G networks,” 2011. [Available online] https://www.computerworld.com/article/2470789/not-so-fast–fake-4g-networks.html; [Accessed] 24/12/2020.

[147] P. Iacopino and A. Popov, “Analysing the behaviour of early 5G users: 10 things to know when planning consumer 5G strategies,” *GSM Assocation*, December 2021.

[148] W. Webb, “5G and Napoleon,” 2022. [Available online] https://www.linkedin.com/pulse/5g-napoleon-william-webb/?trackingId=b7IqYkdoAHxhM58iUEHnjQ%3D%3D; [Accessed] 12/05/2022.

[149] A. Orlowski, “5G is in danger of resembling Napoleon’s retreat from Moscow,” 2022. [Available online] https://www.telegraph.co.uk/business/2022/05/09/5g-danger-resembling-napoleons-retreat-moscow/?fr=operanews; [Accessed] 12/05/2022.

[150] R. Clark, “5G offers little value for the traditional telco,” 2021. [Available online] https://www.lightreading.com/asia/5g-offers-little-value-for-traditional-telco/d/d-id/770886; [Accessed] 03/09/2021.

[151] M. Mroček, “Github/NetMonster Core,” 2019. [Available online] https://github.com/mroczis/netmonster-core; [Accessed] 03/09/2021.

[152] A. Aijaz, “Private 5G: The future of industrial wireless,” *IEEE Industrial Electronics Magazine*, vol. 14, no. 4, pp. 136–145, 2020.

[153] W. Son and M.-s. Park, “The government role in digital era innovation: the case of electronic authentication policy in Korea,” *International Commerce and Information Review*, vol. 19, no. 4, pp. 29–50, 2017.

[154] W. Son, *Policy Success and Failure in the Digital Era: The Case of Korea’s National Public Key Infrastructure for Digital Authentication*. PhD thesis, State University of New York at Stony Brook, 2019.

[155] Andersen, Dave, “5G in South Korea: widespread 5G access plus incredible speeds,” tech. rep., June 2021.

[156] E. Obiodu, A. Raman, and N. Sastry, “Towards a taxonomy of differentiated service classes for the 5G era,” *Proceedings of the IEEE 5G World Forum*, July 2018.

[157] A. Nasrallah, A. S. Thyagaturu, Z. Alharbi, C. Wang, X. Shao, M. Reisslein, and H. ElBakoury, “Ultra-low latency (ull) networks: The ieee tsn and ietf detnet standards and related 5g ull research,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 88–145, 2018.

[158] GSA, “TR-221: Technical Specifications for MPLS in Mobile Backhaul Networks,” 2011. [Available online] https://docplayer.net/2827157-Tr-221-technical-specifications-for-mpls-in-mobile-backhaul-networks.html; [Accessed] 04/05/2022.

[159] 3rd Generation Partnership Project, "Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2 (Release 15) # 23.501 V15.3.0 (2018-09)," tech. rep., 2017.

[160] X. Xiao and L. M. Ni, "Internet QoS: a big picture," *IEEE network*, vol. 13, no. 2, pp. 8–18, 1999.

[161] X. Xiao, *Technical, commercial and regulatory challenges of QoS: An internet service model perspective*. Morgan Kaufmann, 2008.

[162] A. Malik, J. Qadir, B. Ahmad, K.-L. A. Yau, and U. Ullah, "QoS in IEEE 802.11-based wireless networks: a contemporary review," *Journal of Network and Computer Applications*, vol. 55, pp. 24–46, 2015.

[163] B. Teitelbaum and S. Shalunov, "Why premium IP service has not deployed (and probably never will)," *Internet2 QoS Working Group Informational Document*, 2002.

[164] D. Clark, "The design philosophy of the DARPA internet protocols," *ACM SIGCOMM Computer Communication Review*, vol. 18, no. 4, pp. 106–114, 1988.

[165] P. Flavius and P. Ferdi, "Internet service delivery models: evolution and current issues," in *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2011 International Conference on*, pp. 146–153, IEEE, 2011.

[166] A. Odlyzko, "The history of communications and its implications for the internet," 2000.

[167] S. Shalunov, G. Hazel, J. Iyengar, and M. Kuehlewind, "Low extra delay background transport (ledbat)," tech. rep., 2012.

[168] L. Skorin-Kapov and M. Matijasevic, "Analysis of qos requirements for e-health services and mapping to evolved packet system qos classes," *Int. J. Telemedicine Appl.*, vol. 2010, pp. 9:1–9:18, Jan. 2010.

[169] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, *et al.*, "Network slicing to enable scalability and flexibility in 5g mobile networks," *IEEE Communications magazine*, vol. 55, no. 5, pp. 72–79, 2017.

[170] C. T. Marsden, "Zero rating and mobile net neutrality," in *Net Neutrality Compendium*, pp. 241–260, Springer, 2016.

[171] A. M. Kakhki, F. Li, D. Choffnes, E. Katz-Bassett, and A. Mislove, "Bingeon under the microscope: Understanding t-mobiles zero-rating implementation," in *Proceedings of the 2016 workshop on QoE-based Analysis and Management of Data Communication Networks*, pp. 43–48, ACM, 2016.

[172] D. J. Kuss and M. D. Griffiths, "Online social networking and addiction: a review of the psychological literature," *International journal of environmental research and public health*, vol. 8, no. 9, pp. 3528–3552, 2011.

[173] M. Mackenzie, "The automotive sector will drive the demand for 5G," *Analysys Mason*, 2019.

[174] E. Obiodu, A. Abdullahi, A. Raman, N. Sastry, and S. Mangiante, “To share or not to share: reliability assurance via redundant cellular connectivity in connected cars,” in *2021 IEEE/ACM 29th International Symposium on Quality of Service*, IEEE, 2021.

[175] T. V. Nguyen, P. Shailesh, B. Sudhir, G. Kapil, L. Jiang, Z. Wu, and et al, “A comparison of cellular vehicle-to-everything and dedicated short range communication,” in *2017 IEEE Vehicular Networking Conference (VNC)*, pp. 101–108, IEEE, 2017.

[176] M. Gligorijević, M. Bošković, and A. Maksimović, “The role and importance of the “national roaming” in emergencies,” *Nauka, bezbednost, policija*, vol. 22, no. 1, pp. 111–126, 2017.

[177] T. Braud, T. Kämäräinen, M. Siekkinen, and P. Hui, “Multi-carrier measurement study of mobile network latency: The tale of hong kong and helsinki,” in *2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*, pp. 1–6, IEEE, 2019.

[178] A. D. Kora, B. A. E. Ongbwa, J.-P. Cances, and V. Meghdadi, “Accurate radio coverage assessment methods investigation for 3G/4G networks,” *Computer Networks*, vol. 107, pp. 246–257, 2016.

[179] J. Riihijarvi and P. Mahonen, “Machine learning for performance prediction in mobile cellular networks,” *IEEE Computational Intelligence Magazine*, vol. 13, no. 1, pp. 51–60, 2018.

[180] D. Mirkovic, G. Armitage, and P. Branch, “A survey of round trip time prediction systems,” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1758–1776, 2018.

[181] A. Yarali, “Public safety networks from tetra to commercial cellular networks,” 2020.