



**BYDGOSZCZ UNIVERSITY
OF SCIENCE AND TECHNOLOGY**

DOCTOR OF PHILOSOPHY THESIS

**DISCIPLINE:
INFORMATION AND COMMUNICATION TECHNOLOGY**

mgr inż. Adam Flizikowski

**Admission control algorithms for future wireless
networks**

*Algorytmy sterowania przyjmowaniem zgłoszeń
w sieciach bezprzewodowych przyszłości*

SCIENTIFIC SUPERVISOR

prof. dr hab. inż. Tadeusz Wysocki
Bydgoszcz University of Science and Technology



BYDGOSZCZ 2023

Serdecznie dziękuję za nieskończoną cierpliwość, cenne uwagi i nieoceniony wkład w powstawanie tej pracy Promotorowi - Panu Profesorowi Tadeuszowi Wysockiemu, oraz Panu Dziekanowi dr Tomaszowi Marciniakowi za niezapomniane wsparcie motywacyjne i organizacyjne.

Pracę tę dedykuję mojej Kochanej Żonie Dorocie, Jasiowi i Kubie, Kochanym Rodzicom. Dla Was i dzięki Wam. Dziękuję.

Table of contents

List of mathematical symbols used	11
Abstract	13
Streszczenie	14
Key terms	15
1 Introduction	17
1.1 Origins of the topic	17
1.2 Objective and problem addressed in this thesis	20
1.3 Scope of the work	21
1.4 Motivation and rationale	21
1.5 Thesis goals, and the scope of work	25
1.6 The research plan, achievements, and original papers	26
1.7 Papers authored and co-authored by the author	27
1.8 Structure of the thesis.....	28
2 State of the art analysis	29
2.1 Introduction to admission and congestion control role	29
2.2 Admission control as part of RRM framework.....	32
2.2.1 Elements of CAC subsystem.....	33
2.2.2 Admission control stages	35
2.3 CAC - Decision rules	38
2.4 Design of a service class (CoS).....	43
2.4.1 Interaction of CAC algorithms with other mechanisms.....	45
2.5 Traffic demand modelling.....	46
2.5.1 Classess of traffic	47
2.5.2 Conclusions for admission control.....	49
2.6 Capacity modelling in wireless systems	49
2.6.1 Modelling adaptive modulation and coding (AMC)	53
2.7 Admission control – algorithms study	54
2.7.1 Non learning based admission control	55
2.7.2 Learning based CAC algorithms	66
2.8 The future wireless networks	69
2.8.1 Beyond RRM with disaggregated vRAN.....	71
2.8.2 Suitability of optimizations for B5G resource management ...	74
2.8.3 Workload prediction in future wireless networks	75
2.8.4 Multi-RAT solutions	77
2.9 The role of KPIs.....	78

2.10	Real-time video delivery	79
2.11	Overview of selected EU research projects	80
2.12	Identification of guidelines for own research tasks.....	83
2.13	Assumptions made to assure high quality reasearch.....	85
3	Own research methodology	87
3.1	Introduction.....	87
3.2	Quality function maximization – preliminary considerations.....	87
3.3	Methodology of own research	91
3.4	Selection of relevant quality characteristics (<i>HU</i>).....	95
3.5	Wireless system modelling and exemplification	97
3.6	Link to system level mapping (L2S / SLS).....	97
3.7	Map generation	98
3.8	Measurement tools validation	99
3.9	Selection of QoE statistics for evaluating quality of video feeds	99
3.10	Overview of the experiment configurations in chapters 4-8	100
3.11	Indication of original work introduced in the next chapters	101
4	Bandwidth based admission control algorithms	103
4.1	Introduction.....	103
4.2	Complete Sharing Connection Admission Control Algorithm	104
4.3	Declaration based admission control algorithms	105
4.3.1	mDHCAC.....	106
4.4	GBR CAC for 5G ORAN	108
4.5	System capacity estimation - ACM and Symbol Reservation	110
4.6	ARAC - measurement based CAC algorithm.....	112
4.6.1	EMAC - algorithm	113
4.6.2	nscARAC algorithm.....	116
4.6.3	ARCAC algorithm	117
4.7	Test scenarios.....	120
4.7.1	Scenario I – Symbols Reservation Schemes (SRS)	121
4.7.2	Scenario II – SRS with two FEC schemes	122
4.7.3	Scenario III – EMAC, nscARAC comparison	124
4.7.4	Scenario IV – ARAC, nscARAC comparison	126
4.7.5	Scenario V – ARAC, EMAC comparison.....	127
4.8	Evaluation of results	128
4.9	Comparison of SRS SCHEMES and different frame lengths.....	128
4.10	MBAC algorithms validation and performance assessment	129
4.10.1	Scenario 1 – EMAC and nscARAC performance	130
4.10.2	Scenario 2 – ARAC and nscARAC performance	132
4.10.3	Scenario 3 – EMAC and ARAC performance	134

4.11	Simulations parallelization.....	136
4.12	Summary	138
5	E2E modelling of wireless links for admission and congestion control.....	140
5.1	Introduction.....	140
5.2	Congestion control for real-time mobile video streaming - system model	141
5.3	Congestion control algorithms	142
5.4	Discussion of the algorithm choices	143
5.4.1	Adaptive polling service (aPS).....	143
5.4.2	Prototype controller design (MCATS).....	145
5.4.3	Channel based traffic policing	145
5.5	Real network measurements – rationale for congestion control	147
5.5.1	Mobile measurements – first round.....	147
5.5.2	Mobile measurements – second round.....	151
5.5.3	Calculating user-plane throughput with overheads.....	154
5.6	Statistical analysis of delays and losses from field tests	157
5.6.1	Generic flow of the estimation.....	157
5.6.2	General flow of simulation.....	158
5.6.3	Validation of 4G/5G delays and losses (simulation).....	162
5.7	Emulation framework design	165
5.7.1	Network disturber details	167
5.7.2	Baseline trace profiles	170
5.8	Validating network disturber	171
5.9	E2E congestion control mechanisms for security scenario.....	173
5.9.1	Congestion control for security scenario.....	173
5.9.2	Remote loop – prototype.....	174
5.9.3	Local loop – prototype	175
5.10	Results of emulating wireless scenarios.....	175
5.11	Summary of emulator framework	175
6	QoE control algorithms for multi-RAT	177
6.1	Introduction.....	177
6.2	Assumptions for experiment	178
6.3	Approach.....	180
6.4	System Model	180
6.5	Multi-RAT activation algorithm through RAN controller (RANC) ...	181
6.6	Functional Architecture of Multi-RAT Activation	183
6.7	Customised LTE Scheduling for multi-RAT testbed.....	185

6.8	Experimental validation	186
6.9	Scenario1: Intelligent RAN switching between LTE and WiFi.....	188
6.9.1	Experiment Aim	188
6.9.2	Experiment Details.....	188
6.10	Scenario2: Selecting modulation and coding based on link quality....	189
6.10.1	Experiment Aim	189
6.10.2	Experimental Detail	189
6.11	Deployment of the Implemented Modules.....	190
6.12	Experiment Execution Process.....	191
6.13	Results.....	192
6.14	Summary	195
7	5G ORAN workload prediction to support CAC algorithms	196
7.1	Introduction.....	196
7.2	SYSTEM MODEL AND PROBLEM DEFINITION.....	197
7.3	PREDICTION AND OPTIMIZATION ALGORITHMS	198
7.3.1	ARIMA	199
7.3.2	LSTM.....	200
7.3.3	N-BEATS.....	200
7.4	EXPERIMENTAL SCENARIOS AND RESULT ANALYSIS	200
7.5	Data Usage for Workload Scheduling in EMDC.....	203
7.6	Modified approach to CPU/energy consumption prediction	207
7.7	Applicability of the results into the 5G vRAN (and beyond)	211
7.8	Use cases for cloudified, virtualized and disaggregated RAN.....	214
7.9	SWOT Analysis of SD-RAN Deployment	215
7.10	Summary	221
8	Learning based CAC agent design.....	222
8.1	Introduction.....	222
8.2	System Model	222
8.3	MDP model.....	224
8.4	The CAC agent definition.....	226
8.5	Call Admission Control – MDP Formulated Problem.....	228
8.6	Reinforcement Learning	229
8.6.1	Artificial Neural Networks.....	231
8.6.2	Q-Learning approximation by ANN	231
8.7	Validation of the model	234
8.8	Applicability into 5G/beyond networks.....	239
8.9	Summary	240
9	Analysis of research results.....	241
9.1	Overview of achievements.....	241
9.1.1	Quality analysis based on the GoS metric (Problem2)	242

9.1.2	Quality of congestion control based on packet loss and delay models (Task2)	245
9.1.3	Service quality (QoE, QoS) analysis with multi-RAT decision agent (Task2)	247
9.1.4	Quality of predicting CPU consumption to leverage the admission control actions (Task3)	248
9.1.5	Quality of the intelligent CAC agent for wireless networks (Task4)	249
9.2	Remarks	249
10	Conclusions.....	250
10.1	SCIENTIFIC conclusions	250
10.2	Practical considerations	254
10.2.1	Recommendations for video controller synthesis	255
10.2.2	Recommendations for video controller architectures.....	256
10.2.3	Implications of CPU prediction for the RRM in future networks	257
10.3	Recommendations for future work	257
11	References.....	259
12	Annex A	280
12.1	Measurement tools validation	280
12.2	Radio traces.....	281
12.3	Delay measurements	282
12.4	Time synchronisation of mobile terminals.....	283
13	Annex B	284
13.1	Temporal activity, Spatial activity metrics	284
13.2	Evaluation of QoE metrics for surveillance real-time video adaptation.....	284
14	Annex C	286
14.1	Network emulation tool	286
15	Annex D	289
15.1.1	Software	289
15.1.2	Performance	289
15.1.3	Configuration	289
15.1.4	Procedure	291
16	Annex E	293
17	Annex F.....	294
18	Annex G.....	296
19	Annex H.....	299

19.1	EVALUATION OF OPTIMAL APPROACHES TO NETWORK DISTURBER	300
20	Annex I.....	301
20.1	Validating emulator with IP camera (Series 1).....	301
20.2	Validating UDP/TCP use in tests (Series 2)	302
20.2.1	Validation of the “delay smoothing” feature of TBONEX (Series 3).....	303
20.2.2	Validating emulation with both rate and delay enabled	304
20.3	Validating the resource consumption of video processing at the video streaming Server	305
20.4	Validating the influence of TCP use (instead of UDP) with full emulation.....	307
20.5	Validating MCATS	308

LIST OF MATHEMATICAL SYMBOLS USED

Table 1 List of main mathematical symbols used in thesis

B_{total}, C_{max}	total bandwidth of the access point
$B_{used}^{<xyz>}$	used bandwidth of the class of service
R_{req}, B_x	requested bandwidth of the access point
B_{th}	resource utilization threshold
$B_{util}, B_{used}, R_{used}$	currently used resources of an access point (e.g. gNB)
$B_{req}^k, B_{<k>}$	bandwidth required by class k , where k can represent: rtPS, UGS, GBR, non-GBR, BE
U	threshold value of bandwidth
D	CAC agent decision
C_H	is the equivalent capacity determined for n-streams
$C(t)$	value of system capacity in the current time unit
L	packets size
L_B	buffer size
$P_{B,k}$	blocking probability of service class k
$P_{D,k}$	droppings probability of service class k
$P_{DV,i}$	IP packets delay variation (jitter) of a connection i
$P_{loss,i}$	IP packets loss rate of a connection i
P_x^R	Priority in regaining bandwidth share
$d_{E2E,k}$	average level of delay experienced by connections of class k
n_k	number of users of class, where $n \in N, k \in K$
P_{used}	the number of symbols used by connection P
S_{used}	resources used by connection P
S_{req}	Number of requested symbols
S_{rsvd}	Number of reserved symbols
S_{min}	Minimum number of symbols required to serve connection
S_{max}	Maximum number of symbols required to serve connection
S_{All}	number of available symbols (slots)
S_{sched}	number of scheduled symbols
S_{pred}	number of predicted symbols
α	symbol reservation factor in range from 0 to 1
β	Bandwidth reservation factor (used to calculate EBW)
R	TDD frame
$R_i(t)$	is the number of packets that can be carried by one time slot
T_{window}	Length of measurement window for the moving average (regards EMAC, ARAC, nscARAC)
$n_{i,p}^{PRB}$	Number of PRBs of UE=i, allocated to AP p
$r_{i,p}$	Bitrate of UEi allocated to AP p
b_{pk}^i	Bitrate request form UEi

SE	spectral efficiency [b/s/Hz]
n_k	number of k-class users
$n_{k,rej}$	number of user connections rejected in class k
$n_{k,acc}$	number of user connections accepted in class k
F()	operator dependent on the set of parameters $F=F(A)$
$L(), P()$	represents the left and right side of the equation, where the L(*) represents the characteristics of the system, while the P(*) is set of controls which are acting upon the identified system in order to assure meeting quality goals
\bar{H}	performance characteristics as output values, e.g. the quality of the transmission system in wireless networks (QbP/GoS),
\bar{E}	elements of transformation of technical conditions of transmission in wireless networks (QbP)
\bar{W}	\bar{W} – represents interference between different users
Θ	dynamics of the analysed system
$t - t_o$	time
\bar{s}	variable representing controlling, regulating, compensating, monitoring, creating, transformative destruction,
\bar{z}	technological, social, environmental, accidental and other disturbances
A	control algorithms with the use of the considered concept of solving the i-th state of the postulated transmission quality in wireless networks (QbP)
O	delays in the share of the i-th state of the postulated transmission quality in wireless networks (QbP)
N	inaccuracy of elements and realizations in the share of the i-th state of the postulated transmission quality in wireless networks (QbP),
SP_i	i-th postulated state of transmission in wireless networks,
Wt_j	j-th technical conditions described by the solution concept for the postulated state
$\underline{q}(q_1, \dots, q_s)$	set of input signals, $\underline{q} \in P$ and $y \in X$, where $f(q, y)$ – probability density function of random signals q (on the inputs) and y (at output) of a transmission system
y	output signal
$y = [y_1 \dots y_n]$	signals at the output of the system
$y^* = [y^*_1 \dots y^*_n]$	signals at the output of the mathematical model
$\varphi(q)$	$\varphi(q) = [\varphi_0(\underline{q}), \varphi_1(\underline{q}) \dots \varphi_k(\underline{q})]$ - arbitrary function of signals \underline{q} , where $\varphi_0(\underline{q}) = 1$
B	regression parameters such that $B = (b_0, b_1, \dots, b_k)$
$ \bullet $	norm
$ \bullet $	absolute value
ε^{-2}	mean square error
X	Random variable

ABSTRACT

Based on the achievements and scientific methods of telecommunications, an attempt was made to analyse and organize the state of knowledge and practice, an original methodology for researching the quality of admission control in future wireless networks and the foundations of congestion control was developed. Exponential increase in data volumes in wireless systems challenges the current networks, designed originally with respect to busy-hours data loads and the capacity currently used is expected to undergo important evolution towards dense networks which can cope with the currently limiting capacities of the current “macro based” systems.

It has been assumed that the scope of system quality foundations (elements, relationships, models, algorithms) for admission control in wireless systems, wireless network architectures, identifies a strategy for scientific, economic and organizational development of global importance (at least from the perspective of 6G, which is currently at the heart of research). A strategy that distinguishes itself by a clear mission based on the development of knowledge and innovation. A vision that focuses on attaining highest quality (also indirectly efficiency, harmlessness) of a new innovative, modernized, optimized wireless systems and a new telecommunications processes. The above characterized by a purposeful process control in pursuit of learning, implementation and management of innovative data delivery solutions for wireless domain. Author perceives utility of this thesis as a systemic support for the procedures of analysis and assessment of the quality and harmlessness of the product and the process of the telco infrastructure that is becoming the mean for creating and managing the key performance value indicators (e.g. sustainability, energy efficiency, trustworthiness) [1].

The importance of the undertaken topic of reducing harmfulness and increasing the quality of the admission control system (efficiency) is strategic for the deployed wireless systems and through it also to the national economy, where the networks of the future are expected to be ubiquitous (smoothly connected) and become a nervous system of the digital economy. Descriptions, concepts of problem solutions, as well as the material used to implement these concepts are fundamental to telecommunications and information technology science. The contribution to the transformation of the new system consists mainly in the comprehensive approach to processing user requests for resources (and its later regulation/congestion control). Developing the basis for an optimal, modernized and innovative selection of data processing (e.g. admission, congestion) control space features according to the proprietary method is combined in this thesis with developing an original procedure (i.e. supporting the evaluation of the effectiveness of operating algorithms) according to the standards of modern

knowledge. It also involves creative organization of this knowledge in the field of network resource management and control.

STRESZCZENIE

Dla osiągnięcia celu pracy, sformułowano 5 problemów badawczych, które rozwiązano drogą eksperymentu cyfrowego, symulacyjnego, fizycznego z uwzględnieniem warunków rzeczywistych oraz laboratoryjnych, a także uwzględniając warunki idealizowane (np. za pomocą stworzonego stanowiska do emulacji warunków sieciowych w sieciach 4G/5G).

Wszystkie problemy były skierowane na weryfikację funkcji jakości zmiennych systemu komunikacji bezprzewodowej, we wskazanych wyżej warunkach.

Problemy rozwiązano, uzyskując liczne wyniki dotyczące charakterystyk jakościowych tj.: grade of service (GoS), quality of service (QoS), quality of experience (QoE) – uwzględniając metryki takie, jak prawdopodobieństwo blokady, prawdopodobieństwo przedwczesnego zakończenia połączenia, stopień wykorzystania pasma.

Sformułowano wnioski poznawcze dotyczące algorytmów sterowania jakością w kierunku optymalizacji i) jakości sterowania w sieciach bezprzewodowych, ii) jakości doświadczanej przez użytkowników takich sieci, iii) warunków użytkowania algorytmów w różnych konfiguracjach sieci (np. 4G, 5G, WiFi), stosowania algorytmów sterowania przyjmowaniem zgłoszeń, algorytmów kontroli przeciążeń oraz algorytmów sterowania ruchem – powyższe w odniesieniu do zmiennych takich jak m.in.: próg ochronny zasobów radiowych, intensywność napływu zgłoszeń, różne scenariusze jakości sygnału, typy przenoszonego ruchu. Osiągnięto sformułowane cele, wskazano również możliwości doskonalenia jakości systemów telekomunikacji w warunkach mobilnych.

Stworzone zostały innowacyjne algorytmy, a także propozycje ram/procedur badawczych, usprawniające proces ciągłego poprawiania jakości, wg własnego pomysłu, algorytmy te zostały przebadane w odpowiednich warunkach. A zaproponowane rozwiązania mają charakter użyteczny i praktyczny – co zostało potwierdzone na etapie egzemplifikacji i praktykalizacji.

KEY TERMS

Telecommunications - a field of technology and science dealing with the transmission of information at a distance using communication means. The legal definition contained in the Polish telecommunications law defines telecommunications as "broadcasting, receiving or transmitting information, regardless of its type, by means of wires, radio or optical waves or other means using electromagnetic energy" [2]. The term "telecommunications" was first used in 1904 by Édouard Estaunié in the book "Traité pratique de télécommunication électrique" ("Practical treatise on electrical telecommunications")[3].

Congestion – a chronic phenomenon of largely increased intensity of data transmission rates, when utilizing means of communication – that is higher than the nominal, projected (calculated) capacity of the infrastructure available.

Quality of the transmission control subsystem - applies to the quality variables of the admission control system (elements, relations, models, algorithms) in device engineering, information transmission installations. This aspect includes the indicators of the representative model and reality (practice), their distance in research and cognitive space, e.g. Euclidean, accuracy, reliability. The *postulated state* that enables the cognition, description and increase of the utilitarian value of complex technical systems, i.e. subsystems of control, information, logistics, used and operated in modern telecommunications (4G, 5G/beyond). In terms of area, the solutions and research interests of this thesis focus on new technologies: purposeful shaping and selection of the design of solutions; shaping and using the properties of algorithms as well as its optimization, modernization and innovation of transmission system engineering variables (selected elements, relations, models, algorithms). Subject-wise, solved problems and tasks are exemplified on a group of telecommunication networks and its measurement and functional systems with computer-driven processing, measurement and control processes.

1 INTRODUCTION

1.1 ORIGINS OF THE TOPIC

It was recognized by the author, that based on the achievements and scientific methods of telecommunications, an attempt can and should be made to analyse, organize the state of knowledge, practice, and develop an original methodology for researching the quality of admission (and congestion) control in future wireless networks, as well as create the basis for the congestion of wireless links in the modern networks. There is already growing and will further intensify the chronic phenomenon of higher data volumes that the network was nominally designed for (provisioned) which leads to a “data tsunami”, and the networks need to be prepared to properly handle it [4].

Congestion in data processing, measured by the quality of admitting and serving user traffic demands, occurs on some sections of the network and in transmission nodes, especially in the highly populated areas. It is manifested by a large reduction in the average traffic speed, long-term congestion, accidental spreading of traffic towards neighbouring networks. It is difficult to overcome due to the spatial limitations of the overloaded infrastructure and the avalanche of traffic in sections affected by congestion. The conditions for the occurrence of congestion are:

- arrival of more senders or recipients (users) at the same time,
- spending certain resources by users in the form of consumption of technical means, other goods, loss of time,
- the spectral efficiency (bits/Hz/s) must be strictly dependent on the underlying technology capabilities.

The congestion can happen in:

- access, metro or WAN networks (on the links, nodes, points of presence)
- in middle-boxes (systems, platforms)

The congestion in networks will appear in situation where:

- in areas with insufficient capacity called "bottlenecks", this is called a primary congestion; in countries with a high level of computerization, there is a special type of primary congestion (it can be called nodal congestion), when as a result of exceeding the capacity of a node constituting a "bottleneck" in a certain area of the network, ultimately traffic is blocked or limited;
- in places that are not "bottlenecks", but as a result of the occurrence of primary congestion, they become congested as a side effect, because the traffic bypassing the critical section is directed to other places, causing the transmission capacity to be exceeded there as well, this type of congestion can be called secondary.

The scientific development of the foundations for the admission control in radio

resource management domain is related to optimizing, modernizing and innovative solving of problems occurring in the field of engineering and technical sciences. The technical, economic and ecological desirability as well as the technological and applicability possibilities of innovations are based on the rationale of a designer - so strongly related to mathematical models, novelty of solutions and common sense (*Figure 1*).

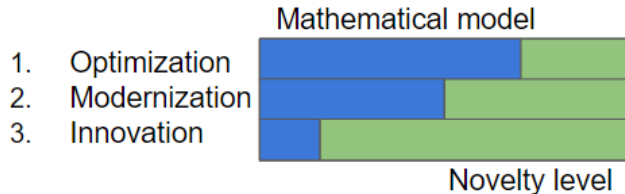


Figure 1 Relationships of mathematical models with novelties in: optimization, modernization and innovation of problem solving (Source: [5])

It is an attempt to answer the question: “*How to reasonably (Figure 1) choose new ideas, features, parameters of the solution (superb, revolutionary, recreational properties), in the broader sense of the novelty of the system (elements, relations, models, algorithms), created in the three creative processes:*

1. *optimization,*
2. *modernization,*
3. *innovation,*

in order to maximize product (solution) quality and process efficiency; minimize environmental damage”. As can be seen from above figure, the procedures for selecting novelties are different for optimization, where they depend mainly on the mathematical model and judging criteria, others in modernization - partly depend on the mathematical model, partly on new solutions, in innovation - they depend mainly on new solutions and fragmentarily on the mathematical model. The topic addressed in this work includes quality, accuracy, reliability in terms of mental cognition, formal description and increasing the value of postulated states of complex technical systems, i.e. data transmission improved by admission/congestion control, its holistic management along with the delivery subsystems (e.g. off-the-shelf servers, operation and management procedures) used and operated in telecommunications. In terms of problems, interests focus on new technologies: deliberate shaping and selection of the architectural means; shaping and using the properties of data services as well as optimization, modernization and innovation of transmission system engineering variables (elements, relations, models, algorithms). Subject-wise, problems and tasks are exemplified on a group of transmission network architectures and measurement-functional systems supported with computer-aided simulation, measurement and control processes. The stages of scientific research, the scope of the basics of the admission control in the engineering of network controls, network designs (and topologies), is a strategy of scientific (*Table 2*), economic and organizational development of global importance as the networks are currently blending with

applications and information systems traffic inside cloud, edge platforms.

Table 2 Characteristics of targeted scientific research strategies

Stages of research	Applied and goal driven strategy: optimization of conditions for improved quality of transmission (minimized congestion)
Title	Influence of technical conditions on operational characteristics of the network
Objective	Analysis, research and development of technical conditions for increasing quality, decreasing congestion in data transmission
Problem statement	What conditions in the technical and data processing domains (\mathbf{W}_{k-p}) are necessary to attain the postulated state
Identifying a hypothesis	The postulated state of quality (admission, congestion) \mathbf{SP} , can be met under conditions of existence of a new technical solution or new data processing $\mathbf{W}_{(k-p)}$
Hypothesis validation and verification	Verification and validation by means of analytics and research, based on relevant sample size, whether under the new conditions $\mathbf{W}_{(k-p)}$, the postulated state can be attained \mathbf{SP}
Drawing a general conclusion	Actuation of new conditions $\mathbf{W}_{(k-p)}$ leads to the targeted postulated state of quality \mathbf{SP} .

A strategy distinguished by a mission based on the development of knowledge and innovation. A vision of the highest quality (also indirectly efficiency, harmlessness) of a new innovative, modernized, optimized product and a new process. Purposeful process control in pursuit of learning, implementation and management of innovative transmission process devices. It is system support with procedures for analysis and assessment of quality and harmlessness.

The importance of the undertaken aspects of reducing harmfulness and increasing the quality of admission control (processing efficiency) is strategic for the national economy. Descriptions, concepts of problem solutions, as well as the material used to implement these concepts are fundamental to science and technology, especially telecommunications. The contribution to the transformation of the new system consists mainly in the discovery of a new means, a way of processing data. Developing the basis for an optimal, modernized and innovative selection of the characteristics of the data processing control space according to the proprietary method and developing an original method supporting the calculation of the effectiveness of the algorithms for operating procedures according to the standards. It also involves creative organization of knowledge in the field of network congestion and nodes.

1.2 OBJECTIVE AND PROBLEM ADDRESSED IN THIS THESIS

The aim of the work is: Analysis, Research and Smart Development (based on knowledge and innovation [6]) and assessment of the quality of control algorithms for accepting and processing requests (congestion) in future wireless networks. To achieve the main and additional goals of the work, the research problems were formulated:

- 1) **Problem1:** Is it possible to introduce substantive models useful in admission (and congestion) control, based on the existing state of knowledge about this aspect in wireless networks, assessed with the quality norm (difference between the model and reality, according to the definition of the standard $Q=||x-y||$) that will allow to: i) assess the status, ii) indicate directions of development, iii) devise alternative solutions, iv) conduct research (with estimation/assessment) and v) propose new control models (trends) towards the transition from 4G to 5G and subsequent generations
- 2) **Problem2:** Can 4G **declaration** (DBAC) and **measurement** (MBAC) based admission control algorithms be reused with necessary modernizations to be suitable for the 5G/and beyond networks, to **further optimize the number of users and decrease probabilities (Pb, Pd)** over existing SOTA?

In order to achieve the goal and solve two above mentioned research problems, it was decided to additionally solve the supplementary research/creative tasks:

1. Task1 (Chapter 5): How to define a complete, balanced, end-to-end evaluation framework for innovative and modernized congestion control algorithms (in 4G/5G) to enable appropriate optimizations adapting transmitted stream of data to the underlying network capacity (e.g. remote monitoring of autonomous cars) for improving capabilities of reliable mobile testing in laboratories, based on field test data.
2. Task2 (Chapter 6): To what extent can quasi-optimization of admission and congestion control parameters and algorithms benefit from combining it with intelligent, QoE-based traffic control in multi-RAT networks using novel RAN controller architectures in the control plane? Multiple networks should take into account the aggregation of multiple access networks, including in particular OFDMA and CSMA-based RATs.
3. Task3 (Chapter 7): To what extent can the introduction of novel Micro Edge Data Center (EMDC) computing platforms combined with AI/ML based load prediction and placement algorithms for CU-UP autoscaling in disaggregated ORAN-based 5G network slices improve the admissible traffic.

4. Task4 (Chapter 8): Currently known Measurement-Based Access Control (MBAC) algorithms, supported by AI/ML techniques with reinforcement learning, MDP, with necessary modernizations, can be used to improve CAC solutions in 4G/5G systems and enable formal frame to continue its evolution towards covering additional characteristics of the future networks?
5. Task5 (Chapter 7): How to design and implement an effective optimization framework that would support the definition of balanced, learning-based RRM CAC/CC algorithms in 5G/B5G networks for use in edge-micro data centres?

In order to solve the above stated research problems, and achieve the main goals, the necessity and possibility of formulating and solving detailed scientific tasks according to the proposed models was considered necessary, in order to assess the impact of selected technology features on the quality of control algorithms for accepting and processing requests (congestion) in future wireless networks. The original research performed by the author in order to respond to the above problems, has been delivered in separate chapters – their numbers are given in parentheses next to each task.

1.3 SCOPE OF THE WORK

Premises for research and foundations of the quality of admission (congestion) control, transformation of the wireless system structure, energy and information environment as transformations of technical conditions into useful and useless states, accumulations, consumption of resources along with emissions to the environment, in telecommunications are identified with the totality of changes, cognition, and description of a selected set, a whole integrated by network admission (and congestion) control.

1.4 MOTIVATION AND RATIONALE

The future networks are in the stage of pre-standardization for 6G. The 5G networks are already standardized (Rel.15-Rel.17) and currently its evolution i.e. 5G-Advanced specifications are being defined by the 3GPP (Rel.18-Rel.20).

Table 3 The evolution of IMT Standards *Source:* [7]

	IMT - 2000	IMT - Advanced	IMT - 2020
Rel. Year	2000	2010	2020
Standards	3G	4G (LTE Advanced)	5G
Air Interfaces	CDMA, TDMA, FDMA	LTE, WiMAX	5G NR
Data Rates	Up to 2 Mbps	Up to 1 Gbps	Up to 20 Gbps
Core Networks	Circuit&Packet Switched core	Evolved Packet Core (EPC)	Software Defined Network (SDN)& Network

	network		Function Virtualization (NFV)
--	---------	--	----------------------------------

The Table 3 shows a simple comparison of the existing ITU standards for the recent wireless network evolution. It is evident that the next generation (6G) stable standards are expected similarly around 2030 to keep the pace of per decade evolution. The key point to notice is a) the fact that with 5G networks are becoming software driven (SDN, NFV) but the data rate is increasing rather due to accessing more spectrum (e.g. dual connectivity, spectrum sharing, multi-carrier, mmWaves with extreme spectrum sizes). The physical layer is largely the same (OFDMA based), but the main changes are touching upon delivery of guaranteed (below 1ms) end-to-end delays so that services like deterministic networking or in general the URLLC (and extreme version of it), massive MIMO with beamforming/precoding and in general the macro access network investments. At the time of this thesis submission the essential research direction concentrates on the existing visions towards the 6G. The latter set of specifications is expected to be delivered in the years 2025-2028. This way the early trials of new generation solutions are expected in the years 2028-2030 and the commercial launches will only be able to start in 2030+. From the perspective of EC-funded research directions the major focus for the scope can be realized based on the topics covered by scientific research and innovation agendas (SRIA) and current and future calls for proposals (Horizon Europe, SNS JU, KDT, etc). The main interest of the future research are the following [8]:

- More **widespread use of AI/ML** algorithms in the edge-cloud continuum (towards deep-edge, with IoT devices) [9]
- Efficient resource allocation for multi-tenant, **multi-RAT** in the more dynamic and changing network topologies and without constraints on particular access-network
- Multiple access techniques such as NOMA and RSMA
- Network **programmability and open APIs** to allow researchers to deliver novel algorithms
- Network architecture is currently driven by the open-RAN initiatives (e.g. ORAN Alliance, TIP Alliance) which introduce the multiple feedback-loop architectures based on intelligent controllers (nearRT-RIC, nonRT-RIC – and soon also real-time RIC will be necessary), where the network interface or control functions are exposed over the north-bound API of the RIC to allow researchers and developers create (and replace) RAN functions by means of dedicated applications – xApps [10]
- There are intense works on the *new communication paradigm called cell-free*, where resource management and allocation are benefiting largely from the fact that current 5G networks are largely virtualized and thus also centralized. Cell-free enables network coverage without the cell borders, as all the access points are belonging to a cluster (or clusters) that shares resources among its APs [11], [12]

In January 2023, the three new research projects (6G-XR, 6G-SANDBOX, 6G-BRICKS)¹ were launched by the EU Commission in order to develop and deliver 6G testbeds which will boost the pace of delivering 6G research solutions to be used by researchers and experimenters all around the world. From the perspective of the status of existing networks, it is predicted that the capacity of legacy cellular systems will soon be threatened without triggering a substantial progress in *densifying the current networks* [13]. In order to be able to deliver the promise of “ultra-dense” networks (UDN) there need to be adaptive (or even disruptive) changes in the way of provisioning and operating wireless systems. The critical foreseen consequences of densification are e.g. the “high intensity of handovers”, “increased interference level” – they need to be addressed to assure successful evolution. The ongoing EU-funded projects (H2020 DAEMON [14], H2020 MARSAL [15]) focus on the introduction of cell-free and distributed cell-free mMIMO - the novel paradigm that can have significant influence on the possibility of introducing truly ultra-dense networks. Together with existing state-of-the-art literature that also highlights the importance of this research direction, it is providing evidence of performance improvements (capacity increase, interference removal) by delivering novel mechanisms and techniques for scheduling, precoding, mMIMO and more (Zhang et al., 2017)[10].

Among all the radio resource management (RRM) mechanisms both admission (congestion) control and scheduling are among the most essential ones that are recommended, but not specified in details by standards of 3GPP - that is why vendors deliver proprietary solutions based on own goals and experiences. Although the schedulers used by major vendors revolve around: round robin, fair schedulers, priority schedulers and so on from ca. 15-20 years already in the cellular systems [18][19].

With the introduction of open-RAN architectures and 5G deployments, it becomes feasible to define 3rd party control applications, so-called xApps, enabling execution within the RAN intelligent controller (RIC) architecture, to manage various 5G/6G RAN functions and mechanisms including: *admission control*, as well as scheduling in the MAC (and other algorithms of radio stack), but also other traffic steering related functionalities [20]. The fact that architectures implementing wireless systems evolve towards open-RAN, virtualization and centralization, creates the need for the combined approach to admission control that manages in parallel: radio resources, computation resources, network resources in combination with elements such as orchestration (e.g. ETSI MANO, Kubernetes) - becomes essential requirement in order to enable sustainable and truly elastic edge computing experiences [21], [22], [23]. Availability of such architectural elements like RIC, E2-interface, xApps contribute to increased potential of innovation in the area of future RAN networks [24].

The standardization of 6G (at time of writing the thesis the 3GPP Release 17 is

¹ <https://6g-Sandbox.eu/>, <https://6g-bricks.eu/> and <https://www.6g-xr.eu/>

being finalized) is still in the planning stage by the 3GPP and is expected to be concluded during the years 2025-2028. Decisions about its features, that can support the massive number of devices (e.g. the massive Machine Type Communication), like e.g. the introduction of non-orthogonal multiple-access schemes like NOMA or RSMA are still to be made, among other waveform adjustments. Meanwhile it will be beneficial to consider hybrid approaches of combining OMA and NOMA in future wireless systems [25], [26], (Liu et al., 2022). The big vendors contributing to 3GPP specification like e.g. Samsung are not widely sharing yet the clear indication of the directions for implementing e.g. the NOMA (RSMA) in the future user terminals. This indication was collected by author in one of the recent IEEE conferences on telecommunication standards in Thessaloniki, Greece (December 2022) by directly asking such question to the head of R&D. The response provided was suggesting that at the moment there is not year clear plan for the next steps related to NOMA.

Numerous whitepapers issued recently by the 5G-PPP Alliance, are targeting ML and AI as the crucial enablers of beyond-5G networks, as well as the novel architectures that are required to introduce future 6G networks build around paradigms of the cell-free, distributed cell-free and mMIMO. Moreover, it is becoming substantial for the future networks to be able to deliver models of various network components, in order to improve the quality of network control and management by digital twinning, prediction and learning (<https://6g-Sandbox.Eu/>, 2023). Besides the recent whitepaper by 5G-PPP identified main pillars for the 6G definition: intelligent connected management and control, programmability, integrated sensing and communications, reduction of energy footprint, trustworthy infrastructure, scalability and affordability [29][30].

Also the future networks consider the aspect of ubiquitous access enabled by the combination of multiple wireless network access technologies in a shape of a multi-RAT (e.g. dual connectivity, LTE/WiFi coexistence (LWA), LWIP, cooperation with NTN networks, cooperation with the WPAN networks like e.g. LoRA WAN). This approach combines resources of many RATs and provides them to users in either access network domain (AN) or backhaul network (BH) domain. Thanks to combining and integrating the networks the user access to data becomes more resilient to disruptions due to coverage, mobility or interference losses in the area (Manjeshwar et al., 2019), [32], [33], [34].

To summarize the above status of the trends in network evolution, the author identifies the need to systematically deal with the topic of admission control and congestion control for future networks. This need results from the following premises:

- **Network densification** leads to novel challenges for controlling network admission due to large increase in small cells, which leads to multiple handovers or zero handovers in the case of cell-free architectures and schedulers
- **User QoE as the essential KPI** should be properly considered in future network control and management. It is especially important for the networks

where multiple-RAT techniques are existing (e.g. OFDMA based and CSMA based)

- As networks become more and more augmented with AI/ML solutions (i.e. data-driven) as well as they become **SW-driven, there increases the role of joint analysis and optimization of multiple resource types**: computing, radio and network level. It is not enough to focus on the selected dimension solely. Especially since new hardware nodes are being designed in Europe (e.g. edge micro data centre) while the virtualized networks become subject to elastic network scaling and migration.
- In addition, it is required to **revisit and validate existing frameworks that address network automation and optimization** in order to consider the critical role of traffic prediction and modelling by means of existing ML algorithms for the virtualized RAN networks.
- In order to accommodate fast growing amount of IoT terminals new waveforms and **multiple-access techniques (e.g. NOMA, RSMA) are recently under investigation**. This trend addresses the needs to boost the ability of networks to deal with intense growth of number of end-terminals, and utilize spectrum resources more efficiently.

1.5 THESIS GOALS, AND THE SCOPE OF WORK

The solution presented in the dissertation concerns the use of machine learning methods (based on MDP) to control the admission of connection-requests in wireless networks 5G/beyond-5G for connections carrying an IP traffic. The research thesis addressed in this dissertation is:

The existing technical conditions in 5G/beyond-5G open-RAN networks, can be improved, e.g. by modifying the algorithms managing admission and congestion control, both: i) based on declarations and ii) supported by measurements - and the expected result will be the improvement of operational characteristics, including the performance (quality) of networks with respect to efficient use of radio, computing and networking resources.

It is possible to define and implement such algorithms supported by measurements (and declarations), based on machine learning and Markov's Decision Processes (MDP) theory, and augmented by QoE feedback, which ensure quality similar to legacy algorithms, but give greater opportunities for the development of efficiency and quality of telecommunications services. In addition, especially wireless systems implemented in accordance with the NOMA and hybrid NOMA/OMA paradigm require specific solutions for admission control.

To prove the above-mentioned thesis, simulation, measurements in the field and analytical tools were used as well as incorporating mathematical apparatus appropriate for the theory of probability and stochastic processes, statistical model definition, regression/non-linear function modelling with artificial neural

networks and machine learning with reinforcement.

1.6 THE RESEARCH PLAN, ACHIEVEMENTS, AND ORIGINAL PAPERS

To make it possible to achieve the above goals (thus proving the above thesis) the author first, planned a set of clear steps towards being able to respond to the research questions identified in section (1.2), covering:

1. Analysis of the **prior art** related to mechanisms of: the physical layer (PHY), link layer (MAC), NOMA multiple access techniques, traffic steering for QoE optimization in multi-RAT environment and in particular CAC algorithms in 4G/5G/beyond-5G networks
2. **Selection, conceptualization, and modernization (with semi-optimization)** of admission control algorithms for the purpose of defining guidelines for own research and methodological assumptions for their more effective use in networks compliant with 5G standards and in next-generation networks
3. Design and **development of a research environment** for pursuing own simulation, emulation and analytical studies (e.g. Matlab, ns2, Linux) and the test methodology supporting the research process
4. **Testing of modified algorithms via analytical and simulation means** and proposing a suitable design and development methodology by modifying (upgrading) structures and parameters (including comparison with the prior art), towards improving the performance indicators of admission control in future networks.

Upon executing and fulfilling the above-mentioned stages and after solving the research problems defined in section 1.2, it will be feasible to assume that the goals of the thesis titled „*Call admission control in the future wireless networks*” have been accomplished. The research work leading to the provision of responses to the research questions mentioned in section 1.2, has allowed author to deliver the following set of original **achievements beyond state of art**:

- **Achievement1:** An analysis of the state of knowledge was carried out, innovative solutions were proposed in the form of the Design and validation of modernized DBAC based admission control algorithms for 4G/5G/B5G
- **Achievement2:** An analysis of the state of knowledge was carried out, innovative solutions were proposed in the form of the Design and validation of modernized, learning based MBAC based admission control algorithms for 4G/5G/B5G
- **Achievement3:** An analysis of the state of knowledge was carried out, innovative solutions were proposed in the form of the Design and validation of the E2E framework enabling evaluation of admission and congestion control algorithms for the 4G/5G/B5G networks, with original HW module for collecting real-life network measurements

- **Achievement4:** An analysis of the state of knowledge was carried out, innovative solutions were proposed in the form of the Design and validation of the QoE based traffic steering algorithm to support congestion and admission control in multi-RAT networks supporting both OFDM-based and CSMA multiple access (e.g. 4G and WiFi)
- **Achievement5:** An analysis of the state of knowledge was carried out, innovative solutions were proposed in the form of the Design and validation of data-driven workload prediction and placement algorithm based on LSTM, N-Beats and ARIMA to support connection admission in EMDC-based deployments of 5G/B5G networks based on open-RAN paradigm
- **Achievement6:** An analysis of the state of knowledge was carried out, innovative solutions were proposed in the form of the 5G vRAN product extension with the proposed admission control algorithm for 5G, combined with intelligent trigger to activate CU-UP autoscaling in the 5G virtual RAN mobile network.

1.7 PAPERS AUTHORED AND CO-AUTHORED BY THE AUTHOR

During the stage of working towards contributing to the research questions supporting the thesis defined in this work, the following original paper by author have been submitted, reviewed and accepted at various conferences and journals:

- Papers demonstrating CAC algorithm modernizations [35] [36][37][38][39][37][40]
- Papers demonstrating authors' achievements related to measurement tools and frameworks [41] [42]
- Papers describing NOMA state-of-the-art [43]
- Papers describing multi-RAT radio controller [44]
- Papers dealing with challenges behind introducing virtualized RAN networks (4G, 5G) into the edge dedicated server designs (e.g. scaling, workload prediction and placement) [45], [46]
- Paper related to sustainability of 6G networks in the view of "loose coupling" theory [47].

Moreover, the topic addressed in this thesis has been successfully applied in the context of research projects e.g.: NCBiR RATfor5G+, ECSEL JU BRAINE, H2020 ORCA, EUREKA-CELTIC MITSU, FP7 DaVinci. In particular the topics related to admission control in various research projects in the past were: i) admission control for 3G in the EuQoS FP7 project, ii) admission control for service oriented overlay network for military VHF/UHF networks in the EDA TACTICS project, iii) admission control for WiMAX networks when the 4G standards were established around 2011, iv) admission control in 5G networks in the H2020 5G Essence/EuWireless, v) admission control combined with the workload prediction in the 6G-SANDBOX for the 6G. This way author has not

only been observing various trends in the network evolution but has particularly been engaged in contributing to such developments, discussions and analyses actively. Summary of the contribution these projects made, into the topic of admission control are presented in the chapter 2.

1.8 STRUCTURE OF THE THESIS

The reminder of the thesis is structured as follows: chapter 2 presents the results of desktop research, chapter 3 introduced methodology underlying the research work, chapter 4 deals with non-learning CAC algorithm for compensation of MCS and mobility related user activity, chapter 5 introduces the framework to deal with lab-based emulation of any 4G/5G/beyond network channel models, chapter 6 deals with the algorithms for multi-RAT based switching of user traffic, chapter 7 introduces the learning framework to support the extension of CAC algorithms with the measurement-basaed prediction of load, chapter 8 introduced learning based design of admission control agents, chapters 9 and 10 perform discussion of results achieved in the chapters 4-8, and eventually the chapter 11 provides list of references.

2 STATE OF THE ART ANALYSIS

2.1 INTRODUCTION TO ADMISSION AND CONGESTION CONTROL ROLE

In each telecommunications network that is using shared resources (buffers, bandwidth, processors), it is crucial to ensure the ability to effectively monitor and control (i.e. regulate) such resource [48]. The lack of mechanisms to ensure effective sharing leads to the degradation of resources (Figure 2). The performance of telecommunication networks [49] (both wired and wireless) can be degraded by overloads caused by e.g. competing for access to scarce resources. The theoretical consequences of such a state affect both the connection and the packet levels. A graphical interpretation of the impact of regulation mechanisms on network throughput is presented in the Figure 2.

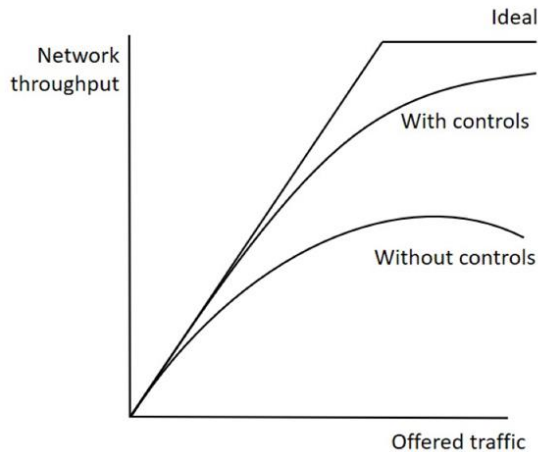


Figure 2 The role of controls in networks [48]

In broadband wireless networks such as 4G, 5G, congestion occurs when the traffic offered exceeds the total capacity of the network, so it is unable to meet network performance requirements and negotiated QoS commitments for current connections and /or new submissions. While in modern wireless systems the wireless network *capacity itself is by definition variable* over time (Figure 12). This way the role of control mechanisms, among others, is to prevent the acceptance of requests that in the current state of resource utilization lead to the occurrence of overload. Two possible strategies for traffic regulation exist [50]: (i) traffic management on the infrastructure side (i.e. traffic multiplexing in routers, base stations, access points, etc.) or (ii) stream adaptation directly on the transmitting side (compression). In the first approach, the regulation of the offered stream or the serviced stream takes place in the intermediary node - described as closed-loop regulation. An alternative is to adapt the traffic at the

source (e.g. adapting the video stream at the transmitter, as described in [51], in order to reduce the demand for bandwidth. In such an approach, additional traffic regulation on the side of the intermediary node (multiplexer) may be inadvisable, because due to the adaptation, the traffic coming out of the source will have low entropy. However, also in this case there is no guarantee of smooth handling of the traffic offered with the assurance that the buffers will not exceed a certain level (L_B) such that $\Pr(X \geq L_B)$, where X is a random process describing the influx of packets to the buffer. The latter approach is referred to as open loop control. The control on the infrastructure side allows for processing a larger volume of offered traffic (maximum load) - i.e. even twice as large as for stream adaptation in the source. On the other hand, the maximum level of offered traffic that a node will be able to handle depends on the acceptable limit value of packet delay in the queues [50, p. 6].

The traffic control mechanisms are necessary to both, guarantee the quality of services provided in the network, and on the other hand, ensure an appropriate level of use of network resources on the side of an operator. The ITU-T organization specification [52] emphasizes that the set of target traffic control mechanisms and congestion control should be implemented using solutions with the lowest possible computational complexity, ensuring the maximum utilization of network resources. Connection Admission Control (CAC) has multiple definitions, e.g.: i) *“the set of actions taken by the network during the call establishment phase (or during call re negotiation phase) in order to establish whether a virtual channel/virtual path connection request can be accepted or rejected (or whether a request for re-allocation can be accommodated)”* [52] ii) *“one of the Radio Resource Management techniques - a set of methods that manage the usage of radio resources and intends to assure QoS and maximize the overall system capacity. The objective of CAC is to maintain a certain level of QoS to the different calls by limiting the number of ongoing calls in the network”* [53]. On the other hand this mechanism is based on a set of actions taken by the network when establishing a connection (or during the negotiation stage of a new connection) in order to decide on acceptance and also includes choosing an end-to-end path. Proper operation of the CAC mechanism guarantees that the acceptance of a new request will not deteriorate the quality of already handled connections [54]. Thus, it allows to control the load in the network and avoid overloads [55]. With regards to the user-plane congestion control in standards of 3GPP there is already the RAN user plane congestion information from the RAN Congestion Awareness Function (RCAF) [56]. In general, congestion in wired networks can be caused by:

- above normative fluctuations of traffic flows,
- network faults.

On the other hand, in next-generation networks (NGN), it is essential to identify and allocate resources for priority connections, especially in the event of network failure or congestion. This assumption is the basis for the traditional

understanding of the role of control mechanisms for accepting requests for priority calls [57]. In turn, from the perspective of classification of the CAC mechanism, it is considered as one of the radio resource management (RRM) techniques. As an essential part of radio resource management (RRM) mechanisms, resource control is needed in various layers of the standard 3GPP protocol stack. The higher layers of the radio stack (PDCP, RRC and MAC) are dealing with the user admission combined with the related resource allocation for more details refer to Figure 4 and section 2.2.2). Mechanisms in the stack important for admission control also reside in the MAC layer of mobile wireless networks. The latter layer among others, deals with resource allocation, modulation and coding scheme selection, power control of uplink transmissions and handover control, respectively. The MAC layer scheduling allocates resources (time, frequency) for both uplink and downlink transmission and operates in each TTI interval, so a timescale of milliseconds range - similar to one sub-frame of standard LTE system. As some novel mechanisms like cell-free, i.e. the cooperative RRM approach, are appearing aligned with the emergence of open RAN interfaces [58], the role of radio access network controller (RAN controller) is growing. Also considering that the 6G will mean broader coexistence of multiple domains (access, core, optical with wireless, cellular with cell-free). Historically it was not always the case, and in the 4G/5G networks it was arbitrarily decided to remove RAN controller and make base station more autonomous – i.e. self-focused. The limited access to controller functionalities, or centralization of processing in today's networks creates situation of competitive resource allocation and thus more interferences. Some recent techniques further empower this trend and rather augment it with various novel techniques like mMIMO, COMP [59]. Also the novel multiple access techniques of NOMA, RSMA are not removing the source of interference but instead they augment the frequency selectivity with additional capabilities to superimpose multiple signals where particular resource (physical resource block) may not get high gain channel at given time [43]. Organizations like O-RAN Alliance [58] are actually understanding the need to have RAN controller, and define the so called near real-time (nearRT) and non real-time (nonRT) RAN intelligent controllers into the open-RAN architecture. The latter is very active trend with vast market share expected in coming years, although many legacy vendors are still not fully supporting this trend due to own motivation. Still technically both ORAN and 3GPP foresee the important role of providing multiple options of functional splits within the RAN stack of 5G and beyond to make the stack more agile and flexible.

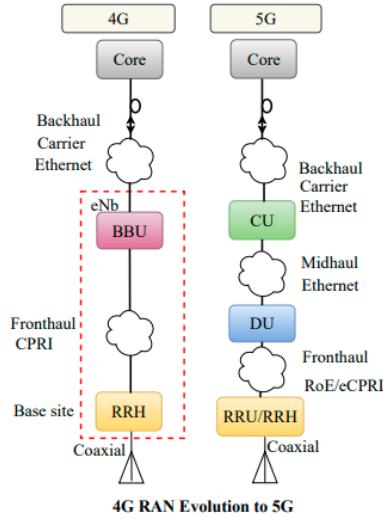


Figure 3 The functional split based architecture evolution from 4G to 5G (Source: [60])

The most recognized splits (Figure 3) are: split 7.2 based on eCPRI (between PHY-high and PHY-low), split 6 (between PHY and MAC layers) and the split 2 (between PDCP and RLC) [60]. Owing to splits in this thesis author will be mainly targeting the split 2, in particular CU-UP function and its offloading, in order to assure additional level of computing resources control especially in the locations at the edge of the network. Irrespective of the splits applied in the networks the RRM mechanisms are operated in different protocols of the stack (Figure 4). The interface between the CU and DUs carries data, configuration, and scheduling related information. The role of scheduler is the fine-grained resource allocation (milliseconds timescale) and the congestion control on the other hand is dealing with some milliseconds-seconds scale, traffic adaptations.

2.2 ADMISSION CONTROL AS PART OF RRM FRAMEWORK

The admission control, congestion control and traffic steering are closely related and they are key consideration of this thesis (Figure 4). The scheduling is natural extension of the CAC algorithms but it will not be the focus of this work. The time scale constituting the time horizon of the CAC algorithms covers the units from seconds, through minutes to hours in the extreme case. Other time scales are expressed in the following ranges: i) cell/packet transmission time (μsec -msec), ii) end-to-end propagation time (msec-sec), iii) connection duration (sec-min) and eventually iv) time between network reconfigurations (days-weeks).

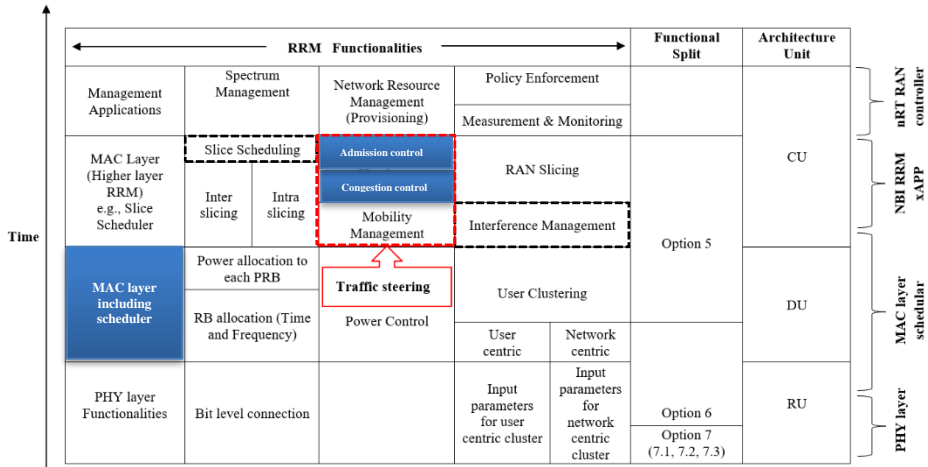


Figure 4 Time scales vs RRM functionalities, functional splits (Source: own)

Regarding the higher layer RRM mechanisms and owing to ORAN specifications e.g., RAN slicing, slice scheduling, admission control, handover management, mobility management can nowadays be developed as special purpose functions, so called xApp – applications. These xApps are located at the north bound interface of the RIC controller of ORAN [61]. Such control applications are executed enabling interactions with the radio stack with granularity of ca. 10-100ms delay (whether regarding DU or CU). The xApp can be run in near-RT RIC by collecting the parameters from the MAC layer and PHY layer through E2 interface.

2.2.1 Elements of CAC subsystem

The general diagram of the functioning of the admission control algorithms, regardless of the type of network in which they operate, is shown in the Figure 5. Depending on the type of radio system, the resources to be controlled will be e.g.: frequencies, time slots, computing resources or spreading codes. Each base station has allocated resources (channels) and their allocation can be static or dynamic.

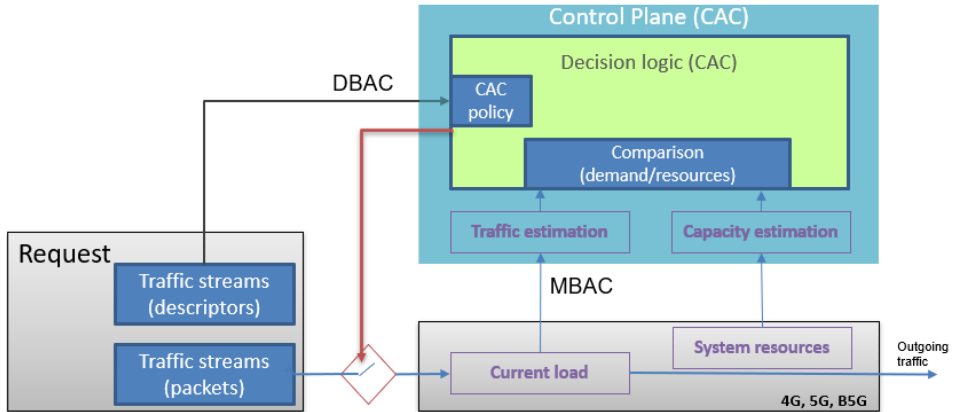


Figure 5 Model of admission control subsystem (Source: author own adaptation based on [62])

Declaration-based admission control algorithms (DBAC) for traffic estimation use the quality parameters of the stream, as indicated in its traffic descriptor (D). Such a solution does not require the presence of a traffic estimator, because the estimation boils down to summing up the nominal QoS values, carried by the traffic descriptors of current connections. Due to the above assumption, the implementation of the DBAC algorithm is simplified, when compared to mechanisms that take into account the real values of instantaneous traffic characteristics. This translates into a greater simplicity of implementation of such solutions, but this is at the expense of efficiency and the efficiency of using the bit rate will be lower compared to systems based on measurement (measurement based admission control - MBAC). The MBAC algorithm performs ongoing measurement of both traffic flows and the real capacity of the system in a given measurement period, a typical measurement includes the following blocks:

- estimation of traffic streams - in terms of system load, characteristics of individual streams, number of streams
- system resource estimation – the task of this functional block is to answer the question of how many resources are left in the system (i.e. at the level of the physical layer), and therefore whether the system is able to handle an additional request. However, the available resources will be measured in different ways depending on the specifics of the system: OVSA codes (for CDMA systems), available bit rate (wired and wireless networks), base station transmitting power, cell interference level, etc.

The final decision to accept or reject a new (or transferred) connection is based on information from the both of above mentioned functional blocks, in conjunction with the information contained in the request's traffic descriptors, and the quality requirements for a given service class (often in the form of target values of traffic parameters). The latter characterize declared parameters that

determine the worst-case behaviour (the so-called deterministic envelope) of e.g. streaming traffic. Examples of such characteristics can be expressed through the parameters of the "Token Bucket" (Double Token Bucket) mechanism. A token bucket can be described by the three parameters {SR, BSS, m }. SR is the maximum (sustainable) bit rate [bps], BSS is the token bucket size for SR [bytes], m is the maximum input bit rate [bytes].

Typical decisions of the CAC control algorithm include (after [63]): (i) acceptance of the requested level of QoS resources, (ii) partial acceptance of the requested QoS resources, (iii) conditional acceptance or finally (iv) rejection of the request. In broadband wireless systems, system load and transmission conditions vary over time. Usually admission considers a Class of service – i.e. an indirect classification of the incoming packet stream to the appropriate traffic handling class in the device. The application control agent (CAC) maintains a database of streams with given characteristics belonging to appropriate classes (the so-called MIB - configuration and management information database) [64]. Therefore, to enable the system to adapt to dynamically changing conditions, the system resource estimator should be based on channel state information, obtained at regular intervals from the physical layer (and MAC) [53, p. 45]. From the architectural perspective, the following are crucial for proper resource management: (1) building and updating a traffic profile in the traffic estimator, (2) current measurement of the bandwidth effectively available in the cell, and (3) a properly designed controller (see the "Decision logic" block in Figure 5). Building traffic profiles has been described, among others, in the works of [65], [66].

In the literature, the term "admission control policy" is understood as a process of regulating the volume of network traffic, while ensuring the quality of connections. Most of the described CAC policies focus on regulating: the level of bandwidth utilization (BW utilization), the total number of connections made and the number of packets/bits sent in a given network per unit of time [67]. When a certain limit of resource utilization is reached, further acceptance of new connections may be blocked, at least until one of the ongoing connections is terminated. Based on such a definition of the CAC policy, the problem of resource allocation in wireless broadband networks can be formulated as the task of identifying the optimal policy, which takes as input the number of connected end users, the number of service classes and the available bandwidth and makes a decision on the efficient allocation of resources.

2.2.2 Admission control stages

In the 4G/5G system admission control takes place at an AP, at a reception of the particular signalling messages listed in the Table 4. The concept of QoS in 5G is based on the notion of flows, whereas in 4G it is based on bearers. The table also indicates what kind of the admission control takes place at the reception of a message. It is important to highlight that the admission control events are

happening (based on 3GPP specifications) at the radio resource control (RRC) signalling connection request, as well as at the handover. There are two distinctive stages:

- **UE Admission Control:** it is executed when the new signalling bearers e.g. SRB1 and SRB2 setup is performed (including PDCP and RLC entity establishment)
- **Bearer Admission Control:** in this case the data connection, so called new data-bearer is setup DRB (including PDCP and RLC entity establishment but also check if the eNB/gNB is able to support the requested bitrate).

The “RRC reestablishment” will not be very interesting here as it is related to failures (e.g. radio-link, handover, reconfiguration). Formally also the handover related requests should be considered when thinking about mobile users switching between cell sites. The bearer level admission will become crucial as the network density is expected to grow in the next years intensively to manage the exponential growth of traffic demand by users [4].

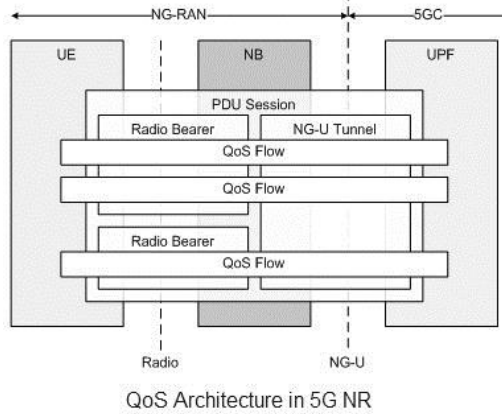


Figure 6 The concept of QoS in 5G (Source : [68])

That is why in chapter 4 author will be addressing the admission control while dealing with more intense changes of cells, when a user is moving across sites which will be smaller due to increased density – e.g. micro or even femto/pico-cells. In such case the intensity of handovers may increase beyond what is experienced in the current networks, without alternatively having implemented complete cooperative RRM with cell-free as an alternative approach relying on 5G disaggregated ORAN networks [69], [70].

Table 4 Admission control – the two stages of execution in 4G/5G (Source: [71])

Protocol and message	UE AC	Bearer AC
RRC CONNECTION REQUEST	Yes	No

RRC CONNECTION RE-ESTABLISHMENT REQUEST	Yes	Yes
S1AP INITIAL CONTEXT SETUP REQUEST	No	Yes
S1AP E-RAB SETUP REQUEST	No	Yes
X2AP HANDOVER REQUEST	Yes	Yes
S1AP HANDOVER REQUEST	Yes	Yes

Typically, an operator defines CAC behaviour based on set of inputs (from UE and RAN) and control parameters – the most popular parameters regarding the commercial networks are summarized in Figure 7. The usual inputs cover the requested quality profile (quality class indicator - QCI) and reason for admission (connection establishment, handover) and these values are compared against the current CPU, memory, cell and networks loads to check the remaining resources. The parameters of admission control usually cover: i) number of UEs per cell, ii) maximum CPU load, iii) available radio resources threshold, iv) maximum load per class of service and/or GBR. The decision rule normally just identifies the requested type of service (GBR and non-GBR) to understand what threshold shall be tracked and compares its current load against the maximum available load. Maximum number of UE per cell depends on the number of configured physical uplink control channels i.e. PUCCHs. The parameters considered for this thesis will be presented in the chapter 3, where own research methodology is provided.

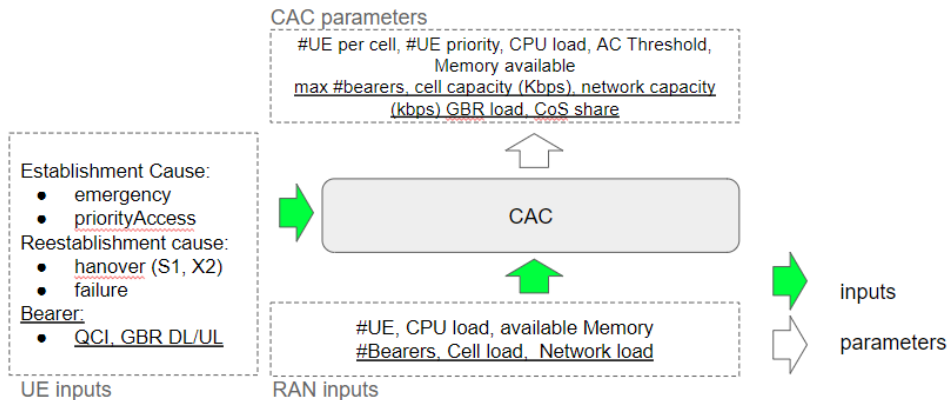


Figure 7 A typical set of inputs/parameters considered for an admission control in 4G/5G (Source: own based on [71])

The baseline admission control rule will control the cell capacity with *number of RBs* in the cell that indicates the maximum available number of physical resource blocks (PRB) and an *average throughput per PRB* as a reflection of spectral efficiency on a set of subcarrier frequencies composing a PRB. The cell-load is

divided between GBR connections maximum load (threshold based) and the overall attainable load in the cell. The crucial fact is that the cell capacity may change as a result of carrier bandwidth change and the change of average throughput per RB.

2.3 CAC - DECISION RULES

The quality of traffic flows (QoS) determines the expectations of the user of a given type of service and depends on the type of traffic. Such requirements are often defined indirectly, i.e. by assigning a stream to a specific class of traffic handling. User expectations are usually evaluated on the basis of two scales: objective (service quality parameters described in the SLA) or subjective (MOS scale for VoIP, PSNR for video, etc.). However, from the perspective of CAC algorithms, subjective metrics and QoS metrics should be considered depending on the type of service, and particular system (4G, 5G, B5G) i.e.:

- for streaming traffic: packet loss level, delays (mean value, variance, quantiles), bit rate
- for flexible traffic: packet loss level, delay (mean value, variance, quantiles), priority

The problem of effectively ensuring the QoS level for connections in packet-switched networks results from e.g. :

- variety of applications (e.g. VBR, CBR)
- reliability of mobility models
- volatility of the number of transferred calls (handover)
- user mobility
- the effectiveness of the worst case description
- statistical multiplexing
- the presence of flexible movement
- lack of a strictly defined connection path
- multicast connections
- the existence of inter-domain connections (causing the need to ensure interoperability and compatibility of the models and mechanisms used to guarantee QoS)
- scalability of network mechanisms.

In order to take into account such diverse requirements, various decision rules are used, adapted to the type of network and the specificity of its operation. The purpose of applying rules is to obtain a satisfactory level of control, i.e. one that allows the maximization of control objectives. Key variables and parameters considered in designing and implementing majority of CAC functions are presented in Table 5 and visually depicted in Figure 8.

Table 5 Main variables considered by CAC algorithms

$B_{used}(t)$	The sum of resources used, guaranteeing the QoS requirements
---------------	--

	for currently implemented connections. In the case of e.g. bit rate, the total bit rate of streams
B_{req}	Resources required to complete the new connection. For connections with variable bit rate (VBF), the peak rate or effective bandwidth is usually used to estimate the bit rate of a new connection. Peak is an inherently conservative estimate, while EBW is best derived from measurements.
C_{th}	The maximum threshold of resource use - the so-called guard band
C_{max}	Total capacity, which is the maximum level of available resources.

The typical rule for admission control is presented in Table 6, besides the other rules consider: transmitted power levels, interference level analysis and analysis of throughput for VBR connections. The main parameters used to describe and define admission control algorithms in this thesis have been depicted in the Figure 8.

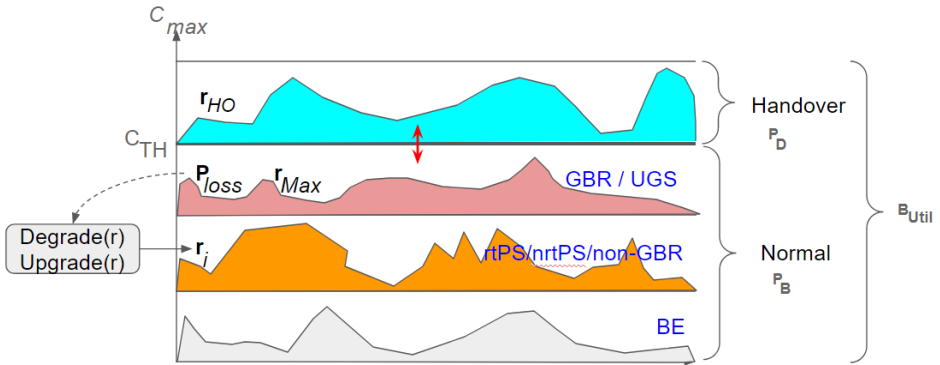


Figure 8 CAC main parameters

The figure shows the total capacity, capacity area dedicated to handover and the non-handover traffic. The utilization of bandwidth is also depicted. The different colours of the time-series refer to multiple classes of service – where the blue series relates to the handover calls. The outlier box indicates the possibility of degrading selected connection resources in order to better support others.

Table 6 Types of decision parameters used in CAC mechanisms [Source: own study]

Rule type	Typical application	Example
Analysis of the level of resources used	Wired networks, wireless (4G, 5G, WiFi)	$D = \begin{cases} 1, & \text{if } B_{used}(t) + B_{req} \leq B_{th} \cdot C_{max} \\ 0, & \text{otherwise} \end{cases}$

Numerous admission control systems that have been developed are based, among others, on the measurement of radio link parameters or on the analysis of the so-called load factor. The reference decision-making algorithms used in the CAC

mechanisms are: complete sharing, measured sum acceptance region [72], Hoeffding bound [73] tangent at peak. The main challenge for CAC control algorithms is that the total capacity (C_{max}) in radio systems varies over time and depends on many factors, including (after [53],[63],[74] the following:

- the scheduling algorithm used in the MAC layer,
- type and configuration of applied correction and protection mechanisms at the level of the physical layer and MAC,
- channel status (and environmental conditions),
- location of the mobile terminal,
- degree of terminal mobility (connection switching - so-called handover)
- maximum level of terminal transmitting power.

Therefore, as various authors point out, a detailed analysis of such systems (especially those using OFDM multiple-access) in terms of obtaining the distribution of buffer occupancy or loads is complicated. However, the system capacity itself, due to the aforementioned variability, should be estimated on an online basis (i.e. $C(t)$). Depending on the type of access network (wired, wireless), the decision rules will take into account different types of resources, including: bit rate, (radiated) power level, interference level but also recently computing etc. Example decision rules for different types of controlled resources are presented in Figure 9.

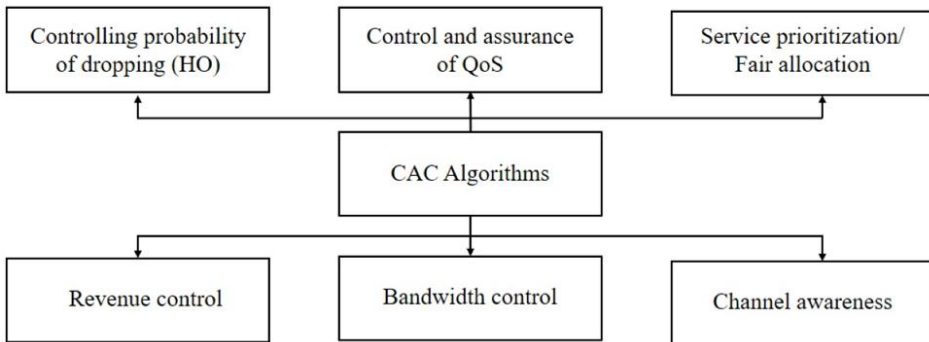


Figure 9 Typical groups of CAC algorithms in wireless networks (Source: [53])

The authors in [75] divide CAC algorithms into two groups: *deterministic and stochastic*. The former are intended to define a situation in which it is possible to guarantee the quality of calls. However, in the second case, the quality parameters of connections (QoS) are guaranteed only at probabilistic level - the taxonomy of such methods is presented in the Figure 10.

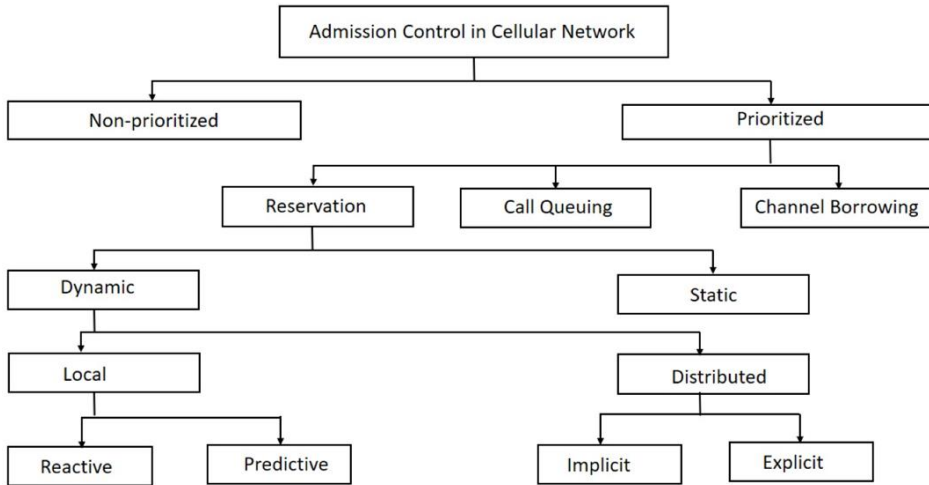


Figure 10 Taxonomy of admission control methods [75]

To deal with the above mentioned rules crucial is the information about the system resources which can be acquired based on declarations (declaration based AC - DBAC) and measurements (measurement based AC - MBAC) listed in the table (Table 7). The literature also includes an additional classification, i.e. CAC algorithms based on experience [76]. The latter is assumed to be a hybrid solution combining PBAC and MBAC approaches. Like MBAC, they rely on measurements to ensure higher network utilization, while these measurements use historical data. In turn, similarly to PBAC, they take into account the nominal values of traffic descriptors. The main disadvantage of mechanisms based on declarations carried in traffic descriptors (traffic contract) is that the traffic description does not include statistical indicators, such as correlation and dynamics of intensity changes, especially when it comes to the phenomenon of bursts [77]. Adoption of such a simplification will lead to degradation of network performance due to insufficient description of the traffic. One of the challenges of CAC algorithms is adapting to dynamic (and often also rapid) changes in the flow of traffic offered [74]. Other approaches to implementing the CAC mechanism include, but are not limited to so-called control preceded by probing the link status [78][63]. Taking into account the history of previous requests in the algorithm optimization process is the basis of learning algorithms that, based on the available data, use statistics, probability and optimization tools to create an appropriate control (decision) algorithm [79]. The approaches to control implementation described above apply to the centralized variant, i.e. one in which the control algorithms are located in a router, base station or access point. However, in the literature there are also decentralized solutions, the so-called endpoint admission control (EAC) [80][81]. Approaches of this type do not require the implementation of virtually any decision-making logic in network nodes, while their main weakness lies in the assessment of network parameters

only from the perspective of end nodes - such an approach is exposed to reduced network efficiency resulting from incorrectly set traffic priorities for CAC queries, and the ease of modifying the decision-making logic in a manner inconsistent with its intended purpose (e.g. cyber attack).

Table 7 Comparison of key groups of CAC algorithms (Source: own)

	Declaration based CAC algorithms	Measurement based CAC algorithms	Intelligent algorithms
Decision rules	Static – predefined Low flexibility of changes.	CAC parameter settings depends on traffic profiles. Relation defined at design stage. Low flexibility of changes.	Decision algorithm can be tuned – based on reward function. High flexibility.
Traffic variability handling	Configured according to predefined network provisioning and capacity planning procedures.	Configured according to predefined network provisioning and capacity planning procedures. However algorithm parameter values can be adjusted based on mathematic equations and traffic profiles.	Traffic handling depends only on algorithm configuration. Can be learned.
Flexibility in considering operator rules / traffic evolution	Any changes require re-starting optimization process (network planning) Policy based networking needs to be implemented to support it.	Preferences can be aligned with equations identifying relation between algorithm parameters and traffic parameters. Policy based networking needs to be implemented to support it.	Decision process and operator preferences can be adjusted fully dynamically based on traffic profiles, environment settings.
Applicability for multi-tenancy	Needs to be defined at network planning and provisioning.	Needs to be defined at network planning and provisioning.	Multi-tenant rules can be learned.
Robustness to cyber attacks	There is no special treatment by the CAC	There is no special treatment by the	Algorithm can be defined in a

(e.g. denial of service)	algorithm possible. Only based on the additional equipment like firewalls.	CAC algorithm possible. Only based on the additional equipment like firewalls.	way where high intensity of flows can be treated as „batch arrivals” or attack vector. Anomaly detection can be aligned with the decision making of an algorithm.
Decision making criteria flexibility	Low – algorithm requires redesign.	Low – algorithm requires redesign.	High – depends on the reward function.

The key point from this section is that it is crucial to address the description of traffic classes by means of proper set of quality parameters.

2.4 DESIGN OF A SERVICE CLASS (COS)

In the 4G, 5G and beyond, the QoS parameters of service classes are identified in specifications (e.g. [82], [83]). Resource Type (GBR, non-GBR, Delay critical GBR), Priority Level (PL), Packet Delay Budget (PDB), Packet Error Loss Rate (PELR), Maximum Burst Size (MBS), Data Rate Averaging Window (DRAW). Where the MBS is defined for the delay critical services as the maximum amount of data that 5G access network is required to serve with certain level of delay (PDB) and DRAW is used to calculate the GBR and maximum bit-rate (MBR). The difference between values of GBR and MBR are subject to relative QoS priority. These values are directly connected with the QoS parameters identified in the IETF and ITU specifications for the IP QoS networks[84]. As can be seen in the figure below the QoS flow accompanied by the appropriate identifier (QFI) is the subject of receiving certain traffic forwarding treatment (scheduling, admission threshold).

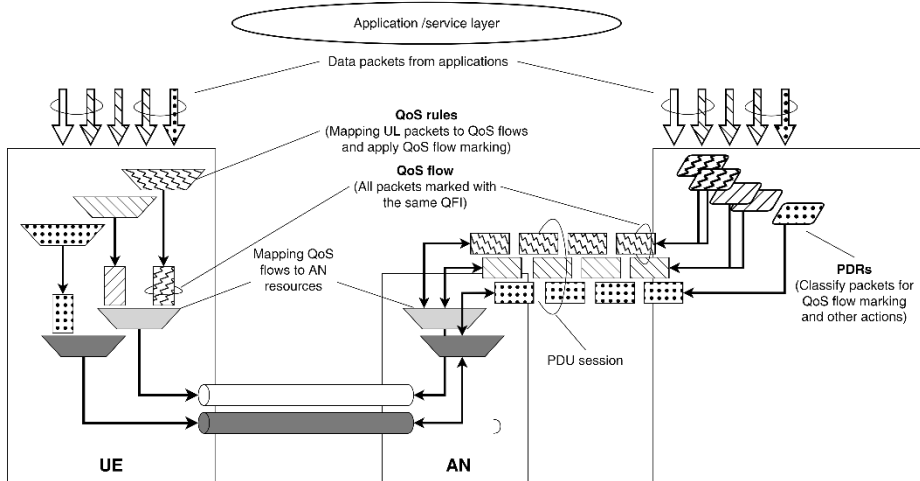


Figure 11 Classification and user-plane marking for QoS flows [85]

The components of the stack important for delivery of the QoS, like the radio bearers of the PDCP (signalling and data) are offering transmission service to the higher layers. From the point of above mentioned principle of user-plane traffic QoS controls, the dynamic admission control algorithms can e.g. be differentiated based on the optimization goal: the probability of blocking new call (P_B) or the probability of rejecting new call (handoff) - P_D . Typically, CAC optimization goals include two cases:

- Case A – minimization of the P_B with a given limit for the maximum PD value, i.e. $PD \leq PD^{\text{tar}}$,
- Case B – minimization of the PD with a given limit for the maximum P_B value, i.e. $P_B \leq PB^{\text{tar}}$.

It has to be noted that in both cases it is the ultimate measure of QoS parameters fulfilment that delineates between accepted and rejected connections. Both probabilities are there to allow finer control of the volume of processed calls perceived by a statistical means. In [86] authors discuss in details the modifications required in calculating the blocking probability especially to deal with various quality metrics (P_b, P_d, \dots) they mention after the ITU-T that “*QoS for mobile services [...] includes different parameters [...] like availability, accessibility, maintainability and user perception of service.*”. Although these metrics are defined in the context of cognitive radio they can be considered for enhancing the QoS definition:

- availability – is the amount of radio channels (PRBs) allocated to users over time
- accessibility – relates available channels with time, and indicates that allocation should be enabling highest modulation support on a given channels, based on RF condition

- maintainability – here is understood as set of technical conditions that need to be properly handled, in order to deal with mobile connections (e.g. handover, speed of communications)
- user expectation – summarizes expectations of quality of service consumed by the user.

Overall the measure of accessibility called grade of service (GoS) is used “as the probability that a call is blocked, or the probability of a call experiencing a delay greater than the predefined queuing time” in the busy hour (Laishram Romesh and Mangang, 2012). It can also be said that GoS measures channels congestion. Recently the European Commission (EC) has been more intensively indicating interest in so called key value indicators (KVI) as opposed to KPIs (key performance indicators) [1]. It is evident that the GoS metric is not there, as the focus is on bringing evolution to performance targets – instead there are targets like: capacity of an area, reliability, trustworthiness targets, as well as purely subjective metric of user quality - QoE (so user subjective metric).

2.4.1 Interaction of CAC algorithms with other mechanisms

The authors in [88] indicate that especially for the purposes of self-organization in networks (here LTE), the design of CAC mechanisms should take into account additional control mechanisms, such as load *balancing*. It should be noted that the **lack of a consistent approach to their configuration may lead to mutual exclusion of optimization goals**, as shown in the article but also in the work [54, p. 153]. The 4G/5G networks have been designed to support mechanisms of this type, an example of which can be the specific role of the X2/Xn interfaces [89], supporting the exchange of information between neighbouring cells (base stations). This approach guarantees the exchange of data that is necessary, among others, for the coordination of mechanisms and procedures from different control levels (CAC, load balancing, interference cancellation and avoidance etc.). In addition, the need for coordination increases as the density of infrastructure elements of future networks increases, as exemplified by networks based on femtocells. The subject of radio resource management in multi-tier networks using m.in femto-cells is discussed in the work [90]. It is worth noting after the 3GPP specification [91] that the 3GPP organization does not plan to standardize self-organization solutions (SON algorithms), but only measurements, procedures and open interfaces enabling the implementation of the SON idea in practice (e.g. X2 and S1 interfaces) are the subject of standardization. Still at time of this thesis preparation the ORAN Alliance and ETSI both deal with specifying aspects of SON in the future networks. The authors in [92] present an analysis of various self-organization schemes in 4G networks (LTE femtocells), in order to ensure the efficiency of continued connections. Authors in [93] indicate that densification of networks in 5G and beyond, increased number of network parameters to (e.g. 2000 parameters for 5G node), increased attention to QoE, coexistence of 5G with its currently operating ancestors among others are crucial

factors regarding self-organization. After [88] self-organization using the CAC approach, i.e. by prioritizing the transferred connections, leads to an increase in P_B , and reduces the P_D . The interaction between load balancing (LB) and CAC leads to a decrease in the failed handover rate (HOFR) due to the reservation of resources by the CAC. In contrast, the presence of load balancing translates into an increased likelihood of transferring the connection from overloaded cells to those with less occupancy. The result of the above is an increase in the probability of rejection (P_D).

An effective policy controlling the acceptance of requests coordinated with spectrum sharing between primary and opportunistic users is considered in multi-tier networks [90]. Such mechanisms play an important role in regulating and ensuring a trade-off between ensuring high bandwidth utilization and individual QoS guarantees at the connection level. The combined analysis of CAC mechanisms and spectrum sharing access mechanisms for multi-plane networks based on multi-access OFDMA allows to regulate loads between macro- and microcells to improve the overall network performance. However, it is important to take into account, in cross-layer mode, QoS measurements both at the connection level (P_B , P_D) and at the MAC layer packet level (throughput, loss level).

2.5 TRAFFIC DEMAND MODELLING

To model multiple quality levels, when defining a class of service performance a maximum allowed value of the link utilization can be calculated applying queueing theory for particular system by providing the buffer size and loss rate. Traffic demand modelling in wireless networks is often based on the simplest streams (stream with an exponential time distribution between requests) due to the many advantages it provides (no consequences, uniformity and singularity). In the literature, many authors point out that in order to model Internet traffic, alternative distributions should be used, m.in. Pareto distribution, or in general a class of exponential distributions, which are better suited for this type of traffic [94]. The consequence of using self-similar traffic (instead of the Poisson distribution) is a drastic increase in queue occupancy and, as a result, also delays, the instantaneous occupancy of the queue can no longer be described by an exponential distribution, but by the so-called "long tail of the distribution". The specificity of the exponential distribution means that depending on the exponent (typically negative), a given function will (a) have a variance and an expected value, (b) only a variance, or (c) both quantities will be infinite. Typical examples of power laws are Zipf's law and Pareto's law [95]. The presence of self-similar traffic means that, unlike traffic modelled using the Poisson distribution, the aforementioned "side effects" of sending this type of traffic in networks cannot be removed by using control mechanisms (because they will occur equally on different time scales). In addition, the estimation of delays for self-similar traffic streams is complex and may not lead to the establishment of a formula describing

this relationship [96]. Typical systems used in literature are e.g. the M/D/1/B, M/M/1, G/M/1. Based on a particular model, then link utilization value for a class of service can be calculated - ρ_k . It will depend on the size of buffer B_k (in packets) and maximum allowed loss probability - P_{loss} . By assuming the allowed level of P_B , Erlang model can be used to calculate corresponding values of the arrival rate for a CoS [97]. In addition to queuing analysis, for modelling radio systems and analysing their performance, approaches using Petri nets (where the model is non-Markov) and integrated approaches using queuing models and Petri nets simultaneously are used, e.g. "Petri Nets including Queueing Networks", interesting applications of this type of approach can be found e.g. in the book [98]. For many years, models based on the intensity of arrivals described by the Poisson distribution have been used in the design and dimensioning of telecommunications networks [99]. However according to various researchers, the queuing systems applicability is limited due to the requirement of steady state of the system (fixed probabilities of transitions in the matrix) [100]. Therefore, the results obtained using this method, for the purposes of identifying optimal solutions (compromise between QoS and power consumption), do not take into account the system dynamics, instead they constitute a reliable estimate of system performance only in the long term. Sample calculations using the above-mentioned techniques are presented in [101] [63].

2.5.1 Classess of traffic

From the perspective of this thesis, quality parameters of video streams and their minimum required values are described in recommendation Y.1541 [102]. The classes of service further considered in this thesis are:

- Streaming class (representing video delivery as well as video conferencing)
- Elastic class (utility of connection is gradually changing as data rate changes)
- Real-time class – with constant bit-rate traffic (CBR) usually modelling the VoIP traffic.

For a streaming class to be modelled an effective bandwidth was coined by Kelly and Gibbens [72] [103] – which is an effective bandwidth of the source lies between its average and peak bit rates; this value for N aggregated sources is lower than the N th multiple of the EBW of a single source (this phenomenon is known as statistical multiplexing gain). The effective bandwidth is a measure that simultaneously determines the characteristics of the packet stream and the QoS quality requirements of network resources necessary to meet the demand defined in this way. The method of determining the effective bandwidth is also described in [104], [105], [50], [73]. The EBW is calculated using the declared traffic descriptor. The second approach is based on the assessment of the amount of EBW based on on-line traffic stream measurements. An interesting extension of the effective bandwidth method is the adaptation of the "many sources

asymptotic" (MSA) approach originally proposed by Courcoubetis (based on Large deviations theory) as an extension of the approach [103] - to networks with variable channel capacity. **More suitable for the prediction of delays and losses are models based on neural networks (ANN)**, which allow one to create effective predictors that can also be used to simply determine the EBW value.

2.5.1.1 Video traffic modeling

One of the well-known approaches to video traffic modelling is the use of the Markov process, e.g. $M/G/\infty$ [106]. There are many examples of using this model to model video traffic, although the authors emphasize that although it is easy to implement, it does not provide a high-quality model. On the other hand, some researchers believe that obtaining an accurate model for MPEG motion is very complicated, if not impossible [107]. In contrast, popular methods for predicting time series based on linear autoregressive analysis (ARIMA) have been used to model video traffic in cellular networks in addition to FARIMA [108] - although it is characterized by long model synthesis times, or simplified variant of the ARIMA model, the so-called "simplified seasonal ARIMA" (SAM) is suitable for modelling the video traffic compliant with MPEG4 Part2/H.264 standards. The model for MPEG2 and H.264 standards differs only in the value of the seasonality parameter, although H.264 traffic, due to its higher compression level, is characterised by greater variability in frame size. Other linear approaches to traffic prediction include the Holt-Winters model, the ARAR method or ARMA. An alternative method is the use of non-linear neural networks, which can be used for time series modelling and prediction, even when the relationships in the underlying data are complex. There are works that use a combination of both approaches, creating neural networks with predictive capabilities, in application to current bandwidth allocation for video calls [109] [110]. On the other hand, the application of a deep neural network approach for time series prediction is shown in [111] The conclusions of the study indicate a significant advantage of using non-linear methods in this type of problems [112].

2.5.1.2 Uplink video delivery in wireless networks

In the article [113] the authors search for optimal values of 4G system parameters for IPTV broadcasting. Video streams of this type offer a tolerance of low packet losses, while they are quite sensitive to latency and its variability [114]. Latency variability can lead to rapid degradation of user-perceived quality - i.e. QoE. When focusing on the uplink direction it can be mitigated by tuning the system parameters (OFDM frame, spatial diversity, ARQ/HARQ tuning). The article [115] addresses the problem of optimising video transmission in LTE networks with QoS guarantees. The authors describe an algorithm for shaping video traffic through an applied additional buffer located in the LTE base station (NodeB). The aim of the optimisation is to minimise the transmission time with the peak bit rate. The proposed algorithm makes it possible to reduce the level of variation in the bit rate of the video stream by about 20% on average, and for video streams

that switch between less dynamic scenes (e.g. news) even up to more than 40% (for the H.264 codec). Similarly, in [116], the authors focus on the possibility of adapting the video traffic stream sent by the terminal in the 'upstream' direction, adjusting its parameters to the current radio conditions. To this end, the architecture and functional assumptions of a *cross-layer* controller (i.e. conforming to the *cross-layer* paradigm) are described. The solution was prepared for the mobile terminal (i.e. for the 'upstream' direction), on which the video server is installed, and not, for example, as proposed in the work on the base station side. Although authors show that the proposed CLO mechanism responds correctly to changes in base station resources, leading to adaptation of the video stream bit rate but they do not provide an exhaustive analysis of the solution, in particular there are no tests showing the response of the CLO mechanism to the channel variability associated with terminal mobility.

Regarding the elastic traffic, since the vast majority of flexible traffic streams use TCP, it is mainly TCP that is considered, in various research works to assess the feasibility of meeting quality objectives in e.g. heterogeneous networks [63]. Comprehensive analysis is presented by [94], which shows that even small UDP background traffic sent together with aggregated TCP traffic (e.g. background traffic) can significantly affect the stability of TCP connections.

2.5.2 Conclusions for admission control

The authors in [117] note that if calls arrive according to the Poisson distribution, CAC algorithms based on traffic measurements and based on measurements of the intensity of requests will allow to accept the same number of calls. However, if the flow of traffic requests, e.g. real-time streaming, does not match the Poisson distribution, then in response to the appearance of a group of requests (bursts), these algorithms secured for such an eventuality will cope better (measurement and arrival aware - MAAC). CAC tests are carried out for self-similar traffic and for arrivals described by the Poisson distribution. Acceptance of the connection for a new session will be performed when, after its admission, the probability of exceeding the limit value of the probability of filling the transmission buffer does not exceed the set value. In contrast, analytical modeling of systems using more complex queuing mechanisms (e.g. than FIFO), more complex CAC algorithms or data source models, quickly becomes complicated [66][118].

2.6 CAPACITY MODELLING IN WIRELESS SYSTEMS

The capacity of systems such as m.in 4G, 5G can be described as the so-called soft capacity [119], mainly due to changes in spectral efficiency over time, primarily in connection with: adaptive modulation and coding (AMC), FEC correction coding and the use of MIMO multiplexing. Due to frequent, dynamic channel changes, the capacity of the channel in the 4G/5G networks, as well as its use, is subject to significant changes. The cell (sector) capacity in cellular

networks is defined as the maximum amount of traffic that can be carried in a cell for the required radio coverage and expected connection quality levels [54]. As shown in the aforementioned study, setting too aggressive level of quality targets for the cell can lead to quality problems, similarly too conservative an approach can lead to an increase in the level of *waste of capacity*. In [120] time-frequency division SISO systems are compared by the author with adaptive capacitance (AMC) systems by using Shannon formula for an AWGN channel system:

$$C \leq W \log_2 \left(1 + \frac{S}{N} \right) \quad 2-1$$

As shown, the highest maximum value of spectral efficiency is obtained by CDMA and TDMA systems with adaptive coding and modulation (6 bit/s/Hz according to Shannon's theorem, 2-3 bit/s/Hz using correction codes). The capacity of radio channel with fast fading, can be expressed as

$$C = E[\log(1 + |h|^2 SNR)] \text{ bits/s/Hz} \quad 2-2$$

A comprehensive study on the analysis of the capacity of various types of channels (AWGN, with slow/fast-changing Rayleigh fading) can be found in the [121]. Furthermore, aspects of channel modelling are well described in [122], [123] among others. However it is worth noting after [75] that the spectral efficiency of different channel access methods (CDMA, FDMA, TDMA) for the Gaussian channel (AWGN), will be equivalent. In the reminder of the thesis the capacity will be calculated following the two equations below (2-3,(2-4):

$$T_s = (1 + G) \cdot F_s / N_{FFT}$$

$$F_s = \text{floor}(n \cdot BW / 8000) \times 8000 \quad (2-3)$$

$$r_{i,p} = 2^{-\mu} 10^{-3} B_{PRB} \log_2 (1 + \text{SINR}_{i,p}) \quad (2-4)$$

Where B_{PRB} is the bandwidth of a single PRB, and it can be computed as $B_{PRB} = 12 * 2^\mu 15 \text{ kHz}$. The μ represents the multiple numerologies in 5G, but in 4G this parameter should be set to $\mu = 0$. The approach shown in (2-4) is followed e.g. by authors in [124] when modelling resources in multi-RAT networks of 5G. In OFDM/ OFDMA systems utilizing the adaptive modulation and coding schemes (see next section) each user can use coding and modulation scheme most appropriate to its channel conditions. Therefore even a constant amount of traffic generated by an application can require different number of OFDM symbols/slots. Therefore achieved transfer rates of a wireless link can vary significantly over short period of time. This adds a "second dimension" to the problem of estimating resources required by an application, since it is hard to predict how users' channel conditions will vary over time. This is in contrast to classic approach to admission control, where capacity of a link in terms of a maximum throughput / number of calls is considered constant. As a consequence, in such an AMC-enabled system, OFDM symbols (or slots for OFDMA) should

be considered a scarce resource, since number of symbols available for a given system remains constant. The authors in [123] give the expression for the total number of symbols in the TDMA frame and for the minimum value, depending on the required minimum bandwidth.

Authors in [96] remind that a frequent practice resulting from the self-similarity of traffic in the Internet, is to design the network assuming that the average expected traffic should not exceed 50-60% of the bandwidth of the system. This results directly from the estimation of the level of delays for the infinite length of the queue, which is affected by the traffic described by the Brownian distribution (i.e. Brownian Motion). While for traffic without signs of self-similarity, the maximum aggregated throughput value is 95%. On the other hand, according to the ErlangB model, an increase in link speed leads to a decrease in the size of the link, while increasing the traffic offered causes an increase in this indicator.

Techniques used to improve capacity can span from header compression (at MAC), silence suppression (it doubles number of VoIP users), improved resource scheduling. The optimal number of MIMO branches 2 (i.e MIMO 2x2), where SNR improvement can be up to 1,8dB. The authors in [120] presented a methodology used in *system-level* simulation to obtain statistically reliable link throughput values as a function of SINR for MIMO links, without the need for full simulation of individual links. This method has been used in system-level simulations in, among others, the BuNGee project [125] and will be followed in this thesis for chapters 4 and 8. It uses the so-called theoretical channel capacity limit ('truncated Shannon bound') to estimate the capacity of a radio system with an adaptive modulation and coding mechanism in next-generation networks. The authors conclude that the **SINR value is a sufficient parameter needed to determine the throughput distribution**. With an approach to capacity modelling presented in for the radio modulation family [73] it is also possible with proposed equations to identify the maximum distance from an AP, where particular modulation scheme can be attained at a given SNR value. For the TDD systems in case resources are fully utilized for a given transmission direction, it is possible to dynamically modify the DL/UL ratio [126]. Here, the admission control first prefers requests from areas with higher² spectral efficiency (bit/s), i.e. higher modulation, although the *scheduling* mechanism itself enables resources sharing in a fair fashion.

In [127], the author focuses on the issue of an analytical approach to estimating the capacity of wireless OFDMA systems with the use of persistent allocation leads to a significant decrease in signalling overhead, especially for a large number of users simultaneously active in a cell (up to 68% decrease). As a result, the average bit rate available in the cell increases accordingly. However, 'static'

² Performance is understood here as the average number of usable bits per symbol within a given subcarrier.

(with memory) allocation is prone to two types of problems: unused areas in the TDMA frame (*resource hole*) and MCS value *mismatch* (MCS *mismatch*).

In the work [128] an assumption was made about the assignment of terminals to concentric circles which are described by a certain bit rate value. In the [75] an exhaustive classification of methods of logical division of the radio cell area into concentric zones depending on the required SNR level is given. Methods are divided into static, where the number and size of zones is fixed, and dynamic, where the number and size of zones is dynamically adapted. In particular, each of such separate regions may have its own specific method of allocating resources (channels). The authors in [Unipd2010] propose an approach for determining aggregate capacity in 4G networks using a Diophantine transformation. The expressions proposed for the determination of capacity in both directions consider the capacity C_{OH} allocated to the signalling data, C_U denotes the part of the capacity already allocated, while $C_{eff} = \beta(C - C_{OH} - C_U)$ is effectively available capacity. The β factor determines the so-called "loss of capacity" due to low efficiency of packaging algorithms, its value will depend on the approach used (and the permutation method whether partial or full utilization of PRB resources – e.g. resource block groups / RBG). It should be noted that the size of the signalling overhead to be taken into account in planning network capacity is about 20% for 4G and about 14% for 5G [129]. The recently performed performance tests in the 5G network deployed in Oman, shows that an average downlink throughput per UE in Oman supported by 4G network to deliver MBB services was 8-12Mbps across 5 tested cities. Whereas the average throughput for 5G in one of the selected cities where 5G has reasonable coverage is 8-18Mbps when compared for two operators [130].

The use of MIMO antenna systems, massive MIMO along with precoding and beamforming techniques, as well as recently very popular reflective intelligent antenna surfaces (RIS) solutions (FR2 frequency band) - will lead to an improvement in the link budget to the level of average 7-9 bit/s/Hz. Online simulator for single user throughput for 5G NR SU-MIMO can be found in here [131]. However, in the long term, even this value of spectral efficiency will be insufficient, to counteract the expected increase in traffic till 2030-35 – i.e. 20-fold increase in capacity needed. In practice, this may mean the need to increase network density by deploying additional small cells up to ca. 177 units per km²[13] in hotspots of the densely populated cities. Average macro site in Poland (assuming 11 thousands AP per operator and ca. 400MHz bandwidth achievable in near future) could hypothetically allow reaching up to 1.4Gpbs per sector. After reaching this boundary there would be no more capacity available at macro sites, and the way to improve the situation would be to start deploying small-cells to densify the coverage [132]. It is expected that such boundary will be reached in 2024/25. On the other hand the introduction of e.g. cell-free architecture and scheduling algorithms can bring significant improvements to the value of SINR across the entire network. Based on the results shown in [12] [11] the capacity

improvement by applying the new paradigm can reach ca. 50%, with overall SINR improvement. And this is possible without investing in new spectrum.

2.6.1 Modelling adaptive modulation and coding (AMC)

As regards the AMC mechanism, its goal is to satisfy the Shannon’s theorem bound by measuring channel capacity in bits per second over available bandwidth and signal-to-noise (SNR) and assign best matching MCS scheme (Sklar & Harris, n.d.). Conventional solutions to the AMC problem includes the fixed look-up table, also called inner loop link adaptation (ILLA) and the outer loop link adaptation (OLLA) technique, which further improves the look-up table by adapting the SNR thresholds [134]. The SNR thresholds allow selection of the correct modulation and coding rate – see Table 8 for example in 5G. The AMC works by measuring and feeding back the channel SNIR to the transmitter, which then chooses a suitable MCS from a codeset to maximise throughput at a given SNIR, here the truncated Shannon model applies to enable modelling of the throughput with AMC mechanism [135]. As a result AMC enables enhancing the effective coverage of the cell. However as authors in [134], [136] show, using the RL-based methods, an important constraint that inner-loop AMC algorithms are depended on (i.e., BLER target) can be eliminated and thus the whole link adaptation mechanism for 5G-NR networks does not depend on predefined fixed parameters. More flexible adaptaiton and higher throughputs (higher spectral efficiency and lower BLER) is thus possible, without pre-tuning for certain scenarios.

Table 8 5G NR - SNR values for MCS switching [131]

CGI index	modulation n	code rate x 1024	efficiency	CGI index	modulation n	code rate x 1024	efficiency
0	out of range			0	out of range		
1	QPSK	78	0.1523	1	QPSK	78	0.1523
2	QPSK	159	0.3177	2	QPSK	120	0.2344
3	QPSK	449	0.8777	3	QPSK	193	0.3777
4	16QAM	178	1.1456	4	QPSK	308	0.6076
5	16QAM	490	1.9141	5	QPSK	449	0.8777
6	16QAM	616	2.4903	6	QPSK	602	1.1758
7	40QAM	486	2.7305	7	16QAM	378	1.4788
8	40QAM	567	3.3223	8	16QAM	490	1.9141
9	40QAM	666	3.9023	9	16QAM	616	2.4903
10	40QAM	772	4.5234	10	64QAM	496	2.7305
11	40QAM	873	5.1152	11	64QAM	567	3.3223
12	256QAM	711	5.5547	12	64QAM	696	3.9023
13	256QAM	797	6.2386	13	64QAM	772	4.5234
14	256QAM	885	6.9141	14	64QAM	873	5.1152
15	256QAM	948	7.4903	15	64QAM	948	5.5547
Downlink – 5G NR				Uplink – 5G NR			

As shown by [137] by adjusting settings of AMC it is possible to adjust strategy for achieving target transmission efficiency based on a user count and their locations. The most recent approaches to controlling MCS switching policy is by applying adaptation to OLLE scheme applying learning to identify optimal switching per user [138], [139]. It is shown that RL-based solutions if not outperform the legacy solutions they are very high in ranking of alternative methods. Another option for further improvements is to use more effective correction codes, such as non-binary nbLDPC codes [140]. Channel coding

schemes are used to help reduce the SNR requirements by recovering corrupted packets that may have been lost due to burst errors or frequency selecting fading. In 5G the LDPC codes are applied to data physical channels whereas the polar codes are utilized to protect the control channels. The use of this type of codes, e.g. compared to RS-CC convolutional codes, allows to decrease the threshold SNR values for switching the less robust modulation in the AMC mechanism. As a result, it is possible to improve transmission efficiency (increase in link utilization, decrease in P_d , decrease in P_b), although the size of the improvement will depend on the radio conditions of users in the cell.

2.7 ADMISSION CONTROL – ALGORITHMS STUDY

With reference to *Table 7* presenting typical groups of control algorithms the following section provides an overview of articles that address the problem of controlling resource allocation. This chapter has been divided into a section devoted to "non-learning" algorithms and a section in which the author collected the most relevant papers describing solutions using „AI/ML learning techniques” in the control implementation process. The problem of assigning resources to users in multi-service systems was originally described, among others, in [118] and [141]. The authors assume the removing from the system requests that have not been accepted, the so-called model with the removal of calls (blocked calls cleared - BCC), because as noted, the analysis of a system in which rejected calls are buffered by BCH (blocked calls held) would be difficult. Telecommunications systems using the BCH mode are popular, among others, in US. A dual problem is the assumption of blocking probabilities for class M and then minimizing the total capacity (C) - one way to minimize the required total capacity is to degrade flexible connections (or VBR media streams), if any are present in the network.

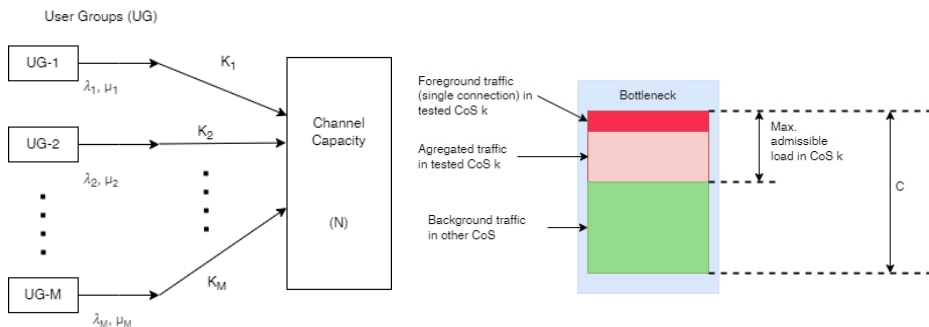


Figure 12 Multi-service system model (a) [141] (b) [63]

A cross-sectional analysis of the performance of multi-service mobile networks supporting user mobility in TDMA networks for various bandwidth allocation

control policy/channel assignment schemes is presented in [142]. For the purpose of the analyses, the authors assumed a system equilibrium in terms of user mobility, i.e. the average number of incoming calls to a cell is equal to the average number of outgoing calls. It was assumed that calls for which there is no room in the system are rejected (BCC) and that a certain number of channels is always available in a cell (fixed channel assignment). Basic methods were presented by the authors just mentioned, including, m.in: complete *sharing*, complete partitioning, *partial sharing*, *sharing with priorities*. [118] presents a comparison between the total bandwidth size requirements (C_{max}) for two different strategies – complete share (CS) and dedicated bandwidth, for each service class. The amount of the required total capacity is examined from the perspective of: different traffic profiles of individual classes (required throughput and intensity of inflow of notifications), and the total volume of traffic entering the network. The authors in [75] point out to three different types of problems arising from the need for control at the level of CAC algorithms:

1. MINO: minimizing the linear function of the target probabilities (P_B, P_D)
2. MINB: for a given number of available channels, minimization of values at a given limit for ongoing connections ($P_B P_D$)
3. MINC: minimizing the number of necessary channels assuming limit values of both probabilities, i.e. (P_B, P_D).

The use of the „MINO” approach allows to minimize the costs resulting from exceeding the permissible levels of both probabilities, which in practice means striving to maximize the operator's profits. The MINB-based approach focuses on prioritizing transferred calls while striving to maximize profits. In turn, the MINC approach transforms the CAC problem into the problem of appropriate network design depending on the assumed traffic conditions (data, user mobility). The analysis of the problem of determining the required capacity, assuming maximum blockage probability levels, is dealt with, for example, by [118], [50], the latter uses Markov chains for modelling.

2.7.1 Non learning based admission control

Admission control algorithms can be classified according to method used to assess current load of the systems (Figure 5). In parameter-based admission control (PBAC or DBAC) information about current state of the system's available resources is based solely on declarations made by applications. Therefore, the performance of this kind of admission control is highly dependent on accuracy of the declarations, availability and types (depending on the system) of descriptors. Another approach is using traffic measurements to estimate the current system load. This technique is used by MBAC (measurement – based admission control) algorithms. PBAC algorithms base most of their calculations on declarations of users. One of the challenges is to estimate the incoming traffic characteristics using only provided descriptors. Thus it can prove hard to estimate

required resources in a system utilizing Adaptive Coding and Modulation (ACM). Applications usually express their bandwidth requirements in bits (bytes) per second. PBAC algorithms seem more suited for systems where it is easy to properly describe flow characteristics (e.g. CBR traffic is usually easily described) and the required slots / symbols of a given flow do not fluctuate significantly over time (due to e.g. variations channel conditions). A technique called Complete Sharing (CS) assumes that all connections are accepted as long as the system has sufficient resources to serve the new call/connection [118] [66]. This technique is the least complicated CAC algorithm and at the same time it is easy to implement. In this technique, the base station accepts calls until the available bandwidth is completely saturated. Parameters such as, for example, the traffic handling class of the service are not taken into account. This technique is easy to implement but is only effective if we only deal with one class of traffic. In networks such as 4G/5G, where we are dealing with different types of traffic (QCI, slice), it can quickly become ineffective therefore, within the scope of this work, the CS-CAC total allocation algorithm will be used as a reference in the measurements of the proposed algorithm modifications.

However even when the equalities from Table 4 is not met it can happen that analysed system will be stable for other time periods, and this way allow handling the congestion. Here some authors to learn optimal policies rely on using an approach enabling modelling of variability based on dynamics modelling using LPV (Linear Parametrically Varying) models - an example of such an approach is the modelling of a system of networked web-services described in [143]. In here, the authors perceive admission control as a congestion protection mechanism that allows calls to be rejected at peak times to provide performance guarantees for ongoing calls. Such approach falls under controller synthesis task similar methods include non-linear approaches, such as MPC (model predictive control), GS (gain scheduling) and LPV. Recent implementations of controllers of this type, based on the LPV approach, make it possible to guarantee performance and stability parameters in the domain of control systems, however they are less popular in the wireless mobile systems control.

2.7.1.1 Measurement based CAC

The problem of estimating available resources of the PBAC approach. can be mitigated (to some extent) by focusing on measurement-based algorithms (MBAC), coupled with appropriate congestion control algorithms. MBAC algorithms seem more suited for systems where flow characteristics are not easily defined (or available traffic descriptors are not sufficient) and the required slots / symbols of a given flow can fluctuate significantly over time (due to e.g. variations in channel conditions). Although new connections requirements still have to be obtained through declarations, the percentage of bandwidth being used in reality by ongoing connections is a known value (usually at a base station level) thanks to measurements of traffic. A classic approach to admission control in cellular networks assumes allocation of dedicated resources for higher priority

calls / connections (so called Guard Channel - GC)[144]. In this technique fixed part of resources always remains reserved for higher priority connections (so called Fixed Guard Channel). Authors in [145] assume a single cell configuration to assess uplink CAC, where the admission criterion of the new user depends on the difference between the total and requested number of Physical Resource Blocks. Other results considering multi cell deployment scenarios are presented in [146] where authors describe and compare static and dynamic CAC in LTE. Additionally, a delay-aware connection admission control algorithm is proposed and evaluated. In [147], the authors present a variation of the baseline approach, based on measuring the intensity of requests instead of measuring throughput i.e. λ_{Σ} is the sum of the intensity of the current traffic in the system ($\lambda_{current}$) and the call intensity for a new session ($\lambda_{new_session}$).

2.7.1.2 Packet loss based CAC

Most CAC algorithms operate based on the controlling the degree of available bandwidth consumption, the number of simultaneous connections or the maximum bit rate in the air interface. Once the operator-defined, acceptable resource utilization threshold (C_{TH}) is reached, a new call can only be accepted after one of the ongoing calls has ended. Alternative approaches focus, for example, on packet loss level (IPLR) control in the case of real-time services. The presence of a resource controller of flexible services in the network makes it possible to regulate the level of real-time connection losses at the required level, at the expense of lowering the bandwidth utilization level, through the use of e.g. protective mechanisms (FEC, H/ARQ). The use of packet loss ratio (IPLR) as a decision parameter is, for example, used in the centralized CAC approach for real-time traffic [148]. Similarly, in the [74] project, the allocation of resources for real-time connections is based on the analysis of the network state - the network state is determined on the basis of the level of packet loss (δ_i) of real-time traffic streams. An increase in the level of losses above a certain threshold (η) is a signal to lower the threshold of maximum allowable amount of *real-time* traffic in the system in favour of the elastic/flexible traffic (NRT). In turn, the authors propose the use of the "packet loss penalty concept" instead of the typical measurement of packet loss levels. The authors show that after reaching a certain threshold of the minimum reserved bandwidth (e.g. VOIP, MPEG) referred to the total capacity of the link, the blocking probability level (P_B) changes abruptly. The authors in [149] show that all tested algorithms poorly cope with predicting stream performance expressed by the means of packet loss level indicator (PER). Simultaneously in the paper [150] are presented reasons for which the control of the level of packet losses is difficult. The problem with this approach is that improper decisions at the CAC level can lead to an increase in the level of losses, which will be up to 10 times higher than the quality objectives set for a given measurement period.

The authors in [151] draw attention to the role of measuring buffer occupancy in the transmitter (UE) from the perspective of estimating the level of packet loss in

a wireless network – as the level of latency can be a valuable indicator of the level of losses in the network (especially during mobility, NLOS, etc.). To counteract the increase in losses, the AMC mechanism is used, but it is also important to affect the traffic parameters on the side of the final application itself (e.g. video codec parameters). The authors in [152] address the problem of sub-optimal video transmission in LTE networks with guaranteed QoS. Authors present algorithm for modelling the video transmission by the additional buffer located in the LTE base station (NodeB) – i.e. downlink direction. The aspect that is optimized is video transmission using the encoded VBR stream. Authors summarize that striving to provide instantaneous compliance of the video stream with the service level agreement (SLA) is less beneficial than the ensuring of long-term compliance with the SLA.

2.7.1.3 Dynamic guard band CAC

In case of dynamic admission control algorithms the implementation of this approach consists of setting a limit on the maximum probability value P_D , while seeking to maximise bandwidth utilisation (*utilisation*), by reducing the blocking probability P_B . This approach is in line with the operator's revenue maximisation strategy (also referred to as “Case A” in section 2.4). Under certain limiting assumptions, the optimal solution is to give higher priority when accepting continuation (*handover*) calls, using a 'guard channel' strategy. This strategy is the optimal fixed CAC strategy for case 'A' presented above [153]. In networks offering multiple service delivery levels (*multi-class/grade*), distinguished by varying bandwidth requirements and associated with varying service prices, the objective of the admission control function may be to maximise operator revenues. In such a situation, a greedy strategy will not necessarily be optimal. The authors in [154] specifically focus on algorithms derived from the underlying 'guard band' algorithm. Such a solution has been proposed by [155], [156] for mobile networks. In this approach, a fixed part of the bandwidth is reserved for handling high-priority calls (so-called static guard band). The DBRAC algorithm, defines a guard band that covers transferred calls and real-time calls together [154]. In addition, it allows an ongoing adjustment of the threshold value simultaneously to the intensity of both call types (VBR, HO). For all VBR calls, a bandwidth reservation threshold is determined, which takes into account the difference between the MSTR - MRTR values, multiplied by the ratio of the β . This ratio determines the effective bandwidth of the wireless system based on quality of service (QoS) requirements and an acceptable blocking probability value (P_B). The authors showed that the DBRAC algorithm can reduce P_D under different parameter configurations: λ, β - with an unchanged number of reserved resources. Furthermore, it was shown that by using the DBRAC mechanism, the same target P_D can be achieved by reserving less resources (bandwidth). Consequently, the level of resource utilisation is improved (B_{util}).

The authors in [157] focus on increasing the fairness of bandwidth allocation to users in a 4G system, by defining a dynamic threshold value for

available bandwidth. This approach is based on the definition of the so-called 'bandwidth allocated to the user by the service provider' (LB). To control the fairness of bandwidth allocation, the algorithm defines a threshold value (TH_i) that must be provided as a "reserve" in the available bandwidth B_a so that requests from users who have so far underused their bandwidth allocation (i.e. S_i) are treated fairly with calls from users making intensive use of their subscription. Therefore, if the condition is met $B_{req,i} \leq B_a < B_{req,i} + TH_i$, the dynamic determination of the threshold value TH_i will make it possible to effectively reduce the probability of a new connection being blocked for connections with a low bandwidth usage rate, i.e. S_i . Thus, the operation of the algorithm appears to be most justifiable for 'higher network load' situations and when there are disparities in bandwidth usage between users, with analysis of the request history (and its prediction) of the other users. The authors compare their proposed solution with the [154] algorithm. The scaling factor for real-time connections (UGS, rtPS) is set $\beta = 0.2$ behind [154]. In contrast to [154], however, the authors do not include continued calls (HO) in their analysis - the latter calls do not affect the fairness of bandwidth allocation, only the total amount of bandwidth (C). In an evaluation of the algorithm and a comparison with the DBRAC algorithm, it was shown that the main advantage of this approach is the introduction of a linear dependence of the blocking probability (P_B) of user connections U_i from the bandwidth utilisation rate η_i in favour of reducing this probability for user connections for which the bandwidth utilisation rate is lower (Figure 13b). While the bandwidth utilisation for both solutions is similar, the aggregate value of the blocking probability for the FCAC algorithm is slightly higher (3-5%) - Figure 13a.

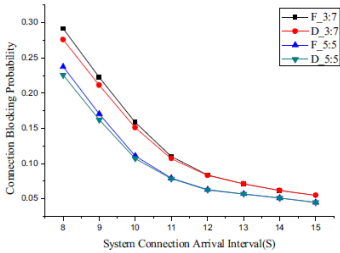


Figure 2. System Connection Blocking Probability of the Two Schemes

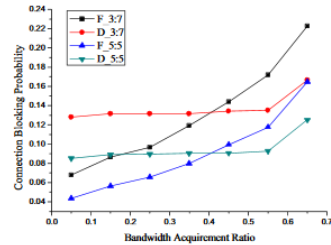


Figure 3. Connection Blocking Probability Distribution (1/2 - 9)

Figure 13 Blocking probability for FCAC and DBAC (a), blocking probability for various bandwidth acquisition ratios β (b)

The author in [53] proposes a systematic approach to the implementation of CAC algorithms for 4G networks, enabling the simultaneous definition of (often conflicting) objectives of service providers and users. The main modification is the inclusion of a dynamically selected objective function (maximising operator revenue and/or maximising utility on the end-user side). Furthermore, the author decomposes the control problem into two sub-problems, i.e. resource control for

each direction separately (converting a two-dimensional problem into a one-dimensional one). The bandwidth (B) is divided into M -independent classes, and consequently the CAC policies are divided into M -independent policies. The author defines a three-phase algorithm based on providing both constraints at the same time: i.e. greedy matching of gain values, assuming fair (weighted) allocation of blocking probabilities between classes. The bandwidth remaining as a result of the algorithm is divided among the remaining classes of service according to the maximum profit criterion.

The approach to dynamically update the threshold guaranteeing an appropriate level of target probabilities (P_B, P_D), is similar to that of the authors in [153] who propose the *Dynamic Channel Assignment (DCA)* algorithm. The authors recall that there are studies in which an optimal CAC strategy for stationary conditions was determined for VBR traffic [153]. However, such approaches are not practical due to: (a) the varying peak bit rates of different types of real-time services, (b) the variability of traffic conditions in the cell ($\lambda(t), \mu(t)$). Therefore, a version of the DCA algorithm for VBR traffic is presented. Dynamic bandwidth allocation algorithms (such as MAAC [117] and the DCA [153]) allow the threshold value to be continuously controlled so that the resulting probabilities are within the assumed limits ($P_B \leq P_B^{tar}$ and $P_D \leq P_D^{tar}$). The main difference in the two approaches is that the authors in MAAC take into account the level of saturation of the TDMA frame by real-time connections in the process of updating the admission threshold. Also, the [74] project implements quality control (QoS - for RT/nRT calls, GoS - for ported calls) using an algorithm that dynamically adapts the resource reservation threshold for handed over calls (Th_{HO}). Measurement of resource occupancy in networks that are controlled by CAC algorithms is typically realised by sliding window, point sample (reading system load every interval T_{window}) and exponential averaging algorithms [73]. The authors in [54] point out that it is the appropriate selection of different threshold values in CAC mechanism settings that should take into account operator policies, map them to specific configuration mechanism settings in the cell, all using a self-optimising approach. The use of a self-optimising approach will ensure, among other things, a trade-off between capacity and quality objectives, expressed through an appropriate set of KPIs.

The authors in [158] argue for moving away from the need to define the maximum threshold explicitly, in favour of dispersing moments of admission with the help of probabilities of α_i . They select the probability level P_D as the main parameter for QoS assurance, its role is increasing due to the increasing number and dynamics of transferred calls, as the trend of micro and pico-cells intensifies. The proposed algorithm is designed to provide stable protection against sudden spikes in call intensity. This approach is a variation of the baseline guard threshold (GC) scheme and is known, as fractional guard band (FGC) [159]. Whenever the channel occupancy exceeds a certain threshold value C_{Th} , the GC policy rejects new calls until the channel occupancy falls below the

threshold. In the fractional-generation method (FGC), new calls are accepted with a probability that depends on the current occupancy of the base station resources (i.e. it decreases as the occupancy increases). Here we are dealing with randomisation, which governs the probability of accepting a new connection [75]. The advantage of the fractional FGC technique over the baseline GC scheme, is that the acceptance of new connections is spread evenly over time, leading to a more stable control process [160]. It has been shown that, due to resource reservation in advance, for CAC mechanisms with reservation, performance in cellular systems can be modelled using an upper bound, even if a constraint on call blocking probability is not explicitly specified. This upper bound is related to call and mobility characteristics through the average number of calls transferred per call. Furthermore, the achievable capacity decreases as the cell size decreases and as the connection time increases.

2.7.1.4 Algorithms using link degradation

The results in [157] show that the rate of fair bandwidth allocation is dependent on the volume of a class traffic, as the number of ertPS/rtPS (and also UGS) connections increases the rate of fair bandwidth allocation decreases. Thus, it is easiest to ensure fair allocation of resource access between classes when the network is dominated by BE and nrtPS traffic, which by definition is traffic prone to resource degradation.

In the work [161] CAC adaptation involves controlling the throughput of flexible links by: (a) reducing the throughput of some connections when the system is saturated and (b) increasing the throughput of some flexible connections when a previously active but terminated connection is terminated. In addition, the authors analyse the impact of link adaptation mechanisms (AMC, HARQ) on the total system capacity (Erlang capacity), showing how to choose a CAC mechanism and modulation scheme to increase the Erlang capacity region.

The ABD-CAC algorithm presented in the book [162] ensures equal priority for handover calls. The algorithm is designed for a heavily loaded network, while guaranteeing high bandwidth utilisation and low wasted bandwidth (*waste*). The algorithm uses the idea of a 'bucket of tokens' to provide delay and bandwidth guarantees. When there is a shortage of resources for real-time connections, resources for a new connection are recovered by degrading nrtPS and rtPS connections. The degradation of individual connections is implemented adaptively - i.e. in relation to the bandwidth requested by the connection. The proposed algorithm, compared to the algorithm with a 'fixed bandwidth degradation step' [156], does not provide an improvement in the probabilities P_B and P_D for either class, while it noticeably increases the bandwidth utilisation of the system, especially for high network loads, precisely by adapting the degradation level to the needs of the request being served. Furthermore, it was shown that for nrtPS, rtPS connections, as the intensity of the requests increases, the profit for the operator (defined by the revenue parameter) increases. The authors emphasise that achieving a fair distribution of probabilities between P_D

and P_B for all types of calls is a task for the network designer. On the other hand, it has also been proposed (following [163]) to use a dynamically modified threshold of reserved bit rate (B_{th}), whose value is subtracted from the maximum bit rate (MSTR) of each class, which translates into an effective degradation level of ongoing connections. The variation of the threshold should follow the variation of the traffic intensity in the cell. The weighted values of the GoS parameter for each class of service, are inserted into the formula for the cost function (CF). The value of the threshold B_{th} is optimised on an ongoing basis, responding to changes in incoming traffic intensity to minimise the cost (CF).

The DHCAC algorithm is described in [164], as a typical control algorithm for accepting requests with bandwidth reservation. For the purpose of the simulation, the authors assumed the existence of three traffic classes (UGS, rtPS and BE) due to the fact that the values of the traffic descriptors for nrtPS and rtPS connections are similar. These two classes differ primarily in their requirements for the level of delay, not in the bandwidth itself. Algorithms based on bandwidth reservation, on the other hand, generally only consider the available and requested bandwidth. Parameters such as delay or jitter are therefore not typically considered. It can therefore be assumed that in the case of admission control algorithms based on bandwidth reservation, there is no fundamental difference between rtPS and nrtPS traffic. For UGS connections, part of the bandwidth is reserved (U), while connections of the other classes are described by the minimum value of the desired bandwidth. The authors assumed that also for BE connections a minimum value of bandwidth is specified. In the case of insufficient remaining bandwidth (i.e. B_a) in order to accept a new connection of classes rtPS and nrtPS with the required bit rate B_{rtPS} , B_{nrtPS} , a mechanism is triggered to **evenly acquire bandwidth from connections of lower classes**, in the order defined according to the priority $P_{BE}^R > P_{nrtPS}^R > P_{rtPS}^R$, where P_x^R denotes the priority in recovering part of the bandwidth. Thus, if the bandwidth of an incoming rtPS connection exceeds the available bandwidth (B_a), a check is performed sequentially to see if, after degradation of the class bandwidth according to the priority order by a value of $\sum B_i^k - B_{min}^k$, it will be possible to accept a new connection B_{rtPS} , B_{nrtPS} . The decision rule of the DHCAC algorithm for a new connection (B_x) of a type other than UGS, is as follows:

$$if (B_{used} + B_x \leq B - U) \quad 2-5$$

The above approach can be problematic and lead to inefficient use of bandwidth. This is because it does not take into account the bandwidth already used by UGS connections. Furthermore, in the worst case scenario, i.e. when a value of $B_a \ll B_x$ excess bandwidth may be received ($B_i^k - B_{min}^k$) of all classes other than UGS. The degradation process itself is a non-trivial issue, as each group of applications represented by a given class of service will be differently susceptible to dynamic bandwidth modifications. TCP connections, due to the built-in congestion control mechanism (AIMD), will respond automatically, whereas UDP connections will

only respond if the applications use feedback protocols to inform the source of changes in network conditions (e.g. RTCP).

In the work [165] the algorithm implements capacity allocation between classes and, at the same time, by monitoring the distribution of incoming traffic, the shares of each class in the available bandwidth are periodically updated. A logical separation of resources for HO connections is foreseen, in order to provide P_D at the lowest possible level for each class in order to increase the level of satisfaction and resources. In the paper, base station (C) resources are divided into three areas: reserved capacity for HO calls (C^h), reserved capacity for new calls (C^n) and shared capacity for calls for which there was no reserved pool space (C_s). The objective of the algorithm is to maximise bandwidth utilisation given constraints on the maximum levels P_D and P_B of each class. The capacities assigned to each class are updated periodically ($T=2\text{min}$) in response to changing traffic parameters ($\lambda_j^n, \lambda_j^h, \mu_j^n, \mu_j^h, h_j$). The results of the comparison with the algorithms (Dynamic Complete Sharing, Dynamic Partitioning Scheme) show the advantage of the dynamic resource allocation mechanism over static approaches, in particular it can be seen that the quality of the connections (expressed by the limiting values of the probabilities) is guaranteed for a higher intensity of the incoming connections. On the other hand, [166] analyses aspects of modelling the performance of a wireless system with bandwidth sharing between classes carrying CBR, VBR and ABR traffic in an ATM network compatible model. The author assumes a CTMC chain as the system model, while network states for which the sum of connections would exceed the available capacity value result in the compression of the bandwidth of selected connections (i.e. degradation).

2.7.1.5 Algorithms incorporating the AMC mechanism

Even though the AMC mechanism is designed as a typical physical layer element, it has a real impact on the higher layers of the ISO/OSI model [167]. Therefore, in wireless networks such as 4G/5G that support the AMC adaptation mechanism, the impact of such solutions cannot be ignored. On the other hand, the level of dynamics of modulation and coding changes will be strictly dependent on the scenario, i.e.: terminal movement speed, cell size and link propagation conditions. Most existing studies on CAC algorithms consider two types of call streams: new calls and transferred calls. However, relatively few authors address the problem of call rejection due to the AMC mechanism. The authors in [168] propose a solution for the CAC mechanism with a static guard band threshold and using adaptive modulation and coding. Additional guard band has been applied to handle calls for which the modulation has changed (e.g. a drop from high performance 64-QAM modulation to low performance QPSK modulation). In contrast, the analysis was carried out for one call type and two modulations (QPSK, 16-QAM). In [169] [170], the authors undertook an analysis of the performance of CAC mechanisms in networks using the AMC mechanism for a wide range of wireless networks with a single traffic type. It was shown that

considering the specifics of the AMC mechanism in the design of CAC solutions leads to improved network performance. In the work [171], the CAC procedure is invoked in the following cases:

- Creation of a new connection
- Changing connection parameters
- Changing the MCS scheme for the connection,

Each of the above conditions can increase or decrease the resource level of a given connection. An improvement in radio performance is always accepted in practice, but the allocation of resources due to deterioration of channel conditions leading to a reduction in modulation value depends on the residual capacity at the base station. A lack of resources to implement a connection with a higher level of protection will typically be associated with the rejection of the connection.

The paper [162] shows that work on guard band threshold adaptation, which considers the dynamics of modulation and coding changes, can lead to a balance between the levels of blocking and rejection probabilities for different links and optimise the use of radio resources. In the situation of the capacity changes described above, related to the operation of the AMC mechanism, it is important to bear in mind the existence of alternative solutions to support adaptation on the network side but also on the application side of the terminal. On the network side, one possibility to counteract such a situation is to proactively use, initiated by the network, the transfer of selected connections to neighbouring BS stations. This will have the effect of recovering additional symbols in the TDMA frame to make calls from for channels with a reduced SNR ratio. Considering the variability due to radio channel dynamics and AMC mechanisms naturally involves the need to take into account the degradation of selected connections if the available resources decrease abruptly or very significantly.

2.7.1.6 CAC algorithms for excess traffic

Although the authors in [172] focus on wired networks, it seems to be a universal approach to take into account in the admission control process, temporary excess traffic resulting from problems in the network (broken link, packet re-routing). In such a situation, the role of the CAC is to prevent such an increase in traffic that would result in a disruption of the quality objectives defined for the supported classes of service. To counter such situations, the authors limit the number of streams that can be accepted in a given time window. One of the authors' final conclusions regarding the manageability of traffic resulting from unexpected network events (flash crowds) is a recommendation suggesting that the CAC mechanism alone may not be sufficient, congestion control algorithms should be used simultaneously. Another typical case of a sudden increase in traffic volume is mass events. As the authors show in [173]for LTE networks, the amount of data per subscriber for the uplink increases more than four times during mass events, compared to a normal day. The increase in resources used is due to the fact that the activity is heavily dominated by the uplink, and since the uplink has, in

general, lower spectral efficiency compared to the downlink, it takes longer to deliver a similar amount of data in the uplink. Therefore, admission control solutions should take into account the possibility of a sudden increase in traffic flow intensity. The authors in [117] propose two CAC algorithms for real-time traffic in 4G networks. Both solutions use a moving average (EWMA) [174] of the number of free OFDMA symbols observed in consecutive frames ($freeSlots_i$) by the serialisation mechanism. The subject of the calculation are the symbols remaining after the traffic offered by the real-time connections has been accepted, in a given frame i . Averaging makes it possible to reduce the reactivity of the CAC mechanism. The first algorithm (MBAC) uses the value of the averaged number of remaining time slots as the CAC decision parameter, and as long as this number is greater than a fixed limit (e.g. 10 slots) a new request is accepted. The algorithm does not deviate from the assumptions of the CS-CAC [118] basic mechanism. However, this approach does not provide protection against an influx of calls that do not follow a Poisson distribution, i.e. in the form of *bursts*, the authors propose an approach that takes into account the continuous adaptation of the dynamic booking threshold (MAAC algorithm). For a newly arriving call, the condition is checked for $B_{reservedBW}^{RT} + B_{req}^{RT} < limit$. If the inequality is satisfied a new real-time connection with a bit rate of B_{req}^{RT} is accepted and the value $B_{reservedBW}^{RT}$ is increased by the MRTR (Minimum Reserved Traffic Rate) bit rate set in the subscription of the newly accepted connection. The setting of the limit value (limit) is implemented periodically, according to a rule known from TCP protocol implementations, i.e. *Additive Increase Multiplicative Decrease* (AIMD) [175]. In addition, the appropriate choice of update frequency of the limit value (limit) is worth considering in such an approach. The authors of the Blue algorithm [176], which controls the size of the packet rejection probability in the router's broadcast queue (Pm), suggest including a variable controlling the length of the period without updates (*freeze_time*). A similar approach to dynamically adjusting the value of the threshold (Th) determining the ratio of dedicated resources to new and transferred connections was used in the work of [74] [148]. These algorithms also use capacity estimation in the next TDMA frame, using the rolling moving mechanism (EWMA). The advantage of this approach is that it replaces the analytical aspect (analysis of the optimal/maximum loss level) with a reactive approach (the threshold determining the RT/NRT link ratio changes when the loss level of real-time regime link packets, exceeds a certain threshold value η).

In the work [177], the authors analyse the congestion control algorithms and uses dynamic multi-model adaptive exponential smoothing (DMMAES), to calculate the optimal smoothing coefficients and weight of each mode to speed up the 'vehicle congestion' prediction. The authors in [178] describe the resource utilisation rate of the base station is by three admissible states: idle, busy, saturated, depending on the system load. With this approach, the authors plan to adapt the adopted CAC strategy to the system state. When an incoming call

requires a minimum amount of bandwidth that is not currently available, it is assumed that normal and on-demand calls can be degraded (classes of service are arranged according to priority). Due to the assumption of the presence of the AMC mechanism, the instantaneous bit rate value is related to the SNR value of the j -th connection in i -th class. In addition, the authors divide the space of possible system states into three subsets using thresholds: ρ_{busy}, ρ_{sat} . Exceeding the threshold value of the system load changes the CAC decision rule according to which the decision process is implemented. Here, a similarity to options approaches [179] is apparent, where the state space is divided into smaller parts.

2.7.2 Learning based CAC algorithms

Recent papers overviewing the use of machine learning to support RRM in general, and slice and user admission control in particular indicate the essential role the learning-based approaches have [180]. There is nice summary of achievements feasible with the use of machine learning (ML): supervised, unsupervised, reinforcement learning as well as deep reinforcement learning (DL) approaches and their potential in supporting the slice management. Authors indicate superiority of DL techniques (e.g. LSTM) over alternative solutions of ARIMA or Holt-Winters for seasonal prediction. The main target for prediction is the “total traffic per slice” in the 90% of the papers. It can be seen that with regards to admission control the promising combination of techniques is the use of prediction (LSTM, GRU, DQN) with the RL model for resource allocation optimization (MDP, sMDP models). Only minority of papers deal with actions different than “admission/rejection” – few of the slice-level admission deal with “scaling resources up/down” when admitting slice. And absolutely few papers at time of writing this thesis refer to the use of digital twin models of RAN networks in order to design optimal resource allocation strategies [81]. It can be seen there are multiple inherent trade-offs between user admission probability and QoS satisfaction, a trade-off between network reliability and resource efficiency, etc. (see section 2.4.1 for more on relation between CAC and other mechanisms). In all of the existing RL solutions, those trade-offs are handled inside the reward function, by defining a weighting factor between two objectives – here the multi-objective RL (MORL) is known to deal with multiple-objectives at time [181]. The reward function is defined as vector not scalar, it can help solve problems with conflicting objectives.

According to the papers above forecasting of end-user traffic is still a missing piece of the network slicing problem, as most researchers focus on slice level aggregated traffic prediction. Another aspect for future research is trade-off for computing complexity and accuracy in the training of DL algorithms used for network slicing. Recent papers on approximate computing which indicates trend of more lightweight computations being also under intense research nowadays [182], [183]. Desired level of data needed to train an agent can be the limiting factor, which requires to carefully design frameworks to perform training

efficiently with the available data. Based on [180] findings the number of studies focused on sample efficiency is considerably low and this is an area where additional efforts are expected. Also combining multi-armed bandit (MAB) with federated learning is foreseen as essential directions towards scalable learning frameworks in the future networks.

Markov Decision Processes (MDP) have a great deal in modelling and solving sequential decision problems for discrete time (MDP) and continuous time (semi-MDP) models. The interest in Markov Decision Processes is solely motivated by the fact that MDP formulated models are quite simple to understand and implement, thoroughly described by rich mathematical theory and have proven to be feasible in sequential decision problems where outcomes are uncertain. MDPs have already been used in a great number of applications ranging from economics, engineering, ecology, medicine, business and communication theory. Also as described in [184] MDPs have proven to derive optimal decisions in discrete-time queuing systems like admission control in ATM networks (Nordstrom & Carlstrom, n.d.) (Nordstrom & Carlstrom, n.d.) and when making routing decisions [145]. Call Admission Control problem in wireless networks is in fact a sequential decision problem where actions are chosen at call arrivals and the outcomes of the action choice are evaluated in order to provide the optimal performance of the system. Therefore an agent/controller is faced with the problem of choosing an action in order to maximize a given cost function (i.e. future income of the service provider or bandwidth utilization). This reflects the situation where a service provider deploys a CAC agent that follows a certain policy and chooses actions on call arrivals. The CAC agent may choose to accept or reject the incoming call and the action choice has an impact on the future sum of rewards (income) received from users or the bandwidth utilization rate. It was proven in numerous studies that MDP models can be used to model the CAC problem in wireless multimedia networks [186] [187] [188]. However, a CAC controller has to take into account the trade-off between the service provider revenue and violation of Quality of Service constraints [186] [189]. To derive an optimal CAC policy and simultaneously achieve better QoS the Markov Decision Processes can be used. Still this method is somehow limited by the Markov property itself which states that the system is memory less. MDP formulated CAC problems make some general assumptions like call arrivals following a Poisson process and exponentially distributed call holding times. As the latter is still true for heterogeneous wireless networks - the call arrival times of handoff calls from different cells are not Poisson distributed as proven in [190]. Thus, the authors in [191] propose a Generalized Semi-Markov Decision Problem model to solve this issue. Another problem that needs to be considered is simply the fact that bandwidth utilization is not constant and may change in time for a constant number of users (i.e. where several users are utilizing VBR traffic). The problem of modelling wireless systems with variable capacity using MDP was described in [192]. A multidimensional Markov Chain was also used in [193] and [144] to solve the CAC agent problem in wireless cellular networks. Also, a Call

Admission Control strategy for integrated WiFi/WiMAX services using Semi-Markov Decision Process was formulated in [194]. However MDP models can be applied to solve different optimization problems like for example MDP-based optimized scheduling algorithms in 4G networks, as proven in [146].

Apart from using general dynamic programming techniques (LP) to solve MDPs, alternative ways are also proposed to derive optimal policies for MDP formulated problems. The authors in [195] solve the MDP formulated problem using Reinforcement Learning [196]. They implement a Temporal Difference algorithm called Q-learning in order to derive the optimal CAC policy. This approach has proven to be feasible with MDP and the calculated Q-values tend to converge to the optimal CAC policy. Still the Q-learning algorithm suffers from the generalization problem but this can be solved using artificial neural networks to approximate the results [195] [196].

Table 9 Overview of MDP research

Group	WN	HH	CT	HPCA	MDP	BP	OBD	RC	OO	SR	O/CT	QC	RTL	STP	COD	LP	RL	OPC	DP	VLC	VH	HN	ANN	WC	
1	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
2	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
3	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
4	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
5	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
6	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green

In the table Table 9 it is possible to check what topics the six representative groups of researchers (rows) are pursuing. The highlights in blue just below the table of Table 9 indicate least covered area and the green ones *Figure 14* on the other hand indicate topics present in vast majority of papers.

Glossary

HN	Heterogeneous Network
WC	4G Compliant
ANN	Artificial Neural Network
OPC	Off-line policy calculation
LP	Linear Programming
RL	Reinforcement Learning
BP	Blocking Probability
DP	Dropping Probability
OBD	Optimal Blocking Decision
O/CT	Optimality/Complexity Tradeoff
QC	QoS Constraint
RC	Reduced Complexity
OO	Online Optimization
RTL	Real Time Learning
STP	State Transition Probabilities
COD	Curse of Dimensionality
CT	Convergence Time
SR	Storing Requirements
WN	Wireless Network
HH	Horizontal Handoff
VH	Vertical Handoff
MDP	Markov Decision Processes
HPCA	Handoff Poisson Call Arrival
VLC	Variable Link Capacity

Legend

Green	YES
Red	NO
Blue	YES (LUT) - NO (ANN)

Figure 14 Overview of topics not well covered in the resource allocation with MDP

For the analysis of the models of probabilistic systems, models are based on the

assumption that there is no dependency between states, i.e. the so-called lack of memory (CTMC, DTMC, MDP, PA probabilistic automata, PTA time-dependent probabilistic automata). Future states depend only on the current state. There are various tools supporting the process of formal modelling and analysis of random systems, e.g. PRISM, ORIS [197]. A formal description of stochastic processes enabling the specification of both quantitative and qualitative properties of the system is referred to as SPA (stochastic process algebra) [98]. One known way to carry out a quantitative analysis of such a system is to use a generalized semi-Markov model. A performance model that describes this type of system family well is e.g. continuous Markov chain (CMTC) - other models are e.g. stochastic automata. The use of these tools, apart from facilitating the design of such systems, enables e.g. determination of probability distributions for transition states. In the [198] authors implement SON agent for eNB using ANN architecture, in order to improve the reinforcement learning performance for the antenna tilt tuning in cellular 5G deployments. ANN is well known for its ability to learn from a vast number of inputs, while the stochastic learning technique relies on a simple action-based probability vector updated based on system feedback. They compare this solution with a simple Stochastic Cellular Learning Automata (SCLA).

Most recent trend in network modelling is the definition of a digital twin for a physical realm, it also regards network modelling. Among the representative references is the paper by [81] which analyses the modelling of 5G network where there is a need to deal with multi-tenant networks where capacity needs to be shared optimally. The traffic generation uses Poisson arrivals and exponential holding times, mobility pattern uses random walk, propagation model is the Urban Microcell. Capacity sharing uses SON model which interacts with DQN-MARL agents in gNBs to share the PRBs appropriately, then at gNB level the RRM policy allocates PRB to UEs. The UEs attach to gNBs based on max-SINR rule. The digital twin concept is relatively new, although already promising as it has been identified by well recognized organizations e.g., ORAN, ESA, EC among others. For example, the ESA has recently launched call for proposals indicating that “challenging aspects of the development of a new generation of mobile networks is the radio technology. Traditionally, it has required a lot of prototyping, testing and measuring inside and outside the lab. A Digital Twin of the physical environment capable of modelling radio propagation in large and dynamic environments would greatly aid this process [199].

2.8 THE FUTURE WIRELESS NETWORKS

In 4G networks, the data and control planes are not separated, without having programmability at the RAN that also limits us to provide global information of the network. The overall transition on perceiving the role and place of ML techniques in next generation networks is shown in Figure 15. It can be noticed that in the architecture part (lower plot) there will be most focus on

programmability, virtualization (5G) and later on the self-sustainability, flexibility and intelligence in 6G. The first specifications by 3GPP are expected around 2028. While in the access network evolution that is expected the prevailing features will be increased speed, network densification, targeting industrial applications with ultra-low latencies and sensing as a new sources of data.

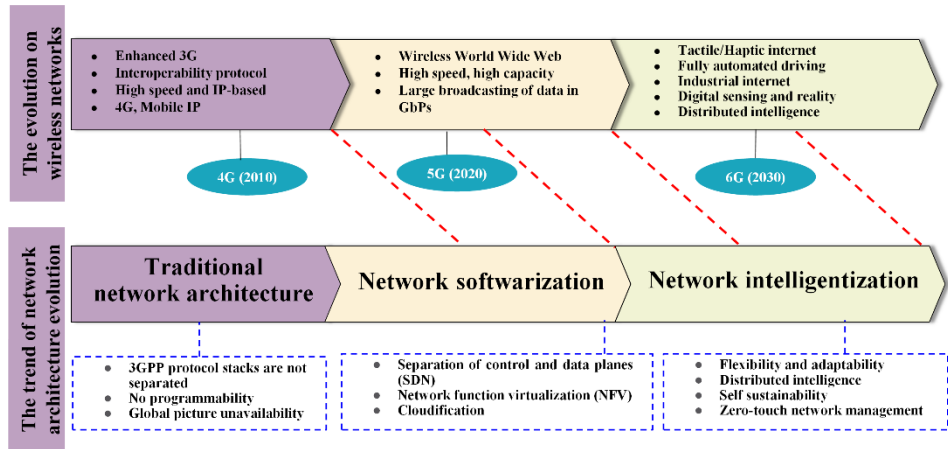


Figure 15 The transition of network architecture towards 5G and beyond networks.

The future 5G network evolution is expected to be extremely heterogeneous where femto-cell, micro-cell, and pico-cell coexist altogether. Furthermore, multiple applications will coexist in the architecture to provide different services to the underlying users. Therefore, the traditional network architecture design has to be software-driven to satisfy the diverse requirements of 5G networks. Figure 15 expresses the vision of ML in B5G networks. As shown in there the disaggregation of software from the hardware with the help of the network function virtualization concept (NFV) is enabling a flexible network architecture of 5G networks. However, the B5G networks should be flexible and self-adaptable where the distributed intelligence is based on applying ML algorithms in the network. The zero-touch network and service management and the influence of the development of the innovative algorithms based on ML will enable dynamic adaptation of the B5G networks that satisfy the demands of the users.

As is known, the user's behaviour is very dynamic and their traffic for the future B5G networks will vary over time. Therefore, it is difficult to predict their behaviours and provide an accurate mathematical model to solve optimization problems. Due to the lack of accurate models, the traditional way of solving the PHY and MAC layer problems based on mathematical models will not be optimal. This is one of the main reasons we need to learn the environment by applying ML algorithms in the networks to make an optimal decision more

accurately and optimize the optimization problems. The ML algorithms for the MAC layer tasks execution will leverage the performance of future B5G networks where the parameters of the network are unknown. Therefore, Table 10 provides a list of PHY and MAC layer problems that could be potentially solved by using several ML algorithms.

Table 10 The example of ML in 3GPP protocol stacks for B5G networks [own summary]

Layer	Problem Type	ML-based Algorithm
PHY	Signal detection, classification and compression	Deep learning with RNN
	Channel encoding and decoding	Deep learning with CNN
	AI-assisted positioning, sensing, and localization	Deep learning with ANN
	Channel estimation and equalization	Deep learning with LSTM
MAC	Dynamic scheduling of radio resources	RL, actor critic learning (ACL), and deep RL
	Power control and link adaptation	
	Interference management	
RRC	Handover Managements	RL, ACL, and deep RL
	Mobility Management	
	Slice admission and congestion control	

From the 5G virtualized networks perspective the edge-computing allows avoiding data transfer to the cloud, and thus is capable of providing low latency and processing large volumes of data. Especially micro-data centres are becoming important recently [200]. Initiative in [201] is promoting the adoption of the edge computing paradigm within the manufacturing and other industrial markets. As it is mentioned in [202] cloud service providers will need to deploy Kubernetes (or alternative edge ecosystem) at large scale with hundreds of thousands of instances at the edge. However, this distributed cloud architecture imposes challenges in terms of resource management and application orchestration. Recently launched H2020 big data processing and artificial intelligence at the network edge (BRAINE) project [22] is dealing with the design of an edge micro-data centre (EMDC) solution that is supporting the underlying innovative use cases of 5G and beyond networks which is discussed in section 7.8.

2.8.1 Beyond RRM with disaggregated vRAN

The wireless system design characteristics introduced by 5G (including among others: service-based architecture (SBA), control and user plane separation (CUPS) splits, functional splits) brings an important change to the so far black-boxed solutions of wireless network nodes. The radio stack software functionalities are decoupled from underlying hardware and allow the MNOs to remotely upgrade settings of the eNB, e.g., update from LTE to LTE-A, or from 802.16e to 802.16m. The tight coupling of software-hardware was a barrier for any modifications. On one hand the hardware (HW) underlying any eNB up till 4G was always following a vendor-specific, dedicated design (as an opposite to

COTS HW architectures). Such a deeply proprietary approach of the past, has effectively blocked extensions and introduction of innovative research results, and has limited it only to the internal works performed by relevant R&D teams of big vendors.

With the 5G, separating control and user-plane is yet another level of flexibility, as processing the two important planes can now happen on separate nodes/systems. As the functions of gNB are now “disaggregated” both in the design phase, as well as at the solution deployment stage, this creates an end-to-end (E2E) open ecosystem. Eventually the ultimate innovation enabler is the O-RAN alliance activities introducing open interfaces and the RAN intelligent controller (RIC). The latter enables plug and play approach to the usage of so called “RAN functions (network interfaces, key performance metrics, etc)” and sitting on top of them SW plug-ins also called xApps³. This way the holistic research process in the wireless systems design can now be “fostered” with easy, programmable access to the internal functions of 3GPP RAN.

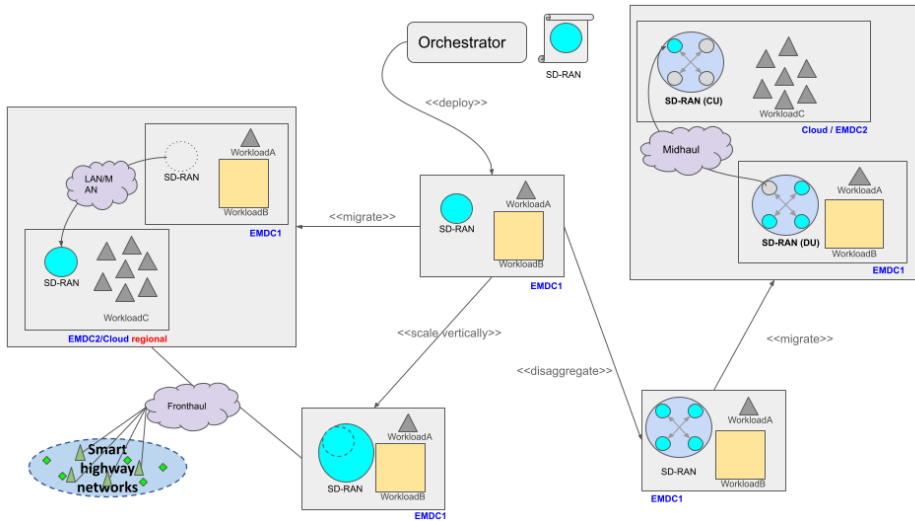


Figure 16 Orchestration of disaggregated SD-RAN.

Therefore the implementations of functions such as radio resource management (RRM), mobility and more, can in addition to the pure simulators also be taken to the new level of “experimenting in the loop of a real system” [203]. The above architectural updates in the wireless systems specifications would be lagging behind, without relying heavily on the virtualization (ETSI NFV), cloud-edge flexibility and performance as well as SDN paradigms adopted to 3GPP/O-RAN.

³ A special type of control plane application available at RAN Controller for RAN Configuration, AI/ML model policy execution, radio resource management model, slice selection, etc as per O-RAN compliance approved specifications specified by O-RAN WG-1 and WG-3

Software-defined radio access networks (SD-RAN) refer to a network software solution that combines the RAN functionalities in a disaggregated manner.

Figure 16 illustrates the innovative options of utilizing disaggregated RAN that can be deployed in the cloud-edge continuum by utilizing orchestrator. In the figure one can see evolution paths of SD-RAN workload from its deployment via an orchestrator, down to various SD-RAN footprints that depend on the required vs available resources and performance metrics. In the flow of the presented life-cycle the SD-RAN work load experiences: (a) migration between EMDCs, (b) vertical scaling within single EMDC, (c) disaggregation before migrating to another EMDC and eventually (d) moving a single SD-RAN subcomponent (e.g. MAC or PDCP) to another EMDC. The key driver here is the pursue for optimizations both at the radio or computing resource side. Smart decision making requires orchestrator to: have access to up-to-date metrics from SD-RAN, be federated with other orchestrators (e.g. across EMDCs), support smooth transition of workloads between EMDCs (e.g. with service mesh functionality), have east-west and cross-domain interaction between SDN controllers (e.g. RIC - transport network controller), have an overall and updated access to resources available across EMDCs (e.g. through knowledge sharing between the EMDCs), deploy suitable service chaining strategies to allow optimal service location and chaining. All these functions require architecture at MANO level that is amended by AI/ML models, thoroughly integrated in a cross-layer fashion, potentially utilize policy-based management with access to semantic information on: infrastructure capabilities, use-case requirements and configuration. The consequences of disaggregated RAN (e.g., functional splits) lays mainly in opening the market for the vendors/providers. However combining it with virtualization and centralization (or distribution) has multiple effects on the radio resource management:

- virtualization allows dynamic scaling or computing resources which in turn removes the bare-metal natural limits of “physical resources” of underlying HW
- software-definition of underlying network links, those which interconnect the edge data centers, brings also scalability and adaptability to the networking infrastructure, which can now get its capacity quickly adjusted by means of e.g. decomposed optical network
- pooling the processing of multiple BBU, introduces among all (i) redundancy gains, (ii) cooperation capabilities between processing nodes, (iii) opportunity for removing certain parts of wireless system signalling like for example handovers in the context of mobility under ultra-dense networks
- network provisioning can become utility based, where multiple criteria can be utilized to optimize network virtualized functions in a dynamic and adaptive way. More generally it becomes easier to create, apply and

tune any intelligent strategies for virtualized network functions (VNF, CNF) orchestration

- softwarization of the RAN opens the doors towards full network programmability. Naturally this is tightly linked with AI/ML increased presence, taking for example novel architectures such as ETSI ENI [204].

Currently, attractive approaches to the implementation of control in networks, mainly due to **open-RAN approaches and virtualization, are based on a centralized approach**. Centralization is beneficial due to the availability of quality parameters from monitoring systems (OSS), easier coordination of conflicting quality goals, a broader view of the network enables a more global approach. In turn, as the density of future networks increases (mainly due to 5G deployments), a centralized approach will have to give way to a more distributed solutions. A trend that is very conducive to the centralization of control is the increasing availability of solutions that enable software control of networks (including SDN, NFV), but also the growing density of base stations per km² in the coming years, due to the increase in the carrier frequency towards millimetre waves (i.e. above 6Ghz). As a result, future networks will be based **on the use of virtualized RAN functionality** deployed in the cloud-edge continuum, which will make it possible to co-locate baseband processing functions, and eventually cooperative control (as opposed to distributed control) will be possible [205]. As a result of this approach, the automation of management in the RAN network using SON techniques will be included in the architecture of future 5G networks, enabling the implementation of various intelligent solutions supported by machine learning algorithms [75].

2.8.2 Suitability of optimizations for B5G resource management

As mentioned in previous section the important novelty of 5G / B5G systems is the enablement of optimizations both at the radio or computing resource side. This way once 5G/B5G are deployed in the edge-cloud continuum both energy consumption, computing and radio resources control can be performed in parallel, just by appropriate set of algorithms placed at the RIC platform [206], .

The process of workload prediction and placement is designed to provide an efficient bridging between the data collection - model training and the ultimate system modernization goals. As such it can be linked with various KPIs, that should be tracked to assure particular performance guarantees expressed inside SLA agreements. It is also assumed that the underlying 5G/B5G system will follow the self-adaptive design. Although it is not specified which optimization best-practices should be followed (e.g. ETSI ZSM, ETSI ENI, etc.). In order to attain optimization objectives (of maximisation or minimization) the AI/ML models should be intertwined with the comprehensive set of radio/computation resource management techniques. Such intertwining should also consider the cross-dependencies between different techniques (e.g. performing resource scheduling and admission control should be aligned). The broadly speaking

resource management and network modernization techniques considered here include (but are not limited to): (i) scheduling of radio resource blocks (RB), (ii) admission control of new connections, congestion control of connections/resources, (iii) multiple-access scheme selection (OMA, NOMA), (iv) multi-access edge computing, (v) network management (that previously assumed availability of SON mechanisms). All these mechanisms either (a) deal with resources allocation based on optimization criteria or (b) influence and support resource allocation indirectly.

2.8.3 Workload prediction in future wireless networks

The virtualized radio access network (vRAN) is introduced as an agile approach of RAN deployment and management that facilitates network operators and infrastructure providers with the promise of better operational efficiency and improved flexibility to meet the exponential demands of 5G enterprise customers [207]. A vRAN architecture enhancements are enabling innovation towards 6G and disrupting business models. In the cellular system, the cell edge users usually face network challenges more severe as compared to the users close to the cell center. This happens due to significant pathloss and greater interference from nearby cells at the perimeter. This issue is now becoming even more challenging due to the increased density of base stations expected. The edge micro data center (EMDC) is a proposed solution for networks deployment at edge which offers customized resources (compute and storage) for applications intertwined with networking infrastructure components (i.e. mix of workloads).

The edge architectures are important for decreasing application response times and improving the ability of operators to collect and process data [208] to further improve the vertical applications as well as network performance. In a typical edge computing paradigm, multiple edge servers are placed close to the end users to support quick computation and required bandwidth. However, the escalated devices will introduce several challenges of resource management and elasticity towards vRAN in the EMDC. The accurate prediction of the future workload considering central processing unit (CPU) consumption is critical to the efficiency of vRAN resource management. Due to resource constraints of the edge servers, precisely predicting the workload of various edge servers can be of great importance in proficiently utilizing the EMDCs. The role of combined compute and radio resource scaling in case of virtualized 5G networks deployments is essential as it allows to perform the paradigm shift from a network provisioning based on a busy hour.

Several artificial intelligence/machine learning (AI/ML) based predictive methodologies and schemes are proposed to estimate the future resources demand [22], (Hu et al., 2023) [23]. One of the European Union (EU) H2020 granted projects, i.e., Big data pRocessing and Artificial Intelligence at the Network Edge (BRAINE) is currently delivering the novel HW architecture and SW middleware for the EMDC - in order to boost the performances of the edge networks. It

provides solution that combines vendor apps workloads together with 5G infrastructure in the same cluster of containerized workloads. Hu et al. in [209] proposes a containerized edge computing framework for dynamic resource provisioning. This framework integrates workload prediction and resource pre-provisioning to provide high utilization of edge resources. Pramanik et al. characterize the computational and memory requirements of virtual RANs with regression models to predict better demands for resources [23]. They use 4G vRAN testbed leveraging the non-disaggregated open-source mobile communication platform and general-purpose processor-based servers. The ML based virtualized network functions (VNFs) prediction and placement in the network edge is investigated in [210] where authors propose a neural-network model to assist in proactive auto-scaling by predicting the number of VNF instances required as a function of the network traffic. This research also investigates the placement of these VNF instances at the edge nodes with a primary objective of minimizing end-to-end latency from all users to their respective VNFs. Authors from Microsoft introduce a user space deadline scheduling framework Concordia [211] for the vRAN on Linux. It builds prediction models using quantile decision trees to predict the worst-case execution times of vRAN signal processing tasks. These predictions are used to calculate and proactively reserve the least number of cores required to perform the vRAN pool operation in the next slot (e.g., 1 ms), releasing the rest of the cores to the operating system (OS) for other tasks. In [212] authors evaluate the use of LSTM network for prediction of electric energy consumption. They show that LSTM provides better RSMA values for two data sets than ML techniques of XGBoost and random forest. They apply the created model to energy saver module that controls energy consumption. Similarly, in [213] authors show that LSTM network is accurate in predicting future requests and the system can allocate appropriate resources ahead of time. It is also used among other predictive methods (ARIMA, GRU, Holt-Winter) for the link adaptation in intelligent MAC experiments by [214]. The LSTM combined with the RL decides whether the system needs to schedule more virtual machines, avoiding unnecessary resource scheduling, especially when requests suddenly appear in peak. The RL makes optimal decisions based on historical experience and current system state.

In [215] it is shown that RL offers a promising perspective for designing cloud autoscaling strategies based on an online learning process. It has been shown that demonstrating that considering the specific characteristics of workflow applications when taking autoscaling decisions can lead to more efficient workflow executions. It is highlighted however in [216] that accurate bandwidth prediction remains a challenging task due to the short-distance coverage and frequent handover properties in 5G networks. In this work both ARMA and random forest are used. They also propose end-to-end congestion control algorithm that adaptively sets the senders' rates to the predicted bandwidth by their algorithm called HYPER.

On the other hand authors in [217] show that embedding the optimization problem in the training pipeline can improve decision quality and help generalize better to unseen tasks – as compared to relying on an intermediate function for evaluating prediction quality. A comprehensive study of various deep reinforcement learning architectures is presented in [218]. Dynamic auto-scaling rules for containerized applications are introduced by [219], they mention that even though CPU and memory utilization is the state-of-art., they propose an autoscaling method which updates resource threshold dynamically based on monitoring data from infrastructure as well as from application. Authors mention that setting optimal monitoring interval is not trivial task.

2.8.4 Multi-RAT solutions

According to [124] “Network selection plays a fundamental role in providing stable connections with an adequate level of QoS. Hence, network operators and providers commonly exploit several advanced techniques to select the best AP to allocate new connections. Among the various techniques proposed in the literature, multiple attribute decision making (MADM) proved to be one of the most flexible solutions to capture user preferences and QoE-related aspects in the decision process [220]. In MADM solutions, the information characterizing the decision-making is made by the so-called attribute values and attribute weights: i) the first ones describe characteristics, qualities, and performances of different alternatives, whereas ii) the latter ones are used to measure the relevance of attributes. By modelling the network selection problem as an MADM, it is then possible to decide the trade-off among service QoS requirements, user preferences, and overall network congestion. The recent example of global user-traffic analyses from the COVID-19 pandemic outbreaks indicate a need for robust and novel solution based on intelligent switching of downlink traffic between two different radio access technologies (RAT) - so called multi-RAT. In [221], the authors propose a multi-path based adaptive concurrent transfer scheme that allows application layer traffic transmission via user data gram protocol (UDP). The proposed method can improve the throughput of the system while satisfying the constraints of quality-of-experience (QoE), the video encoding rate, and the priority of the frame, respectively. Fast-RAT scheduling solution in a 5G multi-RAT scenario is studied by Victor et al. [222] where reference signals received power (RSRP), received signal strength indication (RSSI), and reference Signal Received Quality (RSRQ) based criteria are used to select the suitable RAT. It has been also validated in the paper that the RSRP-based RAT selection provides better user equipment (UE) throughput performance than other methods used for simulation comparison. Sharing capacities between various RAT is attractive method of balancing traffic. The LTE-WLAN aggregation (LWA, LWIP) combines the resources of both RAT’s and provide opportunity of sharing both LTE and Wi-Fi capacities [223]. The LWA provides a mechanism which enables scheduling decisions to enable

aggregated LTE and WLAN traffic at the gNB level. In particular LWA provides capacity to offload non-priority data over WiFi. Interestingly no interaction with the core network is required according to 3GPP specifications. The full control over traffic switching decision is made at the AP. In this mechanism, it is very important that the network should intelligently steer the traffic between two different RATs to satisfy the QoE of the users (or other goals of operators). In a series of works [224], [225] Afaqui et al. showed the design mechanisms and implementation of these (LWA and LWIP) mechanisms through National Instrument software-defined radio (SDR) setup at the Online Wireless Laboratory of Technical University Dresden (TUD). In multi-RAT environments, cell association, radio resource scheduling mechanism, mode selection, and multi-connectivity scheduling solutions have already been investigated to evaluate performance of the networks [226], [227] and (Anany et al., 2019). According to materials in the prior art the most important for assuring this mechanism efficiency is the level of difference between the delays on WiFi and LTE. When too large, the resulting throughput will be degraded as compared to using any of them on a single interface. This way the powerful traffic steering mechanism is required in order to assure that most suitable WiFi AP is selected.

2.9 THE ROLE OF KPIS

The user perceived QoS (or Quality of Experience – QoE) is often considered as the “ultimate measure” of system performance. According to ITU-T one can describe QoS as the ‘degree of objective service performance’ and QoE as the ‘overall acceptability of an application or service, as perceived subjectively by the end user’ [229]. While QoS evaluation is only a matter of measuring crucial network performance parameters, QoE measurements are much more complicated as they usually involve modelling the human perception in the measurement process (in a direct or indirect manner). The user-centric QoE measurement process has been already conducted by ITU-T and captured in Recommendation P.800 [230]. There is shortage of research using QoE metrics which are tailored to the surveillance applications. Most of the existing video QoE metrics are derived from VoD like technologies [231] and thus mainly related to entertainment multimedia. Recently the ITU-T is keeping track of this need in amendments to its P.912 specification for “video recognition tasks”. The methods used so far in improving QoE of video streaming focus mostly on the pixel-level QoE evaluation and can be divided into: Full-Reference, Reduced-Reference, No-Reference categories. These categories differ in the availability of source videos during metric calculation. Major focus (for multimedia scenarios) is put into enhancing and improving image quality frame by frame by evaluating specific, video-oriented metrics such as blurriness or jerkiness. However such approach does not take into consideration contextual requirements of the operators nor variations caused by variable network conditions [232]. From this thesis point of view, most important QoE metrics are: Blockiness, Blockloss and

Freezing. They are important as they directly relate to the critical aspects of video fluency when experiencing wireless network misbehaviour especially in the scenarios with the uplink video delivery. The most relevant No-Reference Video Quality Indicators cover among others: (i) PSNR (Peak Signal to Noise Ratio) – that computes the Mean Square Error (MSE) of each pixel between the original and received images. Images with more similarity will result in higher PSNR values, (ii) SSIM (Structural Similarity) – the main drawback of PSNR is that it does not consider how human perception works, hence in some cases it cannot detect some human perceptible video disruptions. To address this shortcoming, SSIM combines luminance, contrast, and structural similarity of the images to compare the correlation between the original image and the received one and eventually. Regarding the no-reference metric, the handful tool is the set of metrics developed by [233]. The metrics are very clearly defined and also they closely follow the human perception.

2.10 REAL-TIME VIDEO DELIVERY

Video streams are dominating today's Internet traffic share especially in the down-link direction. There is growing need for remote mobile surveillance solutions (area reconnaissance, crisis management operations, area crowd mapping, threat detection and mitigation for trucks on parking lots) and the emerging and disruptive market of autonomous cars strives to capitalize on that feature as well. Remote operation of cars and especially trucks is the big promise for the logistic sector [234].

In order to deal with remote monitoring of the moving objects (e.g. car, drone) environment and provide capabilities for its control by a remote operator one need to consider few crucial substrates: (i) particular scenario requirements in order to assure smooth operation of the car (e.g. being able to follow the view of the street and the guiding lines, detection of obstacles), (ii) network coverage variability (e.g. holes in the coverage) and (iii) channel dynamics while a car is moving (e.g. slow/fast-fading mitigation). On the other hand, it is important to understand that the typical surveillance systems are usually identified as “target recognition videos”. It means that the crucial performance indication is whether an operator can properly detect (recognize) an object or a situation which a camera captures. In the market there are solutions that can be used for real-time monitoring of mobile assets (e.g. police cars, busses). For example, the multipath video streaming solutions are offered by companies like [235] or [236] dealing with security monitoring. Still those solutions do not seem to assume any means of context-awareness in their architecture. In order to build robust video controllers that support teleoperation, it is necessary to deliver relevant emulation of the target network scenarios, so that the controllers can be tuned based on realistic settings of the target environment. Several standalone network emulation tools available on the market e.g. [237]. For the purpose of this thesis dedicated examination of IXIA ANUI [238] which simulates radio conditions and

impairments including signal delays, jitter or packet drops in WAN networks was performed. Unfortunately, tests performed proved it does not support dynamically changing bit rate values (on the order of hundreds of changes per second), which is main disadvantage of the solution given this thesis aims. From the perspective of the video adaptation for wireless systems authors in [239] address approaches to improve the delivery of data such as video over disadvantaged networks. This work focuses on utilizing reliable multicast protocol's hooks i.e. in the NORM protocol to provide a network information service with access to path bandwidth, delay, and lost packets. Resulting network characteristics are derived to drive the video transcoder at the server which chooses from set of profiles that include settings for: video resolution, framerate, and encoding bit-rate to allow the server attempt to fit the video stream into the available bandwidth. The work of [240] relies on context-aware services to adapt system behaviours based on the retrieved context data (context is represented in a way of ontology). Besides choosing appropriate video content based on user profile the dynamic media adaption is performed to improve the video quality perceived by the end user in response to changes in: varying wireless channel quality, available energy of the end equipment, network congestion and application Quality of Services (QoS). The results show that utilizing context may help improving video quality (PSNR) by 2–3 dB. On the opposite [241] suggests that in order to adapt videos sent from UAVs “instead of reacting to packet loss, he uses an increase in queueing delay at the router [or CPE] to detect phases of throughput degradation”. Still few authors focus on the QoE adaptation concerning the needs of the remote monitoring and operation of car or drone.

2.11 OVERVIEW OF SELECTED EU RESEARCH PROJECTS

In the Table 11 an overview of research projects selected due to their convergence with the subject of this dissertation and implemented under the programs m.in the European Commission, the European Defence Agency or Eureka/Celtic in recent years was presented.

Table 11 Review of selected R&D projects related to the dissertation [own study]

Sponsoring Institution / Programme		
European Commission: FP6, FP7, H2020	EDA: Cat. „ad-hoc B”	NCBIR: Eureka/Celtic

<ul style="list-style-type: none"> • FP6 EuQoS (2005-2008) – development of a quality management system in heterogeneous IP networks • F7 DaVinci (2008-2010) – development of new LDPC codes for 4G wireless networks • H2020 5G Essence (2017-2020) – two-tier virtualized 5G network architecture, including RRM elements • SNS JU, 6G-SANDBOX (2023-2025) – development of a testbed for research on 6G, especially in the field of ORAN, 6G, programmable networks, workload prediction and placement 	<p>TACTICS (2014-2017) – development of an intelligent TSI control layer enabling data communication in UHF/VHF tactical networks in SOA architecture</p>	<ul style="list-style-type: none"> • MITSU (2013-2016) – development of new solutions for the implementation of scalable video streaming services in LTE and WIMAX networks • BRAINE (ECSEL JU) (2020-2023) – design and development of the EMDC architecture for edge computing combined with AI/ML subsystems, as well as SOA-based service oriented architecture; elements of virtualization and disaggregation of 5G RAN networks
<ul style="list-style-type: none"> • FP7 Sokrates (2008-2010) – autonomous management and autonomous configuration in wireless networks • FP7 SemaFour (2012-2015) - autonomous management for heterogeneous wireless RAN access networks • FP7 T-NOVA (2014-2016) – implementation of NFV network functions in the service paradigm using virtual network infrastructure • H2020 Superfluidity (2015-2017) – flexible system for the implementation of NFV function management services, in the infrastructure based on the use of cloud computing 		

In the EuQoS project [63] the concept of quality control in heterogeneous networks is based on the resource broker mechanism. In the architectural layer, central resource control was implemented through the resource broker component. The approach also resembles the concept of "control by probing the network" [78]. The project did not address 4G networks, while the approach to the implementation of CAC functions in UMTS networks was based on the use of MDP algorithms. Delivering efficient FEC codes through implementation of non-binary LDPB (nb-LDPC) codes for 4G networks was the main focus of FP7 DaVinci [242]. One of the perspectives of assessing the effectiveness of the system *level simulation* using new codes included verification of the effectiveness of a radio system (WiMAX, LTE) equipped with nb-LDPC codes in connection with the use of CAC algorithms. The approach used in the Sokrates project is very similar to the idea of a dynamically defined guard-band resource threshold, determining the division between resources reserved for *handover* connections and resources for handling new requests presented in [153]. The project analyzed the aspect of CAC algorithms in self-organizing LTE networks, with particular

focus on situations of sudden change in traffic conditions in the mobile network, with time-varying capacity for connections in the downlink direction. Estimation of variable cell capacity is carried out on the basis of work [148]. The proposed algorithm assumes that a sufficiently high number of flexible connections (nRT) is available in the system in relation to real-time connections (RT), in order to ensure flexibility of resource control in the event of temporary overload. The approach proposed in the project, assumes the possibility of degradation of flexible connections (nRT) in order to free up sufficient resources for priority connections. It is similar to the work [162] [243]. In the TACTICS project [244], special emphasis was placed on distributed management of radio resources in the environment of a mobile tactical network. The control of the quality assurance (QoS) subsystem is carried out in conjunction with the objectives of the security subsystem, the current decisions of which, must be taken into account by the QoS controller [245]. A summary of the project's work, along with a demonstration of the performance results of QoS control solutions, is comprehensively presented in the work [246]. The main objective of the MITSU project was to create new solutions for video streaming systems in 4G wireless networks. The main research and implementation works on 4G dealt with development of new methods for detecting and controlling overloads in the 4G environment as well as analysis of the role of protocols like multi-path in optimizing the process of delivering content to the operations center [247]. The resource controller in the transmitting node (on the 4G user terminal side) has been designed to adjust the effective bitrate of the video stream in the uplink direction to the actual link capabilities, based on the measurement of radio channel parameters (CINR, RSSI, upward modulation). The design work there also analyzed the algorithms for controlling the acceptance of applications used in the FP7 DaVinci project. The problem of service allocation to a data-center addressed by the T-NOVA project focuses on the optimal allocation of services to networked data centers. These centers are managed by a single operator [248]. The concept of optimality should be understood as multi-criteria optimization in terms of m.in: economic benefits, QoS, energy efficiency, etc. This type of problem should be treated as an online problem. There is a lack of a priori knowledge about the distribution of *Network Service* requests, requests arrive in a dynamic manner, and after acceptance they can remain in the system for any length of time. Therefore, "service allocation" problems must be addressed on an ongoing basis as applications arrive. The first problem is the selection and allocation of appropriate virtual network (VNF) components to data centers when the orchestrator is requested. To provide a scalable and resilient (robust) solution, new resource allocation algorithms are required to address allocation issues for geographically dispersed data centers, partners use the MDP approach. The benefits of using the MDP approach include among all, the ability to predict future requests (their dynamics) and apply offline learning and then use the optimal strategy for incoming requests in the actual system. The aim of the Superfluidity project was to provide solutions at the network level (infrastructure) and controls that will

allow to provide services "on demand", in any network segment (backbone, aggregation, access) with the possibility of moving the place of service in a flexible manner and not adversely affecting the quality of the service, from the perspective of users using it [249]. The main challenges faced by the project are, among others: enhanced service provisioning periods, over dimensioning to cope with VBR traffic, diverse traffic sources, services, access technologies using multi-vendor devices and diverse end-user needs. The aim of the project was to provide converged services embedded in cloud computing that harmonize with the currently designed architectures of future 5G networks, enabling increasing the role of mobile Edge computing, offering new business models and reducing investment and maintenance costs (CAPEX, OPEX). The work deals with the problem of optimal allocation of network resources (virtualized) necessary to provide services. Of particular importance is the decision to scale up/down the compute resources necessary to start (stop) the virtual machines (VMs) allocated to execute VNFs. A solution was sought that optimizes the level of resources used while meeting the quality requirements specified in the SLA contract. The related task is to ensure optimal load balancing between VMs. The problem was formulated as an MDP problem. The policy controlling the assignment of a task to the queue is activated with each new incoming request and the completion of the handling of an ongoing request. The set of possible controls is described by a vector where the permissible values of the vector include the following actions: „*allocate a new resource*“, "*do nothing*", "*delete machine*". The problem presented in the project concerns the control of task assignment to servers, and therefore in general concerns the problem also raised by the authors in [250].

2.12 IDENTIFICATION OF GUIDELINES FOR OWN RESEARCH TASKS

On the basis of the analysis of the state of knowledge in the field of admission and congestion control, the author concludes that there is a need to conduct own research – responding to the research problems stated and also the research hypothesis identified in chapter 1.2.

Based on the various SOTA items presented in the chapter above, the author would like now to restate key aspects and directions which present high potential solution for the research questions and the research hypothesis presented earlier:

- There are multiple proposals for admission control in the existing prior art but author recognizes that admission control function needs to engage multiple dimensions and features already present and foreseen in the future networks, especially considering :
 - relevant, **customizable and affordable models for evaluating service’s quality**, where the field measurements can be directly introduced in for of models, to offer high reliability testing of new algorithms and solutions (restoring the statistical behaviour of loss, delays and throughputs). As currently its availability is

mainly limited to expensive commercial tools (e.g. Riverbed Modeller) or testbeds which are yet in early development stage and necessitates expanding its availability (e.g. 6G-SANDBOX, 6G-BRICKS) for wider audience

- the **controls for multi-RAT optimization**, that enables exploiting of multiple spectrum bands at same time (e.g. 5G, 4G, WiFi) but also is an efficient way to gradually evolve new network clusters that can be controlled by appropriate admission and congestion control algorithms
- it is worth **validating to what extent existing CAC algorithms that seem suitable to be used in 5G as well as B5G can be utilized** in the emerging networks with necessary modernizations to assure quality improvements for operators and end-users
- to assure keeping pace towards future developments of network controls towards ubiquitous networks, there needs to be **systematic approach for learning-driven and evolving admission control algorithms** (definitions), so that they can be adapted to the new objectives or network capabilities (e.g. semantic enrichment, edge-cloud deployment)
- per-service (e.g. VoIP, IPTV, real-time video), high fidelity and reliable **QoE metrics need to be pluggable into decision making** (e.g. admission control, traffic steering, congestion control) to properly address the KPIs required by the quality targeted by the 5G/6G networks
- The future wireless networks will heavily rely on the SDN based controllers that are offering **high degree of programmability of the control functions**. Here the ORAN driven RAN intelligent controller with xApps and E2 interface is the currently a promising solution that is accepted by the community and represent the potential to also be the optimal solution for currently designed B5G networks. Such controllers are important for these networks but also to coexist with other networks e.g. optical transport networks in the edge.
- **New control actions for admission control are needed** to better capture capabilities and enablers emerging in the novel RAN networks (AI/ML, data-driven controls, cell-free, etc). Considering the growing role of the virtualized, cooperation oriented and disaggregated RANs such actions would target e.g. : i) *auto-scaling of vRAN disaggregated functions* towards federated edge-server or multi-cloud deployment, ii) *activating alternative multi-access scheme e.g. NOMA/RSMA* for selected cluster of users, iii) *offload user-traffic to cell-free or WiFi*, iv) manage energy cost by cooperating with intelligent RU activation schemes in cell-free or cellular networks.

According to the authors in [251], "[...] in recent years there has been a growing

awareness of the need to test and measure new mobile applications and protocols in realistic wireless network configurations, and (6G SandBox, 2023) the [252], [24], [253], [254] network labs have been developed to meet this demand". Previously in Europe, network labs have been federated under the umbrella of the Fed4FIRE projects (nowadays this project concept is continued under slicessc.eu project), this is a pan-European initiative and also brings together 4G and 5G compliant labs (e.g. the NITOS lab in Greece). For example, the paper focuses on measuring the performance of several popular wireless applications in the GENI environment under test. The summary presented in the paper shows rather surprising results obtained using a real network (within the GENI project). For improving signal quality (RSSI), a reduction in application throughput is observed. The authors emphasise that the results obtained using a realistic test configuration may differ significantly from what is expected and may even appear to be inconsistent with what is intuitively expected. For this reason, it is crucial that *laboratories of this type make available reference baseline measurement results*, thus enabling the correct analysis of results obtained in more complex scenarios to be carried out reliably. The author of this dissertation confirms the important role of field measurements, as he also experienced anomalous' results in his study, as E2E delays in the 'upstream' direction decreased with increasing connection bandwidth [42]. When designing various CAC algorithms, several important aspects must be considered:

- Over time, the trend of installing small (micro, pico) cells will become more and more visible, and therefore the number of handover events during the call will increase and will increase; The influence of movement from neighbouring cells is also increasingly predicted [255]
- Due to the differences in the quality of connections of different classes, and the increase in the expected quality of individual connections, it will become necessary to control the allocation of resources more precisely [256]
- The variation in network load on a short- and long-term scale means that CAC mechanisms should adapt their settings to changing movement patterns.

To deal with the key topics identified as missing or not yet properly addressed in the literature author will now present the methodological assumptions taken to maximize research activities outcomes, while assuring its high quality and ultimate relevance towards answering research questions stated in the preliminary part of the previous chapter. By following such approach, author assumes he will be able to validate the main hypothesis of this work as presented earlier.

2.13 ASSUMPTIONS MADE TO ASSURE HIGH QUALITY RESEARCH

The assumptions for the delivery of high quality research outputs which have been selected as most relevant for achieving this dissertation objectives, are

summarized below:

- **A1 (performance evaluation):** it is needed to identify existing performances of the 4G/5G networks, considering different traffic sources, in order to be able to capture appropriate data set for the good understanding of the baseline systems already in the market.
- **A2 (experimentation environment):** it is needed to identify, select and tailor, or define from scratch the relevant simulation and testing environment, that will enable development and evaluation of algorithms and schemes for admission control, suitable for the current and future wireless network generations. At minimum the system level test environment will be required that offers system settings in wide range of parameters that are relevant for the thesis. However, it will be important to also complement simulations and utilize real radio environment considering selected radio interfaces for 4G, WiFi and 5G.
- **A3 (CAC algorithms):** development of novel algorithms, or necessary modernizations, based on the findings from the above-mentioned desk research will be made considering as important criterion its practical applicability into the existing disaggregated and virtualized 4G/5G RAN networks (SW-driven networks), with the potential to apply these solutions to 6G as well
- **A4 (reference scheduler):** considering the main hypothesis underlying this thesis, author decides he will be dealing with reference scheduler algorithms for wireless systems, as addressing the specific novel schedulers like cell-free outside of the scope required to validate the hypotheses. Although author already has contributed to the definition and design of novel cell-free scheduler designs in the past
- **A5 (solid methodology):** a solid and future proof methodological framework that enables introducing new algorithms driven by evolutionary and data-driven concepts, for its application into a virtualized and disaggregated wireless networks provisioning (including admission and congestion control) is essential for assuring scalable methodological designs
- **A6 (intra-slice focus):** it is evident to author that inter-slice admission control can be within the scope of future oriented admission control solutions, however the aim of this work is to focus on intra-slice admission aspects and cover most of this novel extensions, while leaving wireless network provisioning for inter-slice a topic for future research as the targeted number of research questions is relatively high and focuses on the single slice to better focus the research.

3 OWN RESEARCH METHODOLOGY

3.1 INTRODUCTION

In this chapter author introduces the approach to perform research in the next chapters based on the solid foundations. The research is driven by the research questions from chapter 1 and in order to proof the thesis of the work. In the reminder sections author first introduces and analyses most relevant ways of maximization of quality of network performance.

3.2 QUALITY FUNCTION MAXIMIZATION – PRELIMINARY CONSIDERATIONS

The analysis of possible forms of the quality function and problems related to its minimization is conditioned by the use of an appropriate mathematical formalism. Such a formalism may vary and depends on the features of the system to be taken into account. To study the congestion of specific states and transformations of transmission means, depending on the purpose of the research, models reduced to a form useful for specific considerations or calculations should be used, i.e. in the case considered here, to a normalized form and to quality equations (for numerical calculations) [Chodkowski]. When discussing the problems of identification with the use of the quality function, a very general mathematical formalism can be used, assuming that the properties of the technical system are described by a non-linear operator realizing the transformation of a set of input signals (excitations) into a set of output signals (responses) $y = [y_1 \dots y_n]$. The properties of the operator F depend on e.g. from the set of parameters, which can be symbolically written in the form $F = F(A)$. The following form ((3-1) of the operator was adopted for further considerations F , such that:

$$y = F(A) - q \quad (3-1)$$

moreover and it was assumed that the same excitation signals act at the input of the mathematical model and the technical means q . The mathematical model is always an approximate description of the construction of the technical means of transmission, and therefore, in accordance with the assumptions made, a set of signals is obtained at the output of the system $y = [y_1 \dots, y_n]$, different from the set of signals $y^* = [y^*_1 \dots y^*_n]$ at the output of the mathematical model. The value of this difference can be used to assess the quality of the mathematical model. Assuming sets of $y^* \in \text{DCR}^n$ and $y \in \text{DCR}^n$, the quality of a mathematical model can be assessed using a certain mathematical norm $\|y - y^*\|$ expressing the distance between sets y and y^* in the domain D . The type of mathematical operations and the choice of a specific form of the norm $\| \cdot \|$ can be described by an

expression called a quality function (objective function or quality criterion). Functions can be different, e.g.

$$Q^{(1)} = \|y - y^*\|^p \quad (3-2)$$

usually, the value of $p = 2$.

$$Q^{(2)} = \sum_{i=1}^n \|y_i - y_{y^*}^*\| \quad (3-3)$$

$$Q^{(3)} = \max_i \|y_i - y_{y^*}^*\| \quad (3-4)$$

If the set of signals at the output is observed in discrete time instants, i.e. discrete sets of values of each element of the sets y and y^* , than denoting consecutive observations with the index j ($j = 1, \dots, N$), one can get the matrix $n \times N$ with discrete values $y_{i^*}^{(j)}$ and $y_i^{(j)}$ of elements of the sets y^* and y . In this case, the quality functions ((3-2), ((3-3) and ((3-4) are defined by relationships:

$$Q^{(1)} = \sum_{j=1}^N \|y^{(j)} - y^{*(j)}\| \quad (3-5)$$

$$Q^{(2)} = \sum_{j=1}^N \sum_{i=1}^n \|y_i^{(j)} - y_{y^*}^{*(j)}\| \quad (3-6)$$

$$Q^{(3)} = \sum_{j=1}^N \max_i \|y_i^{(j)} - y_{y^*}^{*(j)}\| \quad (3-7)$$

The introduction of the quality function not only facilitates the assessment of the adequacy of the mathematical model, but also enables the implementation of the process of developing the control of accepting applications, the best solution in a specific sense. Optimization may concern the development of both the model structure and its parameters. Then the problem of finding the optimal model can be reduced to determining the optimal set of parameters $A = A^*$; i.e. a set that minimizes the distance between sets y and y^* . This consists in minimizing the quality function Q , written as:

$$Q(A^*) = Q(A) \quad (3-8)$$

Taking into account the equation ((3-1), the relation ((3-2), with $p=2$, can be expressed as:

$$Q = \|y - F(A) \cdot q\|^2 = Q(A) \quad (3-9)$$

i.e. present the quality function in a form that explicitly depends on the parameters A . If the output signals are random functions, then the quality function is usually assumed in the form:

$$Q = \iint_{px} \|y - F(A) \cdot q\|^2 f(q, y) \cdot dq \cdot dy \quad (3-10)$$

where:

$q \in P$ and $y \in X$, where $f(q, y)$ — probability density function of random signals q (on the inputs) and y (at output) of a transmission system. According to relation (3-8) searching for a set of optimal parameter values A^* comes down to minimizing the quality function. Parameters A^* of the quality function depend on the probability distribution function $f(q, y)$. Most often, the effective result can be obtained only if the Euclidean norm (least squares criterion) is adopted. It is then obtained:

$$Q = \iint_{px} [y - F(A) \cdot q]^T [y - F(A) \cdot q] f(q, y) \cdot dq \cdot dy \quad (3-11)$$

where: $q \in P$ and $y \in X$. The next important step in the identification process is the search for the extremum (minimum) of the quality function. The identification algorithm can be determined by specifying the method of solving the system of equations

$$\text{grad } Q(A) = 0 \quad \text{gd } A = A^* \quad (3-12)$$

if there are no constraints on the choice of A and it is known that $Q(A)$ is a differentiable function unimodal, i.e. having only one extremum, with equality at only one point (3-11).

The aspect of selecting the appropriate method of searching for the extremum from the known basic methods, requires careful consideration. The problem of searching for the extremum of the quality function $Q(A)$ it includes the search for the extremum (minimum) of a function of one variable and the search for the extremum of a function of many variables, where the function may be unimodal (having one extreme) or it may be a function with many extremes. The quality function $Q(A)$ used to identify the parameters is usually a function of many variables, because $A = [a_1, \dots, a_r]$.

The quality function of single variable is hereafter denoted as $Q(a)$. To search for the minimum of the unimodal quality function of one variable, the following are most often used [257]:

- Fibonacci method (number sequence),
- golden ratio (Euclidean) method,

- dichotomy method.

The methods listed above enable the use of an optimal, in a specific sense, strategy of narrowing the interval in which the minimum of the quality function $Q(a)$ is located. Each of these methods makes it possible to determine the point dividing the segment along the variable a into two optimal parts. The most common random search methods are:

- stochastic approximation method,
- maximum likelihood method,
- the method of maximum likelihood a posteriori (in other words, the method with a posteriori density of measurements),
- statistical decision method (or minimum risk method).

The stochastic approximation method (for example Q-learning algorithm) is most often used to identify dynamic processes. The form of the regression equation is not arbitrary. There should be e.g. met the condition of linearity with respect to the regression parameters. In the general case, the regression equation may be a non-linear algebraic equation, and in the present case, the linearity of such an equation with respect to the parameters means that the parameters cannot be expressed as exponential terms.

The condition of the linearity of the regression equation with respect to the parameters is related to the process of minimizing the quality function Q . Using the regression analysis method, the quality function can be presented as the relationship (3-11) when the input and output signals of the system are random functions. These aspects will be dealt with in the Chapter 7.

Determination of the quality function $Q=Q(A)$ and its minimization encounter great computational difficulties. Such difficulties are greatly reduced when the regression equation can be represented in a linear form with respect to the parameters. Then, determining the optimal set of parameters A^* is relatively easy, because, for example, in the case of the quality function determined by the relation (3-12) – ((3-11), it is not necessary to know the probability distribution function $f(q, y)$, but it is enough to know the appropriate moments of this distribution [258]. Therefore, when using the regression analysis method, the following regression equation is most often adopted:

$$y = B^T \cdot \varphi(\underline{q}) \quad (3-13)$$

where:

$\underline{q}(q_1, \dots, q_s)$ - set of input signals,

y – output signal,

$\varphi(q) = [\varphi_0(q), \varphi_1(q) \dots \varphi_k(q)]$ - arbitrary function of signals \underline{q} ,

where $\varphi_0(\underline{q}) = 1$ and $B = (b_0, b_1, \dots, b_k)$ — regression parameters.

Thus, the third condition for the applicability of the regression analysis method is to transform the differential equations in such a way that the obtained algebraic equations have the form of dependence (3-13).

One of the important steps in the process of identifying a mathematical model of a non-linear system is replacing non-linear functions with such linearized functions that are statistically equivalent to given non-linear functions. Such linearization is enabled, for example, by the Kozakov-Booton statistical linearization method [257], which from the point of view of engineering applications is one of the most important in the group of correlation methods.

The condition for statistical equivalence is the equivalence of the first and second order statistical moments. I.E. Kozakov adopted two criteria of statistical equivalence, namely the so-called:

- first criterion: equality of expected values and variances of a non-linear function and a linearized function,
- second criterion: the expected value of the square of the difference between the real function and the linearized function is the smallest; it is called the criterion of the minimum mean square error, defined by the relation:

$$\varepsilon^{-2} = E[(y - y_1)^2] = \min \quad (3-14)$$

The statistical equivalence criteria given above enable the determination of the statistical linearization coefficients k_0 and k_{1s} , with $s=1$ or 2 depending on whether the first or second linearization criterion is used to determine the statistical linearization coefficient.

Complementing the previous considerations, it should be added that the conditions in which the identification process is carried out may differ significantly from the normal operating conditions of a wireless network. This mainly concerns the nature of the signals (determined or random) and the points of operation of the forcing signals. In addition, simplifications in terms of taking into account signalling traffic in the network, in particular aspects related to mobility and its impact on additional signalling. In such cases, the mathematical model is identified in conditions that differ from the actual conditions of the wireless network operation, which may result in omitting important characteristics of the dynamics of a real system equipped with a radio interface and additional mechanisms, e.g. HARQ. If, for example, the ranges of real inputs are much wider than the ranges of inputs implemented during the planned experiment, some non-linearities relevant to the operation of the system may remain undetected. The omission of certain qualitative features of excitation signals may distort the real dynamic properties of the model. Taking this into account, during the experiment, the dynamic properties of the system should be carefully examined, which, as a result of the identification, should correspond to the mathematical model of the wireless network.

3.3 METHODOLOGY OF OWN RESEARCH

Analysis, research and intelligent development as well as assessment of the

quality of control algorithms for accepting and realizing requests (congestion) in future wireless networks require more precise assumptions. The universal importance of the criteria, research methodology and development of the transmission system in wireless networks for the implementation of fragmentary goals was assumed.

The transmission system is a set of technical conditions W_j , relevant elements E_1, E_2, \dots, E_m , relations between these elements R_1, R_2, \dots, R_n , which are functions of Θ and t (i.e. operating time and the process dynamics), which are aimed at the implementation of a set of $SP_i = H$ characteristics (3-15) as output quantities on which the assessment of the overall quality, efficiency, harmlessness of data transmission depends, e.g. uniformity, efficiency, high efficiency, reasonable power consumption and nature, specific energy consumption, etc.

$$SP_i = f(Wt_j) \quad (3-15)$$

where:

SP_i - i -th postulated state of transmission in wireless networks,

Wt_j - j -th technical conditions described by the solution concept for the postulated state,

Technical conditions of the concept, architecture and parameters of the 4G/5G/6G network (Wt_j), will be considered by using weights a , b and c , for: algorithms, as ($a \cdot A$), delays ($b \cdot O$), inaccuracies ($c \cdot N$):

$$SP_i = a \cdot A + b \cdot O + c \cdot N \quad (3-16)$$

where:

A - control algorithms with the use of the considered concept of solving the i -th state of the postulated transmission quality in wireless networks (QbP),

O - delays in the share of the i -th state of the postulated transmission quality in wireless networks (QbP),

N - inaccuracy of elements and realizations in the share of the i -th state of the postulated transmission quality in wireless networks (QbP),

a -, b -, c - percentage share of the technical conditions (0-100)% described in the solution concept for the postulated state of transmission quality in wireless networks (QbP).

Four (operational) subsystems (constructs, abstracts) can be distinguished in the modelled system, performing strict and precise functions [259]:

1. **SP process subsystem**, phenomenal functioning, i.e. implementation by the network (system, supersystem) of the set goal (accepting requests, data transmission, resolving congestion, ...);

2. **SC control subsystem**, high-efficiency management and coordination of the construction of subsystems (process, information and logistics) - for low-energy implementation of the objectives of elements, relations and integrated controls of transmission quality;
3. **SJ information subsystem**, self-organization, processing and distribution of information streams within the system (supersystem), from outside the system in accordance with the needs and objectives of operation of other special structures (subsystems);
4. **KL logistics subsystem**, reliable maintenance and supply (deployment) of other subsystems (e.g. MEC, LBO), special systems (and itself), for the reliable delivery of numerous operational goals.

The above subsystems should always be considered in totality and cross-connected when analysing the system, as they co-define the system dimensions comprehensively. The method of analysis, assessment and indication of directions for sustainable development of the quality of the transmission system in wireless networks (QbP) was based on the existing knowledge about the use of the potential of technology, environment, modern technologies, machines and production processes as well as transmission design features, constituted in the design, control, processing, exploitation and post-consumer development, and additionally on the standards of algorithms supported by control signals $\bar{S}(\theta, t)$:

$$\begin{aligned}
 E_1 &= E_1(\bar{W}, \theta, t) \\
 E_2 &= E_2(\bar{W}, \theta, t) \\
 &\dots\dots\dots \\
 E_m &= E_m(\bar{W}, \theta, t) \\
 R_1 &= R_1(W, s, \theta, t) \\
 R_2 &= R_2(W, s, \theta, t) \\
 &\dots\dots\dots \\
 R_n &= R_n(W, s, \theta, t)
 \end{aligned}
 \tag{3-17}$$

With such assumptions, the general model of developmental transformation ((3-18) in the integrated telecommunications environment, allowing for analysis, creative conception, research and evaluation, is as follows:

$$L(H, E, R, \theta, t) = P(s, z, \theta, t - t_o) \tag{3-18}$$

where:

L, P – represents the left and right side of the equation, where the $L(*)$ represents the characteristics of the system, while the $P(*)$ is set of controls which are acting upon the identified system in order to assure meeting quality goals.

\bar{H} - performance characteristics as output values, e.g. the quality of the transmission system in wireless networks (QbP/GoS), (power, energy, emissions, waste), economic, ecological and energy efficiency of processes, harmlessness of the impact of products and transmission processes), harmfulness

to human life and health , e.g. phases, stages of the life cycle of transmission facilities,

\bar{E} - elements of transformation of technical conditions of transmission in wireless networks (QbP), power and environment (carriers, transmission techniques, transmission networks, devices, installations, waste, emissions)

\bar{R} - relationships, connection of transmission system elements (acceptance of orders, movement, transformation, transmission, processing, storage, storage, consumption, development, accumulation, service, etc.), creative impact on transmission objects (optimization, modernization, innovation; use, servicing, repairs, supply, disposal - scrapping; recycling of antenna materials, ...) - R – represents e.g. relations between when e.g. clustering them into cluster served by 4G and another one served by 5G or WiFi. These relationships can be subject to timely evolution.

\bar{W} – represents interference between different users

Θ - dynamics of the analysed system

$t - t_0$ - time,

\bar{s} - variable representing controlling, regulating, compensating, monitoring, creating, transformative destruction,

\bar{z} - technological, social, environmental, accidental and other disturbances.

The left (L) side of equation ((3-18) (model) describes the properties of the feedstock, product, transmission process, its physical characteristics, appropriate for a given class of activities. These properties depend on the characteristics of the elements $E1, E2, \dots, Em$, connections between these elements $R1, R2, \dots, Rn$, and are functions of the Θ and t (which represent the dynamics of the process and time respectively). The unknowns are the elements of the set of characteristics H as output quantities, on which the specific assessment of quality depends, and the overall efficiency, harmlessness - inhomogeneity, ineffectiveness, variable efficiency, unreasonable power consumption and nature, specific energy consumption, etc. Mathematical model of the transmission system in wireless networks proposed for this thesis is:

$$H_u(Q_{prod.}, e_{proc.}, N_{prod.}, N_{proc.}, t_0 - t_1) = f(ZST) \quad (3-19)$$

where:

H_u – operating characteristics change as a function of transmission system variables (ZST)

$Q_{prod.}$ – product quality of the transmission system in wireless networks (QbP),

$e_{proc.}$ – efficiency of the transmission system process in wireless networks,

$N_{prod.}$ – harmlessness of the transmission system product in wireless networks,

$N_{proc.}$ – harmlessness of the transmission process in wireless networks,

$t_0 - t_1$ - transmission execution time.

In order to represent the above-mentioned model and its variables the figure Figure 17 provides visual clue of the input-output relations in the modelled wireless system. Based on the above-mentioned variables defining identified system model (4G/5G/beyond), all the elements have been collected and indicated in the Figure 17. The dashed rectangles in the middle section represent the four subsystems as introduced earlier in the section (SP, SC, SJ, KL), and the $\bar{E}, \bar{R}, \bar{W}, \bar{\theta}$ represent the system characteristics and the target quality metrics are identified as \bar{H} .

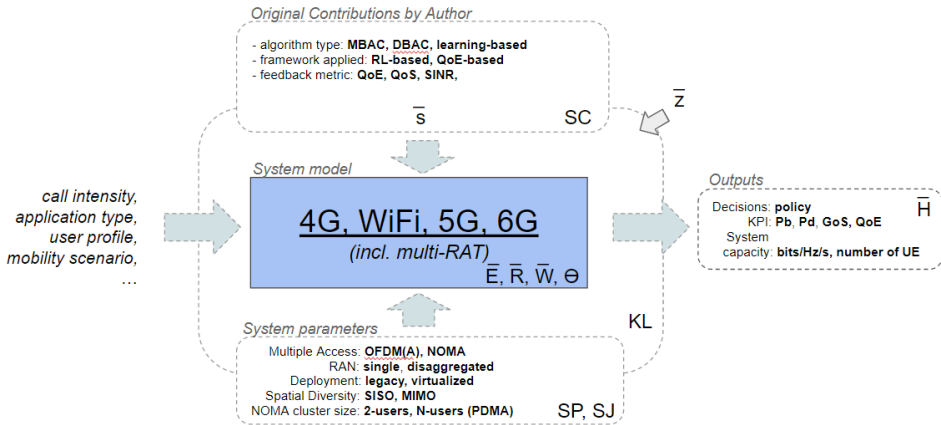


Figure 17 Representation of the modelled wireless system with key system parameters, inputs and outputs

The targeted solution of the control algorithms and solutions for 4G/5G/beyond 5G which has the characteristics of being integrated, advanced, tailored wireless data transmission system can be characterized by:

- Extremely efficient processing of data (data delivery)
- Highly efficient control plane
- Self-organizing information system (e.g. liquidity of RAN resource consumption in response to changing environment conditions)
- Highly reliable network management and operations (by e.g. means of pursuing optimal H especially when system boundary conditions change)

In this way, a drastic, astonishing improvement in the postulated states (goals) and quality is possible.

3.4 SELECTION OF RELEVANT QUALITY CHARACTERISTICS (H_U)

As already indicated in the chapter 2, the most universal quality parameters for admission (congestion) control algorithms that will in the reminder of this thesis contribute to the operating characteristics of the system, and are listed below:

- B_U – bandwidth utilization, represents how well the available BW of the system is utilized in a given unit of time
- $P_{b,k}$ – blocking probability of service class k
- $P_{d,k}$ – dropping probability of service class k
- $D_{E2E,k}$ – delay experienced by connections of class i
- n_k – number $n \in N$, $k \in K$

After [260] “desirable property of an admission control algorithm is that its admissible region be as close as possible to, but not greater than, the simulation curve. In other words, the goal is to utilize resources as highly as possible without admitting more flows than can actually be supported, which would result in violations of the promised QoS”. In order to define key parameters for the next chapters, the following equations are provided below:

$$P_{b,k} = \frac{n_k}{n_{k,rej} + n_{k,acc}} * 100\% \quad (3-20)$$

$$P_{d,k} = \frac{n_{k,drop}}{n_{k,acc}} * 100\% \quad (3-21)$$

$$B_U = \frac{S_{sched}}{S_{All}} * 100\% \quad (3-22)$$

where:

n_k – number of k-class users

$n_{k,rej}$ – number of user connections rejected in class k

$n_{k,acc}$ – number of user connections accepted in class k

S_{All} – number of available symbols (slots)

S_{sched} – number of scheduled symbols

As stated in the analysis of the literature, one of the important quality parameters that collectively combines probabilities P_b and P_d is the Grade of Service. It not only allows to control the level of probabilities in the system, but also facilitates decisions making about changes in the allowable level of these probabilities (β_k). For the purposes of the work, the author chose after [162], the following two metrics as the composite metrics that well define the quality level represented by the H_u :

$$GoS_k = P_{b,k} + \beta_k \cdot P_{d,k}, k \in \{u, r, n\} \quad (3-23)$$

$$CF = w_1 * GoS_u + w_1 * GoS_r + w_1 * GoS_n \quad (3-24)$$

The equation ((3-23) enables combined view on the probabilities determining overall quality, whereas the ((3-24) indicates the cost function is composed of the GoS values per class.

3.5 WIRELESS SYSTEM MODELLING AND EXEMPLIFICATION

International standardization organizations, responsible for preparing specifications (such as IMT-Advanced, IMT-2020) for wireless networks, define requirements for system level simulations for the candidate technologies [261], [262] [263]. The goal behind those documents is to facilitate System-Level-Simulations by providing common methodology to perform such simulations (i.e. for 4G, 5G). According to [261] **cell-level simulations** can be an intermediate step between Link and System-level simulation where the capacity of one cell and one Base Station, providing service for multiple users, is evaluated by means of comprehensive tests. Still the standardized simulation methodology [264] *does not specify how to evaluate system capacity with various connection admission control mechanisms*. Therefore as a first step we focus on the problem of adjusting simulation methodology to facilitate simulations covering CAC with TDD scheme and OFDM for uplink traffic. The applied evaluation methodology is derived from the best-practices in IEEE and 3GPP. However it is worth highlighting that it can be seen that target values introduced in the specifications (e.g. capacity traffic per area) are very challenging and based on available reports and predictions even after largely increasing the current networks' performance with the target traffic increase of x20 by 2030, the expected value is estimated to reach only the 3% of the target (e.g. 300Gb/s/km²).

3.6 LINK TO SYSTEM LEVEL MAPPING (L2S / SLS)

To improve the fidelity level of the simulator and introduce mobile channels, method called Link-To-System interface (L2S) has been implemented. This approach removes constraints that arise when AWGN channel is being used. In particular a method based on mutual information (MI) called RBIR (Mutual Information Per Received Bit | Received coded Bit Information Rate) was selected. It is important, since attempting to simulate scenarios close to reality requires combining admission control and user mobility. The mobility model used is based on traces following the Leavy-walk distribution. Users' movements have been captured for a given geographical area and combined with maps generated by the Radio Mobile radio coverage planning tool [265]. Thus author was able to present results of assessing quality of VoIP (Voice Over IP) conversations also in novel non-binary Low Density Parity-Check (nb-LPDC) coded networks. The corresponding work is described in following chapters. Finally, using L2S allows comparing SUT's performance using either nbLDPC or well-recognised CTC codes. Thus comparison of CTC and nbLDPC codes is provided in terms of resulting system capacity and quality of experience (QoE) performance of VoIP flows.

3.7 MAP GENERATION

In order to deal with evaluation of the performance and system capacity at the cell-level with ACM enabled the following approach based on standard procedures will be followed. In the reminder of this thesis (chapters 4-8) in order to properly evaluate the proposed algorithms and solutions for future wireless systems, the SNR values should not be constant and change in time. In order to evaluate behaviour of 4G/5G system in various propagation conditions we have decided to prepare two distinct geographical maps (with respect to the SNR distribution). Author generated two SNR matrices for rural and hilly terrains limited to 16 square km area. Mobility models follow Levy-walk distribution and are generated using Matlab source files provided by [266]. Both user mobility patterns and SNR maps are combined in order to generate modulation transition trace files functioning as lookup tables for NS2 simulation (Figure 18). The two maps were generated for base station with omni directional antenna, transmit power 42 dBm and of BS antenna's height of 35m.

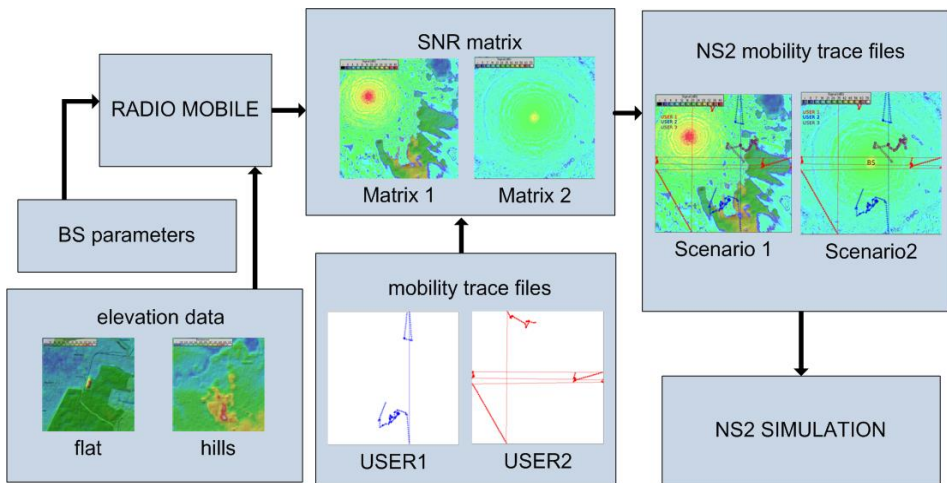


Figure 18 Map generation using Mobile Radio and levy walk model

In order to generate map of SNR values the Radio Mobile [265] was used. It is a freeware software, which main functionality is generating radio coverage areas for a given geographical region and elevation data. In simulations the SRTM project [267] elevation data is used, as it provides the best elevation resolution for the Europe region. In order to generate realistic mobility schemes of simulated users the human mobility trace data measured in New York and Disneyworld were adopted [266]. User movements are facilitated within simulation in the form of lookup tables (imported to a simulator). In this scenario several users are moving within the covered area. It is assumed that each user's channel conditions are changing in time and ACM is enabled. In order to represent characteristics of

different wireless systems the tests will be performed for two coding schemes – Reed-Solomon Convolutional Coding (RS-CC) and the nbLDPC. At each BS, a resource-fair scheduler is assumed. This can be accomplished by the round robin (RR) scheduler which translates to the PS discipline in queuing terminology. To avoid congestion, each BS applies admission control after [268].

3.8 MEASUREMENT TOOLS VALIDATION

This section focuses on the preparatory stage before measurement campaign of video streaming services in 4G/5G networks. Based on the instrumentation designed, developed and validated by author it was feasible to perform field tests that were used to develop models in the chapter 5. Author has successfully prepared a complete environment that relies on low-cost IP performance measurement software. It has been shown that the software provides accurate measurements when validated using professional packet generator/analyser IXIA XM2. It has been also shown that it is quite straightforward and cost effective to realise GPS synchronisation using the self-developed platform based on Raspberry Pi (cost of a unit ca 100 EUR) with GPS module attached. Moreover it has been shown that making the environment more portable with virtual machines is not feasible due to inherent problems with time synchronisation between physical and virtual machines. During this stage of tests a great number of laboratory tests were performed. These results show that current level of delays as well as the throughput is sufficient to enable video transmission in stationary conditions (and good radio coverage). The same behaviour has been confirmed for mobile tests with mobility in the range of 10-30km/h in the range of 1km from base station. The Annex A, introduces the details of the:

- Radio traces collected together with particular parameters collected (e.g. GPS location, SINR, modulation)
- Delay measurements instrumentation (using MGEN tool)
- Time synchronization own QoS probe (using RaspberryPi).

3.9 SELECTION OF QOE STATISTICS FOR EVALUATING QUALITY OF VIDEO FEEDS

The attractiveness of well-defined set of QoE metrics for this thesis lays in the high fidelity of mimicking user perception. It makes is valuable tool for developing video adaptation (control, congestion) algorithms that can adjust QoE (or QoS profile) to the actual context, users' requirement, source of the video stream and the target of use. For example in security environment, for some cases it may be sufficient for security operator to only detect people (objects) entering restricted area, thus it requires more of a smooth, high FPS video, while high resolutions are less important. But in other cases, when it is required to e.g.

identify people and their faces, a higher image quality is more preferable than a smooth, high FPS video. Therefore, based on the comprehensive set of QoE metrics collected over time, it is possible to design and evaluate algorithms and setups to validate the models proposed in the next chapters (5 and 6). For the needs of this thesis the two groups of QoE metrics were identified as most relevant: a) temporal activity and spatial activity and b) blockloss, freezing and blockiness. The temporal/spatial activity inform about the amount of motion in the video as well as the amount of image details – respectively. While the parameters of the second group inform about the video disruption as regards smoothness of playout. The details of the selection process that led to this conclusions are presented in Annex B.

3.10 OVERVIEW OF THE EXPERIMENT CONFIGURATIONS IN CHAPTERS 4-8

In order to allow comprehensive view on the planned research work from the perspective of the wireless system model presented for this thesis in the Figure 17, the summary of foreseen settings of the various subsystems (SP, SC, SJ, KL) is presented in the Table 12. It can be seen that author has planned the experiments considering multiple settings of the environment, namely: admission domain, congestion domain, various traffic types, different directions of transmission (DL, UL), type of feedback used as well as multiple RAT technologies. The particular mapping to each chapter is provided.

Table 12 Overview of topics addressed in the reminder of the thesis

Chapter	Topics covered					
	Admission	Congestion	Traffic	Direction	Feedback	RAT
Chapter 1 (DBAC, MBAC)	declaration and measurement based	Yes	VoIP	DL	QoS, GoS	4G, 5G
Chapter 2 (E2E congestion)	No	adaptive traffic + control	real-time video	UL	QoE metrics	4G
Chapter 3 (multimedia)	No	traffic steering	non-priority traffic	DL	QoE metric X, SINR	4G, WiFi
Chapter 4 (workload prediction)	Yes (computing level)	Yes (computing level)	YT, FTP	DL/UL	CPU utilization	disaggregated virtual 5G RAN
Chapter 5 (AI/ML based CAC)	Yes (RL based)	No	CBR (UGS, BE)	DL	QoS	4G, 5G

In the chapters 4-8 author focuses on solutions centralized at the level of a single cell to implement the CAC function, i.e. exchange of messages between base stations are not used.

3.11 INDICATION OF ORIGINAL WORK INTRODUCED IN THE NEXT CHAPTERS

In this thesis author used various tools for validating the results of novel and modified algorithms and schemes, the enumeration of tools and its modifications done by author is summarized in Table 13.

Table 13 Summary of used tools/testbeds and own modifications (where relevant)

Tool	Author own Modification	Chapter
Own extensions / developments		
Matlab simulator (5G Vienna SLS/LL) – with author extension	Own extensions for CAC algorithms into Matlab simulator - mDHCAC, mFCAC, ARAC, RL-CAC, CS-CAC	4-8
NS2 (patch – author extension)	Own extensions for CAC algorithms into NS2 simulator - mDHCAC, mFCAC, ARAC, RL-CAC, CS-CAC	4-8
RaspberryPi based measurement probe (author own solution)	Designed and build from scratch by author in order to collect: QoS metrics, GPS position and altitude	5
Packet loss/delay statistics model (author own solution)	Wireless mobile channel measurement based, statistical generator for modelling packet losses and delays, developed and validated based on original author concept (released as MS Excel macro)	5
Linux based emulator framework (author own solution)	Original solution of network emulator designed based on a concept by author to aid adaptive controllers design (combines field test results, network simulation for signalling overheads, black-box enforcement of controls)	5
RAN controller (Python) – (author own solution)	Original design of the RAN controller written in the Python language with decision algorithm for QoE based, traffic steering optimization	6
3rd party tools		
QoE evaluation software from the AGH University [233]	Free SW version was used to perform objective measurements of the video frames	5
4G/WiFi testbed provided by the Technical University of Dresden under the ORCA project	The testbed provided by the Technical University of Dresden under the ORCA project (EC H2020), customized and tailored based on a combined USRP-NS3	6

(EC H2020) [269]	platform, that supports the LWA protocol for multi-RAT	
ORAN based 5G RAN testbed (<u>proprietary solution</u>)	Disaggregated 5G vRAN testbed with added CAC xApp and the CU-scaling feature prototype	7

It has to be noted that the structure of the following chapters will be such that at the end of the section there is summary of results. It is later complemented with the overview of achievements in chapter 9, where discussion of results is performed.

4 BANDWIDTH BASED ADMISSION CONTROL ALGORITHMS

4.1 INTRODUCTION

As indicated in the state-of-the-art chapter, the author will focus in this chapter on the admission control algorithm adaptations, to cover the currently missing requirements identified as the gaps resulting from the prior-art overview. The main reference are the algorithms already developed by the author as a result of the prior-art research (Flizikowski, Kozik, Gierszal, et al., 2009), [36]). These four CAC algorithms are based on the concept of reserving bandwidth in the admission process.

These algorithms are: i) the legacy baseline algorithm namely Complete Sharing CAC (CSCAC), as well as other algorithms implemented and modified by the author together with coauthors for the 4G i.e. ii) Dynamic Hierarchical CAC (DHCAC) [271] , and iii) Fair CAC (FCAC) [157] and introduced by the author in [272] iv) the modified version of DHCAC - modified Dynamic Hierarchical CAC, as well as v) the ARAC CAC which is able to compensate resources assessment for the MCS dynamics, as well as for the influence of connections arriving in batches in the averaging window. Bandwidth based CAC algorithms have been chosen, as bandwidth, particularly in presence of the adaptive modulation and coding (AMC) and a variable SNR environment, can be considered a scarce resource. All evaluated algorithms have been described in the following subsections. Moreover, thorough study of the proposed solutions was presented in a number of papers submitted to international conferences [272], [273] [36]. Additionally, the algorithm for admission control considering the requirements of guaranteed bitrate priority connections is also introduced as a reference to highlight the need of congestion control when system load is too high to admit connections of the high priority (GBR CAC). Its purpose is to highlight the role and importance of system load assessment and adjustment by means of e.g. degradation (adaptation) of existing connections which do not belong to the priority class.

Additionally, the author introduces the ARAC algorithm which can properly respond to connection requests arriving in batches. This enables addressing the changes of number of symbols/slots that happens due to changes of the link quality, that is managed by the AMC mechanism that controls the maximum level of the BER/CWER. The Table 14 presents the features of the proposed control algorithms for accepting new calls. Algorithms are divided into those that serve as baseline for comparisons (“baseline”) and those modernized by author (“Modified”).

Table 14 Comparison of the features of the discussed AC algorithms (Source: own)

	QoS guarantees	Decion based on	Working mode
--	----------------	-----------------	--------------

Algorithm name	Deterministic	Statistical	Declarations	Measurements	Preventive	Reactive
Baseline						
CSCAC	•		•			•
nscARAC		•		•	•	
DHCAC	•		•			•
EMAC		•		•	•	
(m)FCAC	•		•			•
Modified by author						
mDHCAC	•		•			•
GBR CAC	•		•			•
ARAC		•		•	•	

In the next sections, the author is first focusing on the reference algorithm that is used for benchmarking purposes (CSCAC) and then introduces the mDHCAC modification as well as the GBR CAC which is simple but with high potential for use in the real networks as it provides extension to the algorithms present inside solutions of 4G/5G vendors. Later the ARAC algorithm is introduced to highlight the requirements of dealing with the intense admission control and handovers resulting from expected network densification[4].

In sections 4.7 and 4.8 the test cases are defined along with results of simulations performed to respond to the research question from chapter 1:

Can selected algorithms for admission control available as prior art in 4G and based on declarations (DBAC) and measurements (MBAC) - be modernized to make them suitable for 5G / NOMA / emerging generations, so as to further optimize performance of admission in 5G and beyond networks?

4.2 COMPLETE SHARING CONNECTION ADMISSION CONTROL ALGORITHM

Complete Sharing (CS) is the least complicated CAC algorithm. Lets assume, that total bandwidth that can be assigned by Base Station is equal to B_{total} , bandwidth already assigned to ongoing connections given by B_{used} and bandwidth required by requested new connection is given by B_{req} . In this case, CSCAC algorithm can be described as:

$$D = \begin{cases} 1, & \text{if } (B_{used} + B_{req} \leq B) \\ 0, & \text{other} \end{cases} \quad (4-1)$$

Where D is the decision, 0 meaning rejecting and 1 accepting new connection. This technique assumes that an AP accepts all connections until it runs out of resources. No other factors e.g. service flow class are taken into consideration. The CSCAC is easy to implement, but it is sufficient only if the system has to

cope with only one class of service (e.g. best-effort). The 4G and 5G standards define several service classes (e.g. Quality Class Indicator/QCI in 5G/4G), which makes classic CSCAC insufficient for appropriate service differentiation. Nevertheless, this algorithm has been assumed in order to provide a reference and validation for the proposed methods of estimating remaining resources of an access point (AP). The concept of the CSCAC algorithm's operation has been presented in Figure 19. As indicated in the chapter 2, the reality of mobile network operator implementations of admission control algorithms limits itself to just differentiating call requests for guaranteed bitrate from the rest of incoming requests would be treated as BE class.

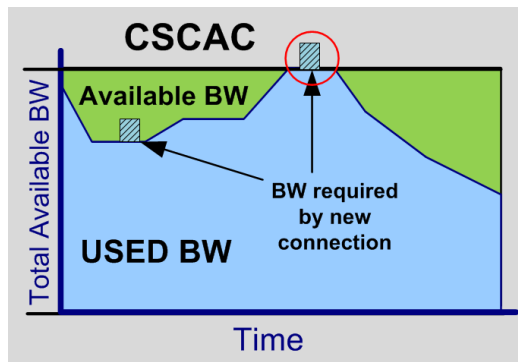


Figure 19 Conceptual example of CSCAC algorithm's functioning

4.3 DECLARATION BASED ADMISSION CONTROL ALGORITHMS

The author has already contributed to the development of modernized 4G CAC algorithms. These algorithms represent various groups of features and that can be utilized as well in the 5G networks (and beyond). The algorithms so far developed and disseminated to numerous conferences were originally targeting the 4G networks. The various aspects addressed by these algorithms that are important to the objectives of this dissertation have been presented below:

- dynamically modified threshold for guard channel (DHCAC/mDHCAC)
- dynamically weighted share of remaining resources to assure fairness among users (and single user data streams) (FCAC/mFCAC)
- various symbol reservation schemes (SRS schemes)

In the evaluation section the results of running these algorithms will be used as a reference and for comparison with the algorithms introduced in the following section.

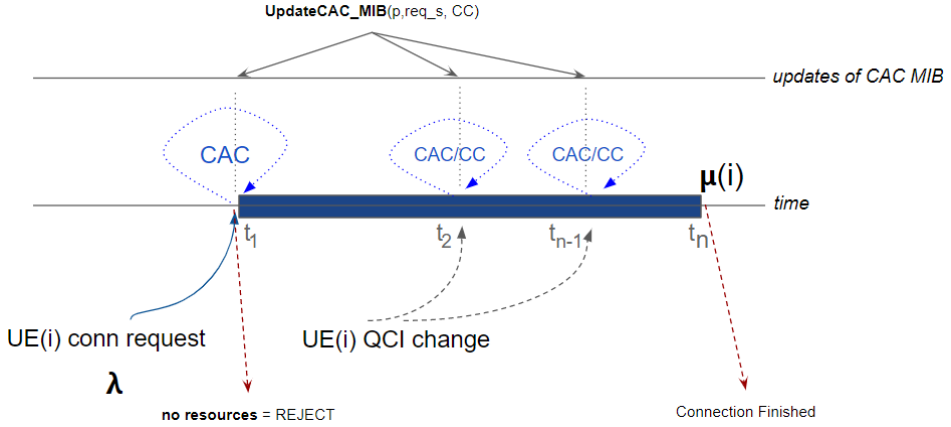


Figure 20 Generic diagram of a UE(i) connection lifecycle

The Figure 20 highlights the moments when admission control algorithm is preliminarily activated (t_1) for a new connection request. In addition the proposed algorithms (mDHCAC, ARAC) they enable admission control to be triggered for a MCS change (t_2, t_{n-1}) in order to especially handle situation of moving from higher to lower SNR value for a connection of UE_i . The blue bar represents the connection holding time (μ_i). The changes of modulations (QCI change) and the original admission control are all leading to an update of the MIB database that keeps register of all the connections statuses. That is why at decision times $\{t_1, t_2, \dots, t_{n-1}, t_n\}$ appropriate symbol reservation scheme will be possible to be applied (see 4.5).

4.3.1 mDHCAC

According to the original DHCAC algorithm referenced in the chapter 2, the condition when deciding to accept a new connection type other than UGS is as follows:

$$if(B_{used} + B_{rtPS} \leq B - U_{GBR}) \quad (4-2)$$

In this thesis author has proposed the following modification of the original algorithm [164] in the article [35]:

$$B_{used}^{non-GBR} + B_{rtPS} \leq B - (U - B_{used}^{UGS}) \quad (4-3)$$

However, the shortcoming of this approach is the unjustified reduction of the "guard band" value as the bandwidth usage of UGS connections increases, and the use of B_{used} instead of the correct $B_{used}^{non-UGS}$. Therefore, the above decision rule leads to an overestimation of the bandwidth available for rtPS connections, if the bandwidth consumed by UGS connections (B_{used}^{UGS}) begins to exceed the threshold value U - the algorithm does not protect against such a situation.

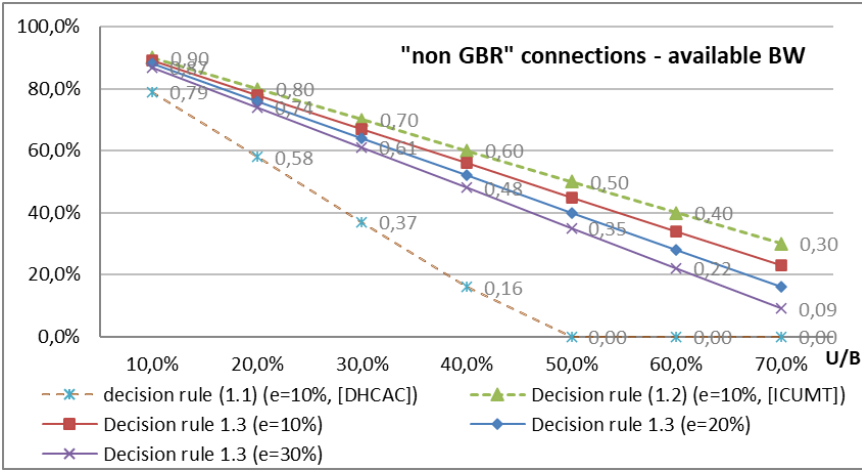


Figure 21 Efficiency of determining the available "non-UGS" bandwidth for the [272] algorithm [source: own]

For the above reasons, the expression specifying the decision rule for non-priority calls should be updated to the form below:

$$B_{used}^{non-UGS} + B_{rtPS} \leq B - \max(U, B_{used}^{UGS}) \tag{4-4}$$

However, the shortcoming of the proposal (4-3) is the problem of the decrease of the "B-U" threshold as the number of UGS connections increases. Therefore, from the perspective of the natural desire to insert this value as a parameter for system analysis, this will be an awkward approach. Computationally, the results (4-3) and (4-4) are equivalent, while the solution (4-4) seems more elegant, because the guard band threshold for UGS connections retains the form proposed by the authors, but the method of calculating the value of the used band (B_{used}) changes. The currently used bandwidth must be analysed separately for UGS traffic (B_{used}^{UGS}) and non-UGS traffic ($B_{used}^{non-UGS}$).

$$B_{used}^{non-UGS} + |U - \max(U, B_{used}^{UGS})| + B_{rtPS} \leq B - U \tag{4-5}$$

The latter inequality (4-5) will cause the bandwidth $B_{used}^{non-UGS}$ to be increased by the value of the bandwidth used by UGS connections only when the amount of bandwidth occupied by UGS traffic exceeds the allowed threshold U. The figure (Figure 21) shows that as a result of the introduced modifications, the decision rule for non-priority connections estimates the resources available more realistically - i.e. in the case of exceeding the level of the protection band U for UGS connections by = 30%, for the protection threshold set at $U = 0.5B$, the decision rule (4-2) leads to an overestimation of available resources by 18%. As the threshold U (set by the operator) is reduced, the available bandwidth estimation error decreases linearly to the level of several percent (for $U=10\%$).

The value of B is set to 1 for clarity of explanation.

4.4 GBR CAC FOR 5G ORAN

According to existing specifications [274] [275], for a GBR QoS Flows, the 5G QoS profile additionally includes the following QoS parameters:

- Guaranteed Flow Bit Rate (GFBR) – for the UL and DL directions,
- Maximum Flow Bit Rate (MFBR) – for the UL and DL directions.

The GFBR denotes the bit rate that should be guaranteed to a GBR QoS Flow after being admitted to the system by the CAC algorithm while the MFBR indicates the bit rate that may be provided to the same flow. The excess traffic may as well get discarded by a rate shaping function if decided so by the MAC resource scheduler or CAC algorithm itself. The algorithm below is author’s proposition of a custom CAC algorithm for GBR applicable to 5G vRAN based ORAN solutions with RIC and xApps. This algorithm includes congestion control and dropping of non-GRB connections (step 4) and eventually even the existing GBR traffic based on priority (in step 5). However its applicability targets early stage 5G deployments where the MAC layer is using the legacy round robin scheduler, which may fail to provide resource guarantees to GBR users. Such CAC algorithm can mainly provide benefit of introducing custom congestion control to assure that non-GBR resources are degraded in a controller fashion.

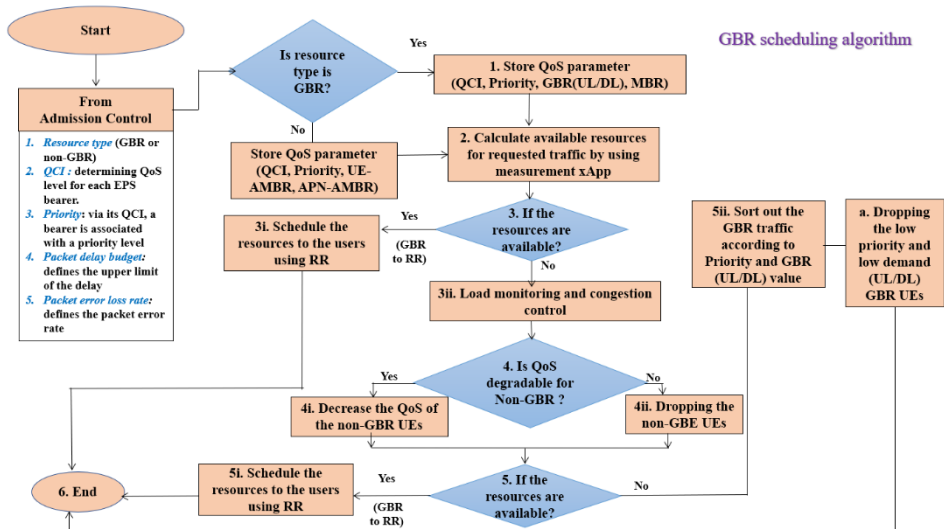


Figure 22 GBR CAC for GBR and non-GBR services

The steps of the simple algorithm which could be used by 5G ORAN network equipped with the RIC and xApps are described below:

1. After getting the parameters for admission control, the algorithm first separates and stores the requested sessions into GBR and non-GBR. The parameters considered consider:
 - a. Resource type - type of radio bearer needed by the connection
 - b. QCI – determining complete set of QoS descriptors for a given class
 - c. Priority – priority level for a connection
 - d. Packet delay budget – provide upper limit for the E2E latency
 - e. Packer error rate – the maximum loss limit
2. Then, it calculates the total resources required from all requested traffic.
 - a. Cell load calculation – in the chapter 7 the load estimation and prediction based on LSTM technique will be discussed in details.
 - b. Cell capacity calculation - refer to the section 4.5 for the capacity calculation
3. Check the availability of resources (e.g. based on rule described by equations: (2-3, (2-4):
 - a. If available, then schedule all resources using any conventional algorithm (e.g., Round robin).
 - b. If not available, then monitor traffic load and apply congestion control.
4. Check if the QoS can be degraded with minimal influence on QoS from the non-GBR traffic (use the information from system MIB about non-GBR traffic).
 - a. If yes, then degrade the QoS of the non-GBR traffic - comprehensive analysis for adapting non-GBR traffic of a video traffic class has been proposed in the next chapter 5.
 - b. If no, then, drop the existing non-GBR traffic e.g. based on the “worst-SNR” ranking criterion.
5. At this stage, proceed to session processing (scheduling data) or search for resources to be regained from GBR traffic (worst-case):
 - a. If yes, then schedule all resources using any conventional algorithm (e.g., Round robin).
 - b. If no, then sort out the GBR traffic according to priority level and GBR (UL/DL) value, packet delay rate drop the low priority and low demand (low QoS) GBR traffic to admit/process high QoS GBR traffic (high priority based on use case/application scenarios).
6. End

The steps: 3a, 3b and 4a have been highlighted as they are addressing the important aspects of calculating the load and capacity of a wireless system. These are crucial for any wireless system admission control proper operation. As regards the load calculation in step (3a) and the congestion control in step (4a) the next section will introduce the symbol reservation schemes for the OFDM-based systems, including congestion control aspects. These schemes will be further used in the algorithms proposed below to tune the resource allocation strategies. Furthermore, the congestion control topic will be addressed in the next

chapter 5, where the complete framework will be described on how to design closed-loop controllers for the surveillance use-cases or more generally for the recognition tasks use-cases (see [276] for reference). The later specification identifies the QoE requirements that must be met in order to be able to identify objects in the video stream with certain level of reliability and precision.

4.5 SYSTEM CAPACITY ESTIMATION - ACM AND SYMBOL RESERVATION

As already introduced in the section 2.1, to keep the CWER $< 10^{-3}$ e.g. to mitigate channel dynamics due to e.g. user mobility or alternative reasons, the allocation of optimal modulation and coding scheme can lead to a change in resources required for a connection. And especially the resources accepted at the connection arrival will be increased due to higher protection needed, e.g. more robust modulation and coding. This way changes of the channel can lead to increased demand for radio resources (e.g. PRBs). The capacity measured in bits is expressed in 4G and 5G systems in symbols or PRBs (see section 2.6).

In OFDM-based systems the overall capacity can be expressed by the number of available OFDM symbols/slots. In the 4G/5G wireless systems however this does not provide reliable capacity information in terms of bits per symbol, as the capacity depends on a modulation and coding scheme (MCS) used. Moreover, for systems with adaptive coding and modulation (ACM) an MCS index for a given connection can change over time, which means that number of symbols required meeting particular connection's QoS demand will vary and thus - influence capacity. This fact can be expressed after [277] in the following equation:

$$C_i(t) = N_r * R_i(t) \quad (4-6)$$

where: $R_i(t)$ is the number of packets that can be carried by one time slot and is determined by the channel quality of connection i via AMC as in table (Table 8). Here both - $R_i(t)$ or $C_i(t)$ indicate the channel quality or capacity.

The process of estimating and reserving an adequate number of symbols at CAC level in this chapter is referred to as Symbol Reservation Scheme (SRS). Although estimating and reserving number of symbols required to serve connection is a crucial part of the QoS provisioning stage in mobile systems, it has been given very little attention when researching admission control algorithms. Methods that could be used in order to cope with this problem have been presented below. Moreover, simulator used by the author in this thesis, which has been used to evaluate Admission Control algorithms is – to my best knowledge - the only open source SLS module that provides support for symbol reservation schemes.

“Worst-case-scenario” SRS (WCSRS): One method to cope with this problem could be to assume “worst-case-scenario” - estimate number of required symbols according to most robust MCS, therefore reserving maximum number of symbols for a connection. If a connection uses less symbols than reserved maximum, free slots could be used by low priority connections e.g. BE connection. Although this approach ensures no connections will be dropped due to lack of resources, bandwidth utilization for this method would not be optimal in case most terminals use less robust MCS and there are few BE connections.

SRS with reservation factor (RFSRS): Other method could be to estimate number of required symbols according to formula:

$$S_{rsvd} = S_{min} + \lambda * (S_{max} - S_{min}) \quad (4-7)$$

where:

S_{rsvd} – is the number of reserved symbols.

S_{min} - denotes minimum number of symbols required to serve connection.

S_{max} - denotes maximum number of symbols required to serve connection.

α - symbol reservation factor in range from 0 to 1.

Although this method could result in higher bandwidth utilization and lower connection blocking probability, it is necessary to find an optimal value of α . The study of influence of the SRS schemes on the CAC performance will be provided in section 4.9.

SRS with Congestion Control (CCSRS): Another method dealing with system capacity is reserving number of symbols required for the particular MCS as requested at time of creating new connection. In turn an admission control algorithm would be triggered each time the MCS changes [171]. If there is not enough resources to serve connection using a new MCS, connection is dropped. In this approach admission control algorithm works as combined Admission Control and congestion control (CC) algorithm. As proposed in [171] switching to higher order modulation (less robust) would always be accepted, as it requires fewer resources (higher spectral efficiency). This method should result in higher bandwidth utilization and lower connection blocking probability of new connections as we are always reserving only a number of symbols that are required to serve all connections at a given moment. The disadvantage of this method is the fact there is a possibility that already accepted connections could be dropped. To present the importance of the symbol reservation method let us consider scenario in which user (U) using VoIP application moves from location in POSITION I to location in POSITION III (Figure 23). When a mobile terminal is changing its location from area B to area A the CAC working also as congestion control algorithm is triggered. As user U is switching to higher order modulation (less robust) his connection is maintained, and a part of symbols previously reserved for user U is now available for newly arriving connection requests. Later when user U is crossing from area A to area B, his connection gets dropped, because when he was in area A (using higher order modulation), an AP may have

already reused the unused symbols by accepting new connections from other users. Therefore, user's connection may be eventually dropped. This could be avoided by performing a congestion control algorithm that would use different criteria e.g. drop lower priority connections. Such approach is introduced into the ARAC algorithm presented in the next section. The changes in connection switching between areas of higher/lower modulation (spectral efficiency) are connected with events of triggering CAC/CC function to evaluate available resources (this time points refer to the time points t_2 , t_n in the Figure 20).

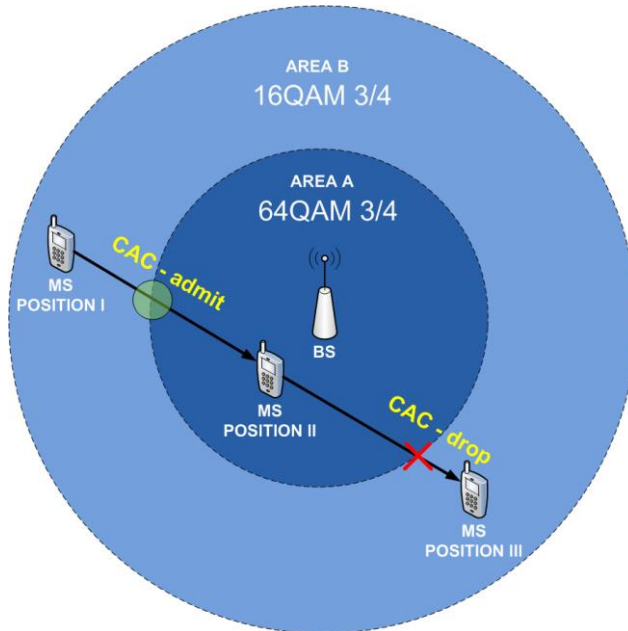


Figure 23 Changing SNR environment and CCSRS (Source: own)

The above mentioned symbol reservation modes (e.g. min, max, ratio) are used whenever it is needed by the admission algorithm to estimate required symbols for allocating resources required for a new connection or an updated one, e.g. due to change in quality of channel (*SNR change* -> *QCI change* -> *modulation change* -> *required number of bits changes*).

4.6 ARAC - MEASUREMENT BASED CAC ALGORITHM

This section introduces the CAC algorithm called ARAC that is updated version of the nscARAC and based on the moving average EMA statistics for the connections that either just appeared or recently concluded in the current averaging interval. The main rationale is to provide responses to the three major questions when designing a proper CAC algorithm:

- *Does the performance of measurement-based CAC change, if the system experiences situations, in which connection requests arrive in large batches?*
- *To what extent does the capacity change if some users follow the VoIP traffic pattern with silence-suppression enabled – depending on the admission control algorithm used (EMAC, ARAC)?*
- *How to improve resource estimation, especially when considering connection requests arriving in large batches?*

The above questions have been assessed by following the approach that assumes worst case user mobility [278]. In simulations with admission control, we decided to follow an approach similar to the one presented in [171]. This approach assumes that admission control could be triggered not only by the arrival of a new connection request. Such an approach seems logical in a system utilizing adaptive coding and modulation, since resource requirements of a given connection can change over time. Therefore, admission control is triggered in situations when:

- new connection request arrives.
- peer's MCS (Modulation and Coding Scheme) changes,
- parameters of a given service flow have been changed.

Since admission control is triggered also when parameters of a given flow have been changed, *admission control algorithms are functioning also as Congestion Control algorithms*. At the beginning of this section author first introduces design and design deficiencies of the two existing measurement-based admission control algorithms (EMAC, nscARAC) - for comparison with the original proposition of the author. Author introduces the modernized CAC algorithm which is aware of the current network state and is able to cope with the problem of batch arrivals as well as compensates evaluation of resources for the dynamic MCS changes. The algorithm is called Arrival Rate aided Admission Control (ARCAC or ARAC) and refers to a design approach of the algorithm presented in [117].

4.6.1 EMAC - algorithm

The first presented algorithm for controlling the acceptance of new calls is a simple algorithm based on the calculation of an exponential moving average (EMA - Exponential moving average). This means that in each sample (i.e. measurement), an appropriate weight is assigned. The weight is calculated on the basis of an exponentially decreasing function. The algorithm in relation to the 4G/5G network was presented, for example, in [117]. The EMAC algorithm in measurements uses the commonly used method of the so-called sliding window. Methods used to measure average resource consumption (from [101]):

- the method of the average value from the sample (Point Sample)
- the method of the maximum value in a time window (Time Window)
- exponential moving average (EMA) method.

It should be noted that in the classical approach to the control of call acceptance control, the MBAC methods usually focus on calculating the average bit rate of

the collective stream [101]. Given the certain bitrate request b_p^i from UE_i , it is possible to compute number of resource blocks to be allocated by AP p to satisfy the request:

$$n_{i,p}^{PRB} = \lceil (b_{pk}^i / r_{i,p}) \rceil \quad (4-8)$$

In 4G/5G systems, system resources should be expressed in OFDM symbols (for TDMA) or slots (for OFDMA). In the reminder of this thesis both solutions are considered for capacity modelling – the (2-3) prevails in this chapter whereas in the Chapter 6 the (2-4) is mainly considered. This is motivated by the fact that the future networks may need coexistence of various technologies in the same time (e.g. LWA, Dual Connectivity or multi-RAT in general). The study related to multi-RAT is performed in the chapter 6.

Therefore, MBAC algorithms dedicated to 4G/5G systems are based on measurements of the average symbol / slot utilization rather than measurements of the average bit rate of the collective stream. In order to track the consumption of slot/symbol resources it is useful to follow the changes of the current resources available/consumed not directly based on instantaneous value but considering some averaging and smoothing with the help of moving average like the one below:

$$EMA = Metric(t) \times k + EMA(y) \times (1 - k) \quad (4-9)$$

where: $Metric(t)$ represents the metric of interest for EMA (e.g. symbol used), t represents time in the unit of interest (e.g. today), y refers to previous value of a unit (e.g. yesterday), k represents the weighted multiplier, and N is the averaging interval given in number of units (e.g. days). The calculation of k is as follows $k = \frac{2}{N+1}$. The algorithm based on EMA averaging has been presented in [117] and in this chapter is referred to as EMAC. Below we present the pseudocode of EMAC Table 15 and the naming convention there is aligned with the diagram of the algorithm main aspect presented in (Figure 24).

Table 15 Main algorithm behind the EMA based algorithms (EMAC, nscARAC, ARAC)

Reference pseudo-algorithm for EMA-based admission control	
Step 1	Assign: $T_{window} = X$, $S_{All} = 0$, $S_{used} = 0$, $S_{req} = 0$.
Step 2	Take frame R from the current measurement window.
Step 3	Assign: S_{All} = the total number of available symbols in frame R, S_{used} = the sum of symbols used by connections in frame R, S_{req} = the number of symbols required to handle a new connection (e.g., based on MSTR).
Step 4	Calculate the average predicted bandwidth consumption for frame R using the formula $S_{pred} = (S_{used} + S_{req}) / S_{All}$
Step 5	Store S_{pred} in a data structure.

Step 6	If there are more frames in the measurement window, go back to Step 3.
Step 7	Calculate the Exponential Moving Average (EMA) from the samples stored in the data structure.
Step 8	If the average bandwidth consumption EMA is lower than the allowed maximum bandwidth consumption, accept the new call; otherwise, reject the new call.

Since EMAC does not provide protection against problem of estimating resources when connections start arriving in large batches (EMAC underestimates number of used symbols - Figure 24), in [117] authors propose a threshold – based solution. Value of guard channel (threshold) is adjusted based on the value of the declared Minimum Reserved Traffic Rate (MRTR) of existing connections and recent bandwidth utilization.

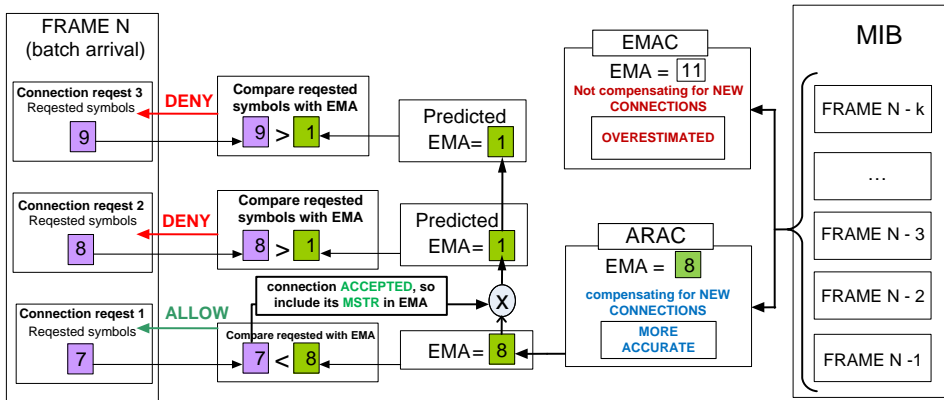


Figure 24 EMAC vs. nscARAC – example of the process of estimating resources for four frames (Source: own)

The disadvantage of the EMAC algorithm is the fact that as the length of the measurement window K increases (K is the number of samples/measurement intervals used to calculate the mean value), the impact of newly accepted calls on the mean EMA decreases.

Table 16 Pseudo algorithm for EMAC calculation of resources used (S_{used})

S _{used} calculation for EMAC	
Step 1	$P_{used} = 0$
Step 2	Take connection P from frame R
Step 3	P_{used} = the number of symbols used by connection P
Step 4	Add P_{used} to S_{used}
Step 5	If there are more connections in frame R, go back to Step 3

Consequently, if a large number of new calls arrive in a period t_n , short in relation to the length of the measurement interval K ($t_n \ll K$), because the EMAC algorithm is based on the EMA value, it becomes unreliable (it does not take into account newly received calls). We deal with a similar situation when the Adaptive Modulation and Coding (AMC) mechanism is activated, in the case of changes in the SNR level and, as a result, the value of the MCS index of users. It should be noted that the impact of the AMC mechanism on the estimation of the average resource consumption may be significant, and this will translate into the performance of the EMAC algorithm. This is because changes in the MCS value can happen much more often (even every few super frames) compared to the probability of batch arrivals. Therefore, modifications are needed that will allow the CAC algorithms to correctly take into account the impact of recently accepted requests appearing in parallel.

4.6.2 nscARAC algorithm

As mentioned in the previous chapter, the EMAC algorithm can estimate the momentary level of system resource utilization, but it is vulnerable to rapid changes in the intensity of new call arrivals. The nscARAC (no state control Arrival Rate aided Admission Control) algorithm, supported by measurements of the average frequency of new calls, is able to adapt to a situation where there is a sudden increase in the number of new calls.

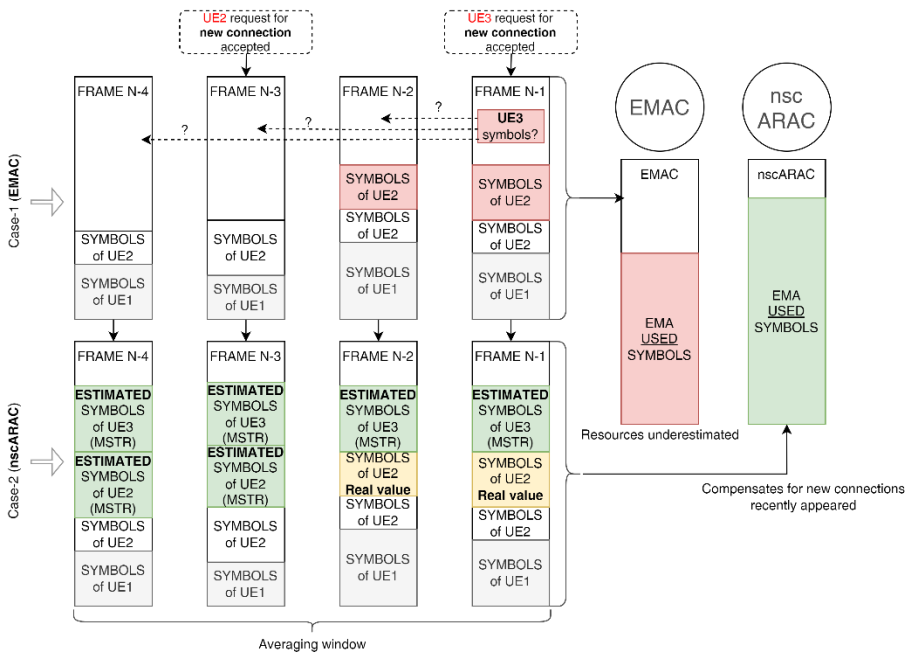


Figure 25 Example of EMA estimation for four frames, for two algorithms - EMAC and nscARAC (Source: own)

The nscARAC algorithm assumes that in 4G/5G systems the AP has the ability to measure the instantaneous speed of new calls. Consequently, in a situation where in the period t_n short in relation to the length of the measurement interval K ($t_n \ll K$) there are several new calls, nscARAC calculates the average EMA and takes into account not only the actual consumption of resources for the last measurement interval, but also motion descriptors of newly accepted calls (Figure 25). Otherwise, the potential impact of newly accepted calls on average resource consumption would be missed. This would be because newly accepted connections (those whose lifetime $t_e \ll K$) do not have much influence on the average EMA (little or no data transferred). The pseudo-code of the algorithms of nscARAC is presented in Table 17.

Table 17 Pseudo algorithm for nscARAC calculation of resources used (S_{used})

S_{used} calculation for nscARAC	
Step 1	$P_{used} = 0$.
Step 2	Take connection P from frame R.
Step 3	T_{conn} = connection's duration.
Step 4	If the connection has been active for less than the measurement window length, proceed to Step 5. Otherwise, go to Step 6.
Step 5	If the connection already existed in the current frame R, then P_{used} = the number of symbols used by connection P. Otherwise, P_{used} = the average predicted bandwidth consumption for connection P. Proceed to Step 7.
Step 6	Assign: P_{used} = the number of symbols used by connection P. Proceed to Step 7.
Step 7	Add P_{used} to S_{used} .
Step 8	If there are more connections in frame R, go back to Step 3.

Although the nscARAC algorithm can estimate the degree of resource consumption in the case of newly accepted calls, it is not able to estimate the potential changes in the degree of bandwidth consumption caused by the change of modulation and code rate of already existing connections.

4.6.3 ARCAC algorithm

Instead of using predefined thresholds, the proposed ARCAC takes an advantage of the fact that Base Station (BS) can monitor information about current arrival rate. Based on this value BS calculates, if the measured EMA of resources used does take into consideration recently accepted connections. If connection requests start arriving in large batches, in order to predict future value of average free symbols ARCAC also takes into consideration QoS parameters (e.g. MSTR) of

connections that have already been accepted, but do not exist long enough to influence average symbols utilization (Figure 27). Below we present the pseudo – code of ARAC (Table 18).

Table 18 Pseudo algorithm for ARAC calculation of resources used (S_{used})

S_{used} calculation for ARAC	
Step 1	$P_{used} = 0$.
Step 2	Take connection P from frame R.
Step 3	T_{conn} = connection's duration.
Step 4	If the connection has been active for less than the measurement window length, proceed to Step 5. Otherwise, go to Step 6.
Step 5	If the connection already existed in the current frame R, then P_{used} = the number of symbols used by connection P. Otherwise, P_{used} = the average predicted bandwidth consumption for connection P. Proceed to Step 7.
Step 6	P_{used} = the number of symbols used by connection P. Proceed to Step 7.
Step 7	If the MCS (Modulation and Coding Scheme) value has changed during the current measurement window, estimate \hat{P}_{used} based on the number of transmitted bytes and the new MCS value, then assign $P_{used} = \hat{P}_{used}$.
Step 8	If the connection has ended during the current measurement window, assign $P_{used} = 0$; otherwise, assign P_{used} = the number of symbols used by connection P.
Step 9	Add P_{used} to S_{used} .
Step 10	If there are more connections in frame R, go back to Step 3.

This chapter introduces a modified nscARAC (i.e. MAAC) algorithm called Arrival Rate aided Admission Control (ARCAC or ARAC). It is modernized approach to the measurement-aided admission control (MAAC) algorithm presented in the same article [117] as aforementioned EMAC. Since EMAC does not provide protection against problem of estimating resources when connections start arriving in large batches (EMAC underestimates number of used symbols – Figure 27), in [117] authors propose a threshold – based solution. Threshold is adjusted based on Minimum Reserved Traffic Rate of existing connections and the most recent bandwidth utilization.

Unlike nscARAC, the ARAC is able to estimate not only the resource consumption of newly accepted calls, but also the potential changes in bandwidth consumption due to modulation and code rate changes of already existing connections. In addition, the algorithm takes into account not only the impact of newly accepted calls and changes to the MCS of existing connections, but also takes into account the resources freed up by connections that ceased to exist during the last measurement window. The ARAC measures arrival rate and also it improves the EMA calculation by excluding connections which do not exist anymore although their past statistics still influence EMA. Moreover, ARAC

compensates for connections that have changed their MCS during connection, meaning their current bandwidth requirements have changed e.g. due to mobility.

Instead of using predefined thresholds, ARCAC takes an advantage of the fact that AP has the ability to monitor and identify information about current arrival rate. Based on this value an AP estimates, if measured EMA of free resources does take into consideration recently accepted connections. If connection requests start arriving in large batches, in order to predict future estimate of free symbols, ARCAC also takes into consideration QoS parameters (e.g. MSTR) of connections that have already been accepted, but do not exist long enough to influence average symbols utilization (Figure 27).

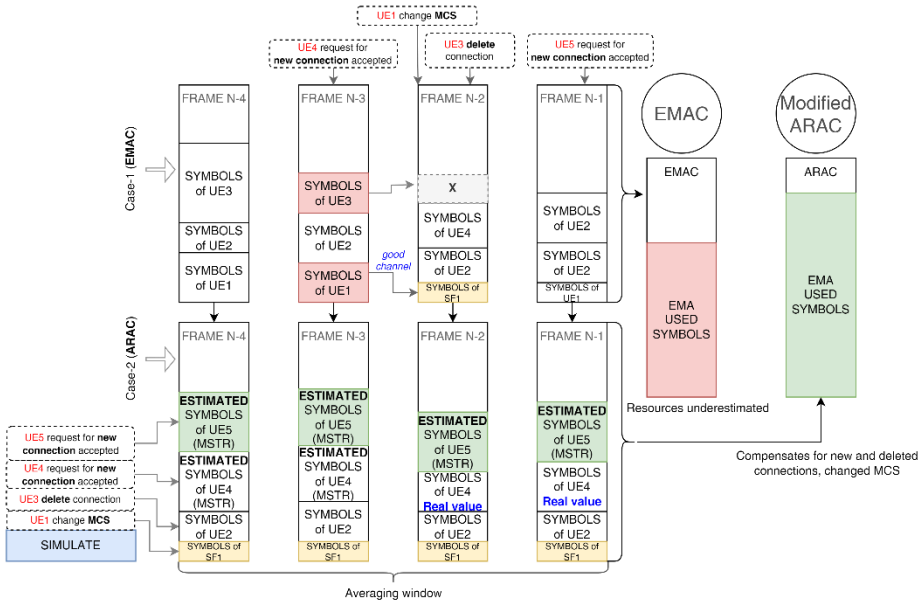


Figure 26 Example of estimating (EMA) for four frames and two algorithms - EMAC and ARAC (Source: own)

The Figure 26 shows a comparison of the ARAC algorithm with the EMAC algorithm. It should be noted that although the ARAC algorithm should be able to estimate the average value of resource consumption in the most accurate way among the presented MBAC algorithms, it is happening at the increased cost of computational complexity of the algorithm. This is because the algorithm must take into account all MCS changes of existing connections, while MCS changes may occur with high frequency (e.g. once every few super-frames).

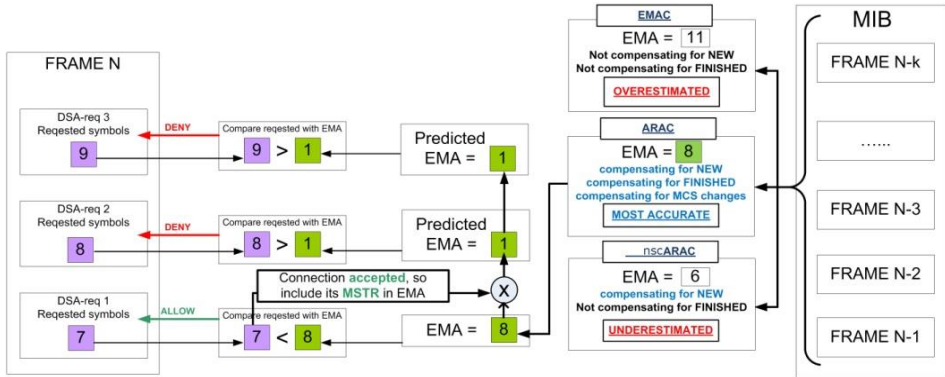


Figure 27 Comparison of concepts behind algorithms of EMAC, nscARAC and ARAC (Source: own)

In summary the comparison of the three concepts behind the algorithms of EMAC, nscARAC and ARAC is presented in the Figure 27. It should be noted that from the perspective of the approach to estimate required resources in case connection requests come in batches, the proposed ARAC is designed to be most robust. It not only compensates resources for the new and finished connections but also compensates for the dynamic MCS changes that can happen quite often in case of more mobile scenarios. It should be noted that the k frames observed by the EMA (represented by the MIB database entries in the figure) is considered for calculating the influence of connection requests coming/going in the current frame N . It should be noticed that the EMAC overestimates resources needed for new connection while the nscARAC underestimates these resources. It happens as the first is not considering compensation for the recent new or finished connections, while the latter only compensates for the recently appearing new connections that are in the averaging window too short to be noticed by the EMAC.

4.7 TEST SCENARIOS

This subsection describes simulation scenarios used to verify, evaluate, and compare implemented Connection Admission Control algorithms. All test scenarios assume single-cell environment. To minimize simulation time, the bandwidth for test scenarios has been set to 2MHz unless indicated otherwise. Connection arrival rate (for scenarios where applicable) varies between simulations from 0.1 up to 3 connections per minute. These values have been chosen accordingly to bandwidth (and therefore available resources) and average duration of existing connections. For scenarios with more resources (e.g. bandwidth 3.5 MHz) higher values of arrival rate could be chosen. Results obtained for described scenarios are presented in 4.8. The list of test scenarios together with their aims has been presented in Table 19.

Table 19 List of scenarios

Test	Objective	Channel
Scenario I	performance comparison of CAC with ACM and two Symbols Reservation Schemes (SRS)	flat channel
Scenario II	performance comparison of CAC, ACM with Congestion Control for two FEC schemes	CWER < 0,01
Scenario III	performance comparison of the EMAC and nscARAC	Map based SNR traces based on Leavy walk model
Scenario IV	performance comparison of the ARAC and nscARAC	
Scenario V	performance comparison of the EMAC and ARAC	

4.7.1 Scenario I – Symbols Reservation Schemes (SRS)

This scenario aims at measuring efficiency of CAC with ACM and two symbol reservation schemes. Measurements are conducted assuming varying SNR environment and CSCAC is used as admission control algorithm. Therefore, two test scenarios for this configuration are assumed:

- Case 1A – CAC and “worst-case-scenario” SRS (WCSRS)
- Case 1B – CAC and SRS with Congestion Control algorithm (CCSRS)

Bandwidth utilization (B_U), blocking probabilities (P_B) and dropping probabilities (P_D) have been measured for two symbol reservation schemes - “worst-case” SRS (WCSRS) and SRS with Congestion Control (CCSRS). For WCSRS we used CSCAC. In the case of CCSRS author is using CSCAC as combined admission control and congestion control algorithm. CSCAC does not prioritize connections due to their service flow class and here interest is mainly in the number of symbols required by connection due to MCS changes. Therefore, only CBR traffic is considered.

Table 20 Parameters used in Case 1A / Case1B

Parameter	Value
total bandwidth (BW)	2MHz
Arrival Rate (AR)	0,1 - 3 conn./min
avg. connection time	8 minutes
MAC scheduler	Round-robin
cyclic prefix (CP)	0.125
Modulation (MCS)	BPSK $\frac{1}{2}$ - 16QAM $\frac{3}{4}$
frame duration	20ms
traffic	CBR - 200B

Table 20 describes system parameters used for this configuration. All connection requests are generated according to the Poisson process. Holding times of

connections are exponentially distributed with mean value of eight minutes duration. Modulation and coding schemes for all users adapt during simulation in response to their movement and the underlying SNR map. We assume user cannot have higher modulation than 16QAM $\frac{3}{4}$ for sake of simulation execution time. For simplicity we assume SNR values are the same for both uplink and downlink. ACM thresholds for the Reed Solomon Convolutional Coding (RS-CC) codes were used (Table 23) which comes from the specification [279]. The three symbol reservation schemes (reserve Max; reserve Min; CSCAC with Congestion Control) - have been validated during simulations. For sake of the simulation time MCS has been limited to range between QPKS $\frac{1}{2}$ and 16QAM $\frac{3}{4}$. All users are mobile (“worst case” mobility model) and move along paths on generated map (Figure 28). For more information about SRS, paths and the maps used they are introduced in section 4.7. The detailed simulation parameters have been presented in Table 21 and Table 22. The map has been generated for a village near Warsaw (Poland). Process of map generation and idea of modulation transition files functioning as lookup tables for the simulator, has been described in section 3. SNR thresholds are set for Reed-Solomon convolutional coding (RS-CC) as defined in standard [279]. The map showing SNR distribution and the history of movements of three exemplary users has been presented in the Figure 29. Results together with analysis for this scenario have been presented in the section 4.8.

4.7.2 Scenario II – SRS with two FEC schemes

In this scenario author considers simulation cases for two distinct maps with two FEC schemes – DaVinci nb-LDPC and Reed-Solomon convolutional coding (RS-CC). Therefore, four test scenarios for this configuration are assumed:

- Case 2A – Reed Solomon Convolutional Coding (RS-CC) with Map 1
- Case 2B – DaVinci nb-LDPC with Map 1
- Case 2C – Reed Solomon Convolutional Coding (RS-CC) with Map 2
- Case 2D – DaVinci nb-LDPC with Map 2

Table 21 Simulation parameters		Table 22 Traffic characteristics	
Network configuration parameters	Value	Network configuration parameters	Value
AC algorithm	CSCAC	Arrival rate	1 to 30 (conn/min)
SF class	UGS	Average Connection Time	30s – exponential
CBR — declared throughput	320 kbps	Guard Channel	NONE
Link2System (L2S) abstraction	OFF		
Simulation time	1800s		
Scenario Repetitions	10		

SNR	Changing (Map1)
Target CWER	0.01
Carrier frequency	3.5 GHz
Bandwidth	3.5 MHz
Frame length	5 20 ms
Number of sub-carriers	256
Number of data sub-carriers	192
Cyclic prefix	1/8
MCS	QPSK 1/2 - 16QAM 3/4
Transmission direction	uplink

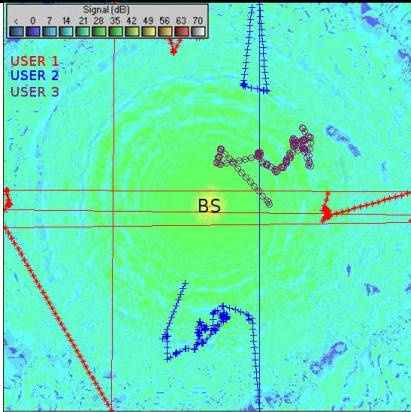


Figure 28 SNR map/user mobility models for scenario IV (Map1)

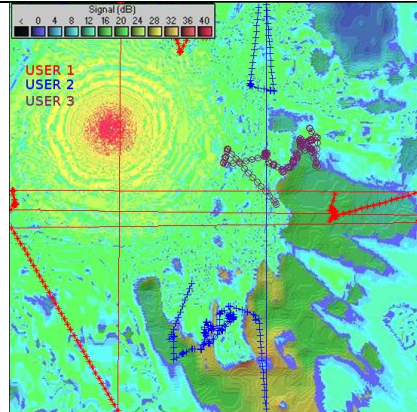


Figure 29 SNR map/user mobility models for scenario II Case C / D

In order to be able to compare the gains of two different FEC codes the two sets of ACM thresholds were taken (Table 23). The thresholds are optimized for two particular FEC codes – Reed Solomon Convolutional Coding (RS-CC) and nb-LDPC (DV). The first one comes from the specification [279].

Table 23 SNR thresholds for scenario II – Case A - D

Modulation	BPSK	QPSK		16QAM		64QAM	
Coding rate	1/2	1/2	3/4	1/2	3/4	2/3	3/4
nbLDPC	-0,88	1,79	4,77	6,75	10,62	14,18	15,78
RS-CC	3	6	8,5	11,5	15	19	21
nbLDPC	3,88	4,21	3,73	4,75	4,38	4,82	5,22
Gain							

The above ACMs were used against two different SNR scenarios to present results of system capacity assessment with ACM enabled and various FEC codes. Moreover, CAC algorithm with SRS behavior is studied for varying SNR environment. To provide more realistic simulation parameters authors have proposed to combine the output of the radio coverage planning (SNR matrices) with the realistic mobility patterns of mobile users (taken from field measurements facilitating DTN networks). For the nb-LDPC we assume code word error rate is always less than 1% ($CWER < 0,01$). First map utilized here comes from Scenario I. Figure 29 shows second SNR map used. The map has been generated for a village near Katowice (Poland). Map 1 represents good SNR conditions (on average) whereas Map 2 mimics a bad SNR environment. As mentioned before, user mobility patterns are generated according to the Leavy-Walk model [280]. The arrival rate of user requests follows the Poisson process. The CSCAC is configured to handle both admission and congestion control algorithm. The code word error rate (CWER) for both FEC schemes in presence of ACM is assumed to be 1%. All simulations have been repeated 20 times to assure statistical reliability of results. All figures present average values together with 95% confidence interval. Results together with analysis for this scenario have been presented in 4.8 Results together with analysis for all of the aforementioned scenarios have been presented in next subsection of this chapter.

4.7.3 Scenario III – EMAC, nscARAC comparison

The Table 24 shows the settings of the system used in the simulation. Mobile users move around the map with the SNR distribution generated for the suburban area (Warsaw area). The SNR values picked from the map is transferred to the L2S (Link to System interface) as white noise. An example of using the map is shown in Figure 30.

Table 24 System settings

Parameters	Value
AC algorithm	EMAC nscARAC
Carrier frequency	3,5 GHz
Bandwidth	3,5 MHz
Number of sub-carriers	256
Number of data sub-carriers	196
Cyclic prefix	0.125
MCS	QPSK, 16-QAM, 64-QAM
coding rates	1/2, 2/3, 3/4, 5/6
code word length	48, 96, 144, 288
speed (for mobile users)	0.83 m/s
Transmission direction	uplink
L2S	RBIR

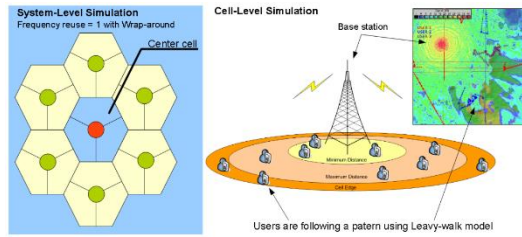


Figure 30 Example of usage of map-based SNR

Table 25 presents the parameters used during the simulation of the first scenario. Simulations were carried out for mobile users - they move on a map with designated SNR areas. In order to be able to measure the performance of MBAC algorithms, in addition to VoIP CBR traffic (treated as UGS), VoIP VBR traffic treated as rtPS class traffic was also introduced. In the case of VoIP VBR traffic, we use two codecs - G.711 and G.729. 33% of users use UGS, while 66% use rtPS (including 33% G.711 and 33% G.729). The algorithm used as AC also functions as an overload control algorithm.

Table 25 Scenario 1 - parameters

Parameters	Value
AC algorithm	EMAC nscARAC
measurement range	0,4 s
Intensity of new arrivals	25 to 250 conn./min (Poisson) per traffic class
average duration of UGS rtPS calls	20 s (exponential)
frame duration	20 ms
SNR	Changing (Map1)
ACM	ON
overload control algorithm (CC)	ON
forward error correction (FEC)	nbLDPC
L2S	ON
CWER	10^{-3}
rtPS – traffic - VoIP codecs	G.711 G.729
call type (rtPS VoIP)	unicast
queuing algorithm	Round-robin
UGS VoIP	64 kbps
Simulation time	200 s

Scenario Repetitions	8
----------------------	---

4.7.4 Scenario IV – ARAC, nscARAC comparison

The scenario tested and compared the performance of two algorithms - ARAC and nscARAC. Table 26 presents the parameters used during the scenario simulation. Simulations were carried out for mobile users - they move on a map with designated SNR areas. In order to be able to measure the performance of MBAC algorithms, in addition to VoIP CBR traffic (treated as UGS), VoIP VBR traffic treated as rtPS class traffic was also introduced. In the case of VoIP VBR traffic, we use two codecs - G.711 and G.729. 33% of users use UGS, while 66% use rtPS (including 33% G.711 and 33% G.729). **The algorithm used as AC also functions as a congestion control algorithm.** In addition, a 10% guard band for VBR traffic has been introduced. It was also assumed that the delay of the backbone network is 50ms, while the maximum allowable delay for VoIP connections is 150ms (100ms in the access layer). If the average delay of a given connection was above 150ms, it is considered that the QoS requirements of the connection have not been met.

Table 26 Parameters for Scenario 2

Parameter	Value
AC algorithm	ARAC nscARAC
measurement range	20 ms
SNR	Changing (Map1)
ACM	ON
overload control algorithm (CC)	ON
forward error correction (FEC)	nbLDPC
L2S	ON
CWER	10^{-3}
rtPS – traffic - VoIP codecs	G.711 G.729
call type (rtPS VoIP)	unicast
queuing algorithm	Round-robin
UGS VoIP	64 kbps
Simulation time	200 s
Scenario Repetitions	8

The scenario consists of two variants (Table 27):

- Scenario 2a - performance of AC algorithms for short measurement intervals (0.4 s) and average call durations (20 s)
- Scenario 2b - performance of AC algorithms for long measurement intervals (4.0 s) and short connection durations (2.0 s)

Table 27 Traffic characteristics - scenario 2a and 2b

Parameters	Scenario 2a	Scenario 2b
measurement period length	0,4 s	4,0 s
average call duration	20,0 s	2,0 s
Intensity of new arrivals	25 do 250 conn./min (Poisson) per traffic class	

4.7.5 Scenario V – ARAC, EMAC comparison

The scenario tested and compared the performance of the EMAC and ARAC algorithms for a variable length of the measurement interval at a constant frequency of new calls. Table 28 presents the parameters used during the scenario simulation. Simulations were carried out for mobile users - they move on a map with designated SNR areas. In order to be able to measure the performance of MBAC algorithms, in addition to VoIP CBR traffic (treated as UGS), VoIP VBR traffic treated as rtPS class traffic was also introduced. In the case of VoIP VBR traffic, we use two codecs - G.711 and G.729. 33% of users use UGS, while 66% use rtPS (including 33% G.711 and 33% G.729). The algorithm used as AC also functions as an overload control algorithm. In addition, a ten percent (10%) guard band was introduced for VBR traffic. It was also assumed that the backbone network delay is 50 ms, while the maximum allowable delay for VoIP connections is 150 ms (100 ms in the access layer). If the average delay of a given connection was above 150 ms, it is considered that the QoS requirements of the connection have not been met.

Table 28 Parameters for Scenario 3

Parameters	Value
AC algorithm	ARAC nscARAC
frame duration	20 ms
SNR	Changing (Map1)
ACM	ON
overload control algorithm (CC)	ON
forward error correction (FEC)	nbLDPC
L2S	ON
CWER	10^{-3}
rtPS – traffic - VoIP codecs	G.711 G.729
call type (rtPS VoIP)	unicast
queuing algorithm rtPS	Round-robin
UGS VoIP	64 kbps
Simulation time	200 s
Scenario Repetitions	8

new calls frequency	140 con./min (Poisson) – independently for UGS and rtPS
average duration of UGS rtPS calls	20,0 s
measurement range	0,2 s – 10,0 s (10 – 500 superframes)

4.8 EVALUATION OF RESULTS

This subsection presents results obtained for scenarios introduced in section 4.7 to properly validate the proposed algorithm modernizations. Connection Admission Control algorithms and resource reservation techniques introduced in section 4.5-4.6.3 have been evaluated in terms of blocking probabilities (P_B) per class of service, bandwidth utilization (BW utilization as perceived by Base Station's schedulers), E2E bandwidth utilization (BW utilization as perceived by users) and Dropping Probabilities (where applicable). Moreover, Connection Admission Control algorithm has been evaluated together with two reservation techniques in varying SNR environments.

4.9 COMPARISON OF SRS SCHEMES AND DIFFERENT FRAME LENGTHS

Figure 31 presents average blocking probabilities for all tested algorithms for both 5 and 20ms frames. Figure 32 and Figure 33 present blocking probabilities for 5 and 20ms TDD frame length respectively (for a clearer view). As in previous simulations (e.g. see [281]) “best case” SRS (reserveMin) is characterized by lowest blocking probabilities. At the same time “worst case” SRS (reserveMax) is characterized by highest blocking probabilities. Next figures present bandwidth utilization. Best case SRS relies on connections' initial declarations and assumes the best possible MCS. This is reflected in the bandwidth utilization perceived by AC algorithm (Figure 34-Figure 36), which is lowest for all algorithms and results in lowest blocking probability. Nevertheless, this leads to poor QoS of connections – refer to Figure 37. This happens because mobile users are often required to change to lower MCS (more robust scheme) to cope with channel conditions, therefore leading to increase in required symbols. CSCAC working with the proposed CC algorithm achieves moderate blocking probabilities and is able to provide appropriate QoS levels (for ongoing connections). This is achieved at the cost of some of the connections being dropped due to insufficient resources - Figure 38. The *worst-case* SRS variant, is able to provision QoS without dropping ongoing connections, but this in turn results in high blocking probabilities of new connections.

4.10 MBAC ALGORITHMS VALIDATION AND PERFORMANCE ASSESSMENT

This section presents a quantitative analysis for the following modernized algorithms based (MBAC) on measurements:

- EMAC (EMA based Admission Control) - an algorithm supported by measurements of the average bandwidth consumption.
- nscARAC (no state control Arrival Rate aided Admission Control) - an algorithm supported by measurements of the average frequency of arrival of new calls without considering MCS changes.
- ARAC (Arrival Rate Aided Admission Control) - an algorithm supported by measurements of the average frequency of new calls, considering MCS changes.

All graphs shown in next sections (4.10.1-4.10.3) measured mean values with 95% confidence intervals.

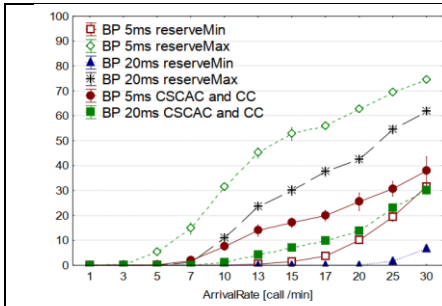


Fig. 1: Average blocking probability for frames 5 ms and 20ms

Figure 31 Average blocking probability for frames 5 ms and 20 ms

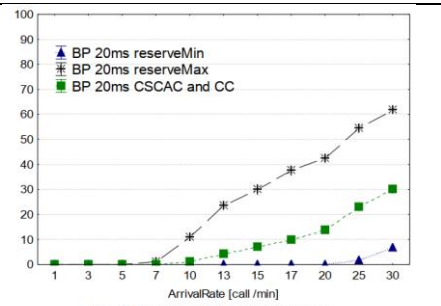


Fig. 2: Blocking probability for 20ms frame

Figure 32 Blocking probability for 20 ms frame

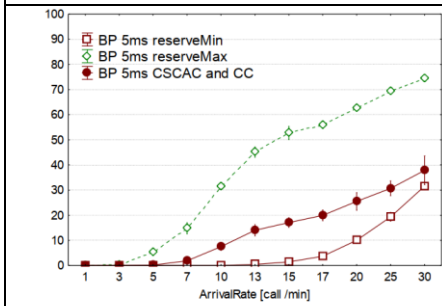


Fig. 3: Blocking probability for 5ms frame

Figure 33 Blocking probability for 5 ms frame

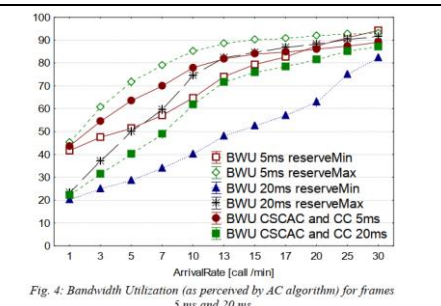


Fig. 4: Bandwidth Utilization (as perceived by AC algorithm) for frames 5 ms and 20 ms

Figure 34 Bandwidth Utilization (as perceived by AC algorithms) for frames 5 ms and 20 ms

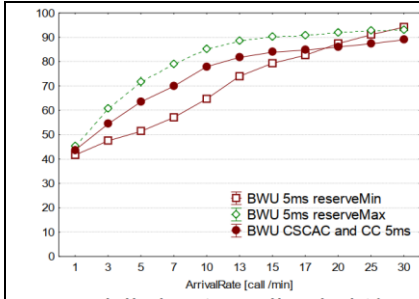


Fig. 6: Bandwidth Utilization (as perceived by AC algorithm) for 5ms frame

Figure 35 Bandwidth Utilization (as perceived by AC algorithms) for 5 ms frame

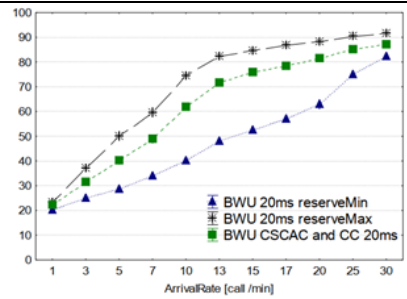


Fig. 5: Bandwidth Utilization (as perceived by AC algorithm) for 20ms frame

Figure 36 Bandwidth Utilization (as perceived by AC algorithms) for 20 ms frame

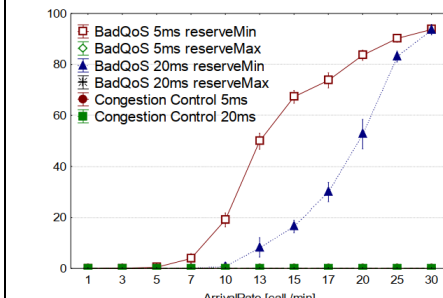


Fig. 7: QoS for tested SRS

Figure 37 QoS for tested SRS

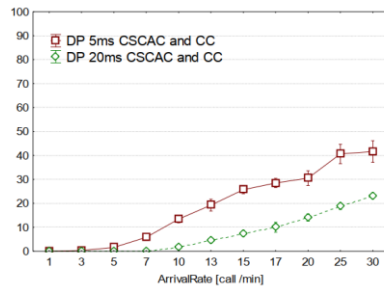


Fig. 8: dropping probability for SRS "CSCAC and CC" for frames 5 ms and 20 ms

Figure 38 Dropping probability for SRS "CSCAC and CC" for frames 5 ms and 20 ms

4.10.1 Scenario 1 – EMAC and nscARAC performance

The figures Figure 39, Figure 40 and Figure 41 show the average delays experienced by VoIP traffic for both tested algorithms. Interesting is the figure Figure 44 which indicates the intensity of the MCS changes experienced in the test scenario. It is easy to see that the influence of FEC codes is directly related with less MCS changes due to more robust codes. It can be observed that for nscARAC all types of VoIP calls experience lower latency than when using EMAC as the new call acceptance control algorithm. The nscARAC algorithm is able to more accurately estimate the resources consumed when many new calls arrive in a short period of time. This becomes even more evident the higher the frequency of arrivals. This is because with the increase in the frequency of calls, the length of the interval Δt_{req} between successive calls decreases, so the ratio of Δt_{req} to the length of the measurement window K decreases. The difference in delay for the G.711 codec and high call arrival rates is about 23%, while for the

G.729 codec the difference is 21%. In addition, differences in average latency are most pronounced in the case of codecs with silence detection enabled (G.711, G.729), which are treated as real-time (e.g. rtPS/GBR) traffic. This is because CBR traffic, treated as a UGS class, always receives a higher priority than rtPS connections. This means that bandwidth will always be allocated to UGS traffic first. Hence, in the case of mixed traffic (UGS + rtPS), the lack of resources will have a greater impact on the latency of lower priority connections (rtPS). The probability of rejecting a call for the EMAC algorithm is lower than the probability of rejecting a call by the nscARAC algorithm (about 2% for high call intensities - Figure 42). Assuming that ACM is active (to assure appropriate level of CWER) and each MCS change triggers the AC algorithm, EMAC has a clearly lower probability of premature connection termination (about 14% for high arrival rate AR - Figure 43). Although the observed delays are acceptable from the point of view of voice connections, it should be noted that the scenario assumes zero delay on the side of wide area network – i.e. from the Gi interface towards caller “B”. Hence, if the backbone network latency would on average oscillate around 80ms, ARAC should be considered as a more efficient solution.

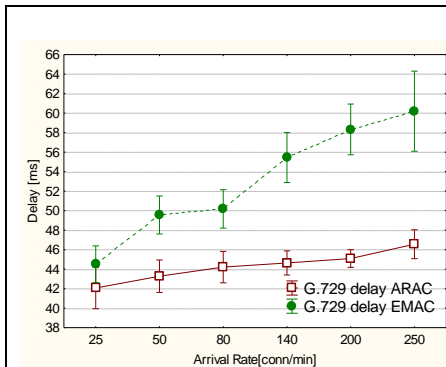


Figure 39 delay G.729 VoIP for ARAC and EMAC (rtPS)

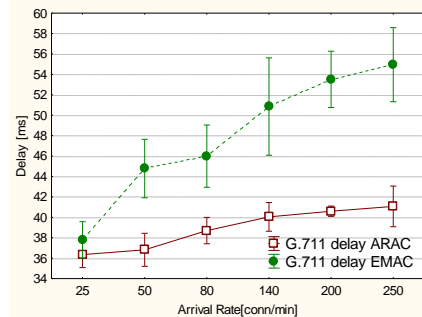


Figure 40 G.711 VoIP delay for ARAC and EMAC (rtPS)

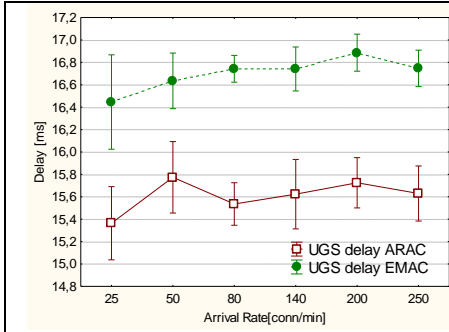


Figure 41 VoIP CBR delay for ARAC and EMAC (UGS)

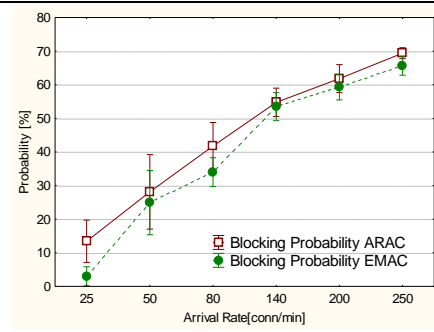


Figure 42 probability of rejecting a new call for ARAC and EMAC

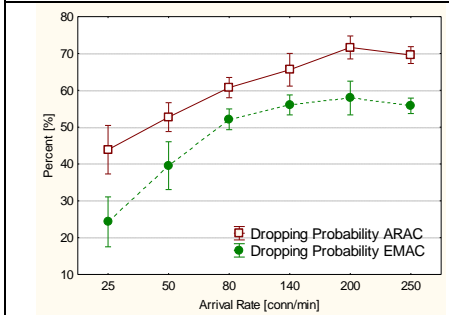


Figure 43 probability of premature connection termination for ARAC and EMAC

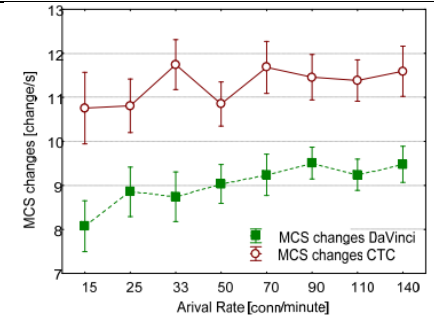
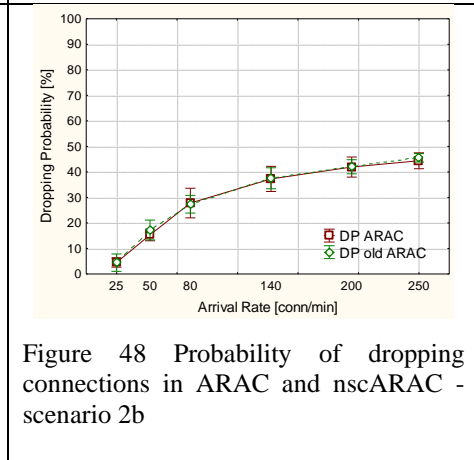
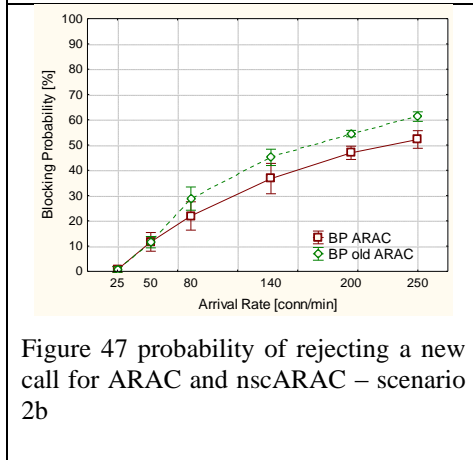
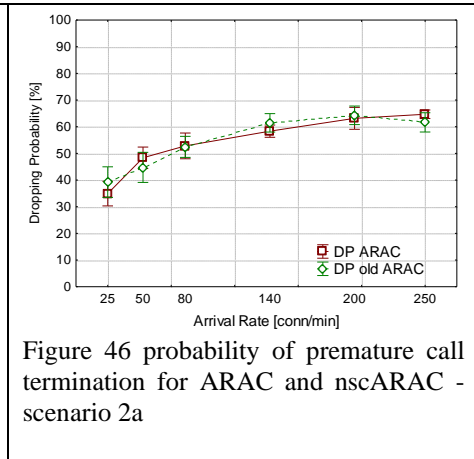
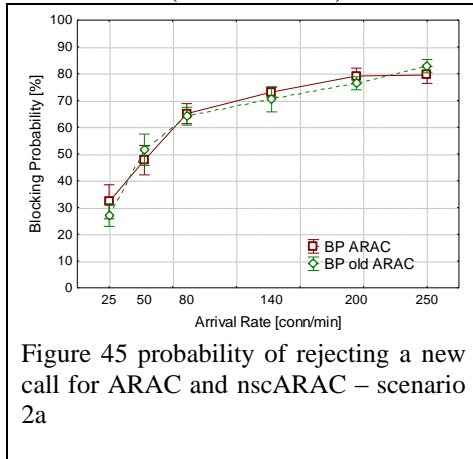


Figure 44 MCS changes for LDPBC and CTC Map1

4.10.2 Scenario 2 – ARAC and nscARAC performance

The performance of the ARAC and nscARAC algorithms for scenario 2a is almost identical both in terms of the probability of rejecting a new call (Figure 45) and the probability of premature connection termination (Figure 46). This is because for the average frequency of calls and call durations as well as relatively short measurement intervals, the differences in estimation between ARAC and nscARAC are negligibly small. This difference becomes visible only when the average connection times t_{conn} are significantly lower than the time of the measurement interval K , i.e. for scenario 2b, in which the ratio $t_{conn}:K$ is 1:2. In this case, we are dealing with differences of up to 10% in the case of the probability of blocking new calls (Figure 47). ARAC can account for the additional resources released by recently terminated calls. At the same time, the probability of premature connection termination remains unchanged for both algorithms. This is due to the fact that the change of the current MCS usually

takes place to the closest neighbouring MCS (e.g. 16QAM 2/3 to 16QAM 1/2 or 16QAM 3/4) as it can be seen in chapter 4, so that the difference in the required symbols *for applications with a low bit rate remains unchanged* (for high modulations) or changes only slightly (for low modulations). Hence, the difference in available resources - although it exists - becomes negligibly small. The difference could be noticeable only for very long measurement intervals. It is worth noting that both algorithms successfully provided QoS guarantees for all allowed calls (see Table 30).



For long measurement intervals and short connection durations, ARAC provides a lower probability of connection dropping than nscARAC (Figure 47). Despite this, ARAC remains a more computationally complex algorithm as it also monitors MCS changes and considers the impact of call termination. Therefore, the longer the measurement interval, the more time it takes to calculate the average resource consumption value. Moreover, monitoring MCS changes by the CAC algorithm, does not significantly affect performance in terms of the

probability of premature termination of the connection especially when the measurement period is short as compared to an average connection duration. This means that ARAC can be considered as a better solution in systems where we are dealing with long measurement intervals (or relatively short connections), and at the same time computational complexity is not a problem. It should also be noted that the complexity of the algorithm can be significantly reduced if we disable the MCS change monitoring option.

4.10.3 Scenario 3 – EMAC and ARAC performance

For *short and medium* measurement intervals K , the EMAC compared to ARAC, is characterized by a lower probability of call rejection (Figure 49) and a similar probability of premature call termination (Figure 50). This is because for newly accepted calls, ARAC will calculate a less optimistic value of available resources than EMAC would in the same situation. Because of this, a connection requesting an MCS change is more likely to run into a situation where there are no resources available to accept the change. However, for long measurement intervals, ARAC begins to outperform EMAC. This is because with long measurement intervals, EMAC errors in the estimation of resource consumption begin to become apparent. Figure 51 shows the percentage of connections for which the QoS requirements were not met. The graph shows that ARAC managed to ensure the appropriate level of QoS for all connections in each simulated case. However, in the case of the EMAC algorithm, errors in the estimation of available resources make it impossible to guarantee the appropriate level of QoS for all accepted connections. For long measurement intervals (500 super frames, which corresponds to 10 seconds), EMAC is unable to provide the required QoS level to almost half of the accepted connections. The unfavourable effect of errors in the estimation of available resources can be minimized by shortening the measurement interval. However, even in the case of very short measurement intervals (10 super-frames, which corresponds to 200 ms), EMAC is not able to ensure the required level of QoS for all connections (guarantees not met for 1% of connections - Figure 51), and the deterioration further grows when we extend measurement interval to 100 super-frames – here 10% of connections will not have its QoS assured due to too optimistic EMAC estimation of available traffic.

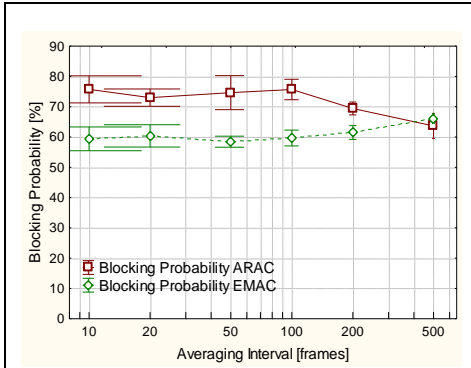


Figure 49 probability of rejecting a new call for ARAC and EMAC

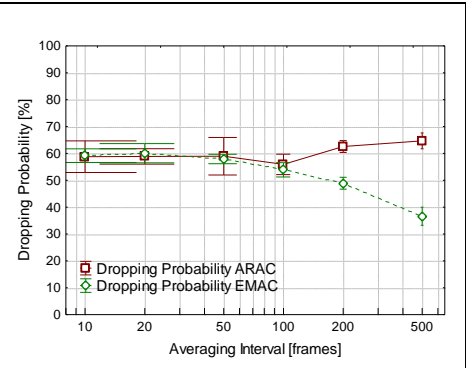


Figure 50 Probability of call dropping for ARAC and EMAC

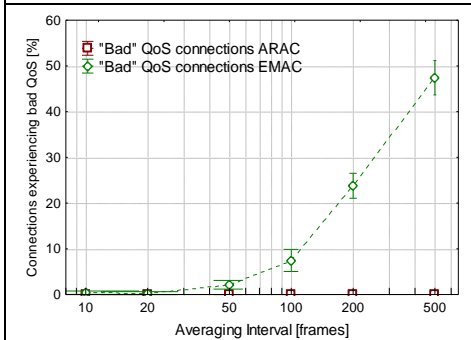


Figure 51 Percentage of connections whose QoS level was not met - ARAC and EMAC

The table below presents a summary of the results for the above test scenarios, but this time in terms of the analysis of the quality of algorithms and for a fixed value of the call arrival intensity $\lambda=280$ calls/min.

Table 29 Simulation results for the intensity of new calls $\lambda=280$ calls/min

Scenario	Scenario1		Scenario2		Scenario3	
	ARAC	nscARAC	ARAC	nscARAC	ARAC	EMAC
Percent of connections meeting the QoS target [%]	100	100	100	100	100	52
Blocking probability [%]	72	70	38	47	65	65
Dropping	59	61	38	38	62	38

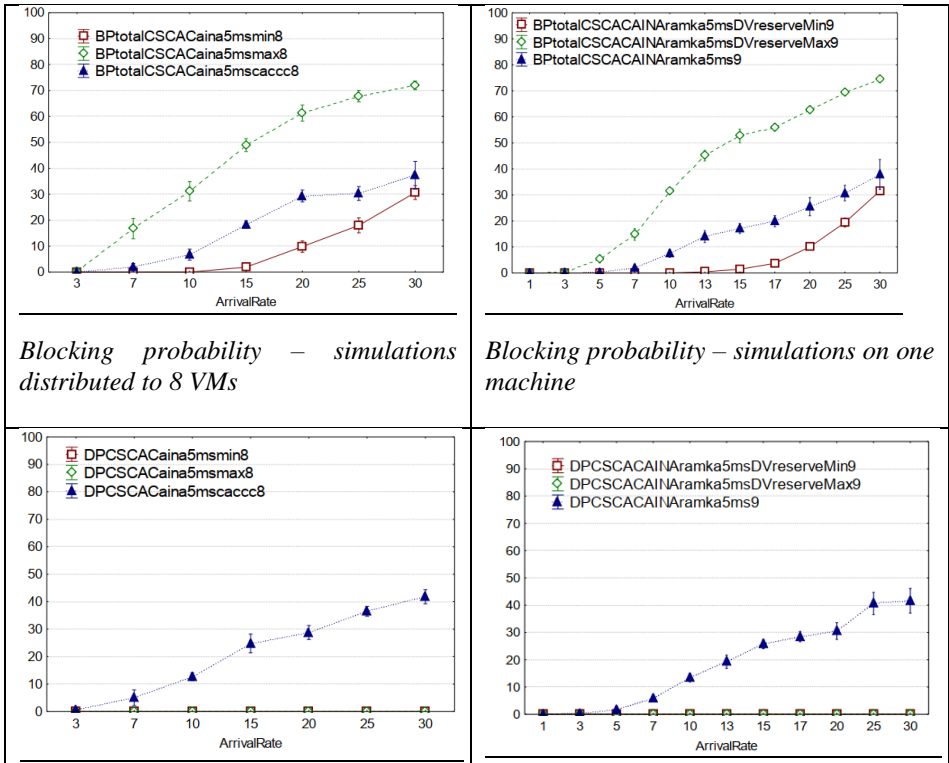
probability [%]						
-----------------	--	--	--	--	--	--

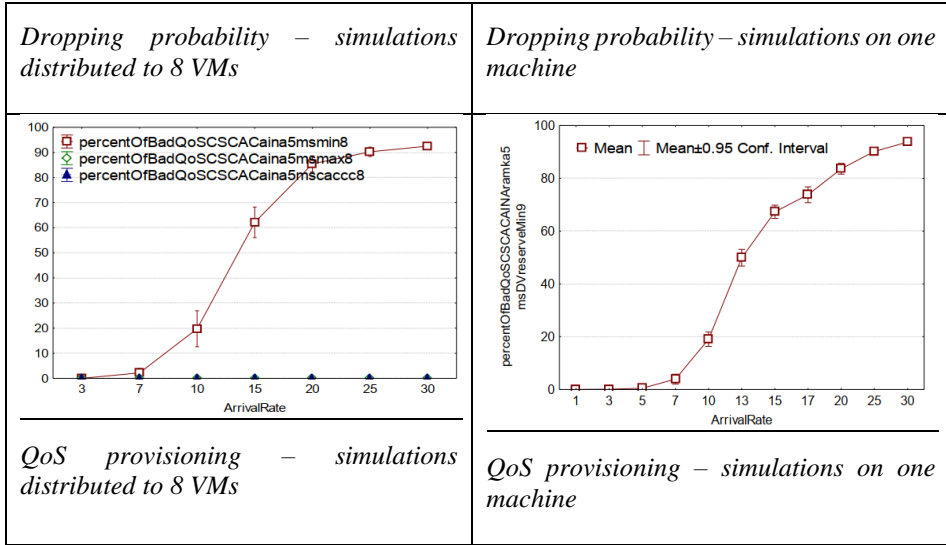
As it can be seen from Table 29 the connections admitted by the ARAC and nscARAC algorithms don't face the problem of QoS violations during connection lifetime. On the contrary the simpler EMAC algorithm decision rule does not give guarantees for the QoS levels of the connections as the "Scenario3" above shows that only 50% of connections meets their QoS requirement. Blocking probability between ARAC and nscARAC will be dependent on the scenario but also on the settings of averaging period for measurements (K).

4.11 SIMULATIONS PARALLELIZATION

To speed-up the calculations for various scenarios, author has implemented the parallelization architecture for performing simulations. The exemplary charts showing the operation of the simulator in the version for one machine with the simulator and for the parallelized version are presented in the Table 30. In the "virtual machine" case the calculations per arrival rate were distributed among a certain number of machines working in parallel.

Table 30 Validation of simulation parallelization solution implemented





As can be seen from the graphs, the results obtained for the simulation on one machine do not differ from the results obtained when distributing the simulations among machines. The environment prepared by the author to distribute simulations has been depicted in the Annex E. The plots (Figure 52. Figure 53) indicate the level of achieved simulation time gains. The gains are between 4-16 times decrease of simulations time, depending on the number of activated parallel machines (for 8-30 machines respectively).

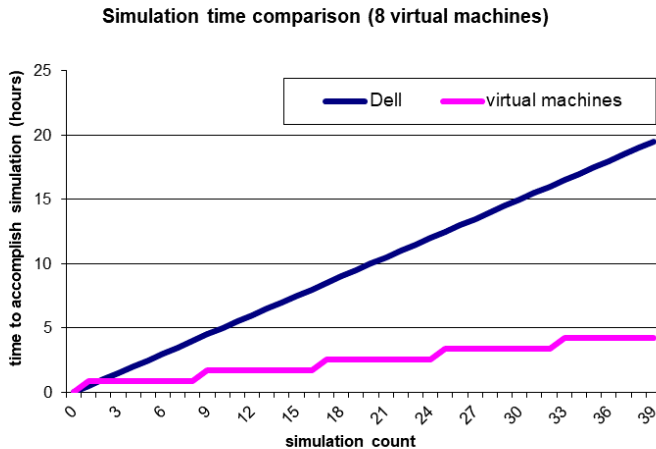


Figure 52 Comparison of simulation time for single/multiple machines

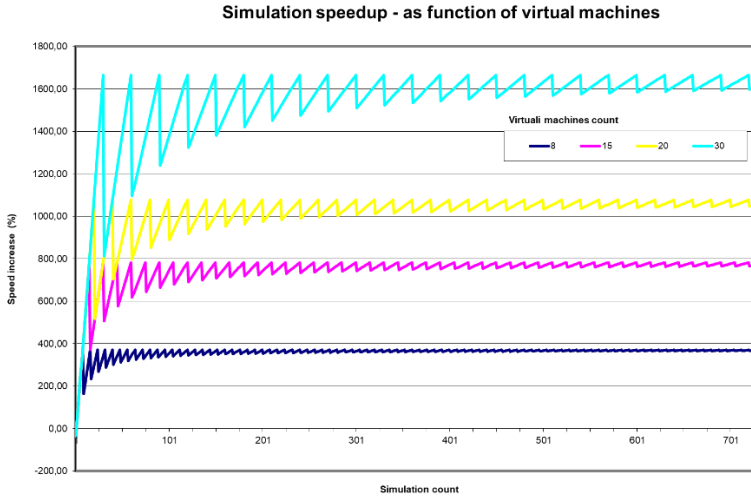


Figure 53 The Level of speed-up boost by parallelization of simulations

4.12 SUMMARY

In this chapter bandwidth-based admission control (and congestion) algorithms have been evaluated. It is worth reminding that in simulations the scheduling algorithm used was round-robin, which is good benchmark but does not allow maximization of resources per user nor per system. Still it provides simple baseline for comparisons and allows focusing on the efficiency of CAC algorithms and not the proficiency of the scheduler. In similar fashion the CS-CAC was used in order to provide a reference point to other algorithms.

The above sections show that a *moving average* based CAC algorithms are interesting solution to control the connection admission, especially if the dynamics of admission requests can be expected higher due to higher mobility or more aggressive settings of the AMC SNR thresholds. The proposed modifications of the ARAC mechanism enables dealing with MCS changes but also considering the recently added/removed connections – this is valid approach when number of relatively short connections is growing (or which duration is short in a cell due to the trend of densification) – here the connection duration is compared to averaging period (K). The reason for higher intensity of short-lived connections would be connected with the length of a connection directly or due to more connections switching to another cell due to smaller radius of the dense networks in the future.

Moreover it has been shown that resource reservation schemes SRS (it applies to both symbols and PRBs) can be combined with particular admission control and scheduling to enable pre-reservation of radio resources to account for future changes in users' channel quality, especially when more robust modulation is necessary. Here the three types of such reservation were evaluated: RFSRS,

WCSRS, CCSRS and optimal scheme. The WCSRS reserves resources assuming worst channel conditions so it is most conservative. The CCSRS reserves the actual amount of resources that directly result from a new modulation i.e. switching towards more robust modulation. Whereas the RFSRS enables reservation of resources based on the preselected ratio (β) of symbols (or PRBs) that is configured for a class of service or the whole system. The discussion of results will continue in chapter 9.1.1.

5 E2E MODELLING OF WIRELESS LINKS FOR ADMISSION AND CONGESTION CONTROL

5.1 INTRODUCTION

This section introduces the original framework enabling the design and evaluation of the video adaptation policies, applied to a user generated content (Figure 54). The main use-cases targeted here are representative to an emerging market of autonomous cars and drones. In both cases it may be required to provide remote assistance of an operator upon failure of an autonomous steering mechanisms. This way auxiliary traffic adaptation mechanisms should be present at a vehicle location, in order to enable adjusting video stream to the instantaneous capabilities of the channel. It will be even more challenging if the vehicle in question would be mobile while tele-operated.

Identifying appropriate admission control strategy (see Chapter 4) will be complemented by means introducing additional feedback loops which echo performance metrics back to the video source (i.e. UE terminal). In this case the QoE or QoS metrics are considered, in parallel with the characterization of the radio interface indicators (e.g. SINR, loss). The contributions presented here result from very pragmatic observations of perceived video quality experienced during various drive tests with a video camera attached at the UE side.

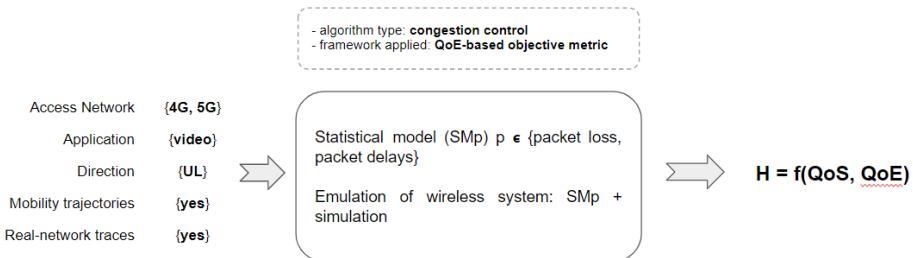


Figure 54 High level concept of the chapter goals

The goal in this chapter is to (i) propose a systematic approach to a design and evaluation of congestion control solutions for the 4G/5G networks, but with potential to be universally applied for future use-cases, (ii) perform baseline validation of a QoE of relevant video feeds, by replaying variability of mobile channel (at low speeds) based on traces collected from real 4G/5G networks across Poland and (iii) focus on selected set of QoE metrics (freezing, blockiness, blockloss) and evaluate quality for multiple settings. As it has been shown in chapter 2, there are multiple evidences in literature, of 5G network performance evaluations. The main parameters of such tests is the a) city of operation, b) modem type used, c) mobility levels, d) type and traffic direction etc.

To validate the proposed architecture author has designed test cases that validate usability of the proposed solution for the uplink surveillance and remote

monitoring traffic adaptation. A systematic approach is proposed in order to combine real network traces, network modelling and emulation, video transcoder as a service and the use of no-reference QoE metrics in order to assure effective means for video controllers design and tuning. Author has described architecture as well as results of comprehensive tests of the relevant QoE metrics in uplink for mobile 4G/5G networks. The results prove that the proposed framework provides valuable mean for development and evaluation of controller algorithms, answering the demand of emerging scenarios for mobile surveillance (e.g. use-cases inside European projects like the ECSEL JU BRAINE – use case 2, intelligent camera use-case, interest of various stakeholders including e.g. service providers like Italtel).

5.2 CONGESTION CONTROL FOR REAL-TIME MOBILE VIDEO STREAMING - SYSTEM MODEL

The overall system model for the congestion control algorithms considered in this chapter has been introduced below. Main aim here is to focus on developing a logic inside the “Controller agent” box which will be able to reason about the necessary adjustments of transcoder installed inside the Server node, in order to respond to temporal variability of a wireless network.

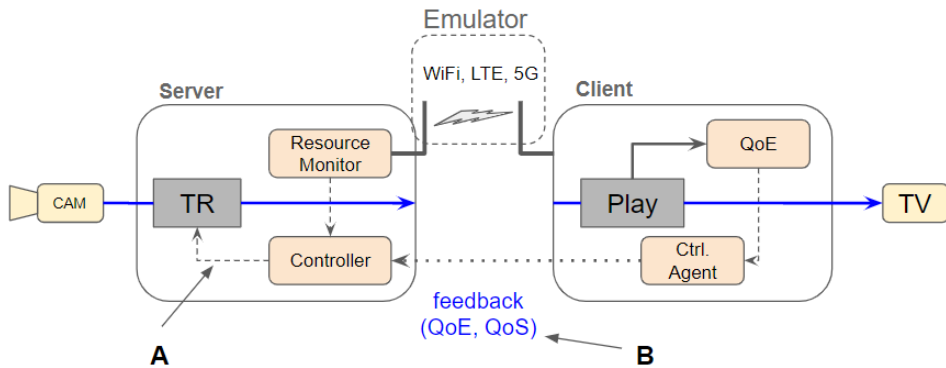


Figure 55 System model of adaptive video delivery in uplink [own source]

Figure 55 depicts an architecture diagram of a wireless system representative for a security scenario, where video is delivered in an uplink direction. This scenario is considered in the thesis due to growing importance of user-generated video. It considers scenarios where individuals are producing content for leisure as well as the relevant public services which can be delivering video in uplink in relation to emergency situations, public surveillance, mass events and others.

Such scenarios require real-time and adaptive delivery of video streams. The streams that are produced are subject to disruption caused by mobility or harsh environment where e.g. access points are not populated densely enough. Relevant

environment could be a rural area or location where coverage is to some extent limited. The Emulator box represents a 4G/5G network for which traffic traces has been captured in advance in order to allow reliable testing of different network scenarios for video streaming purposes. Point “B” in the figure represents the decision logic at the receiver that enables capturing the channel information together with quality indicators in order to provide evaluation of the video stream quality and the needs for adjustments on the side of Server. Inside the Server the controller algorithm assumes combining the two key enablers for congestion based video control:

- **Local feedback “A” loop** – this feedback is enabled by shaper adjustment (and potentially transcoder adjustment) based on the local view on radio resources (wireless channel) provided by the “Resource monitor” component which is symbolically depicted by the “A” reference point. In contrast to “remote feedback” this mechanism focuses on the cross-layer approach to optimizing parameters of TCP protocol which could be used on the transmitter side in order to adjust the traffic based on smart TCP optimizations which are properly adjusted to channel capacity. For shaping options at the Video Streaming source (Node#1), please consult figures (Figure 57 and Figure 58)
- **Remote feedback “B”** – this feedback builds on the measurements performed at the far end – i.e. the “Controller agent” - which is then delivered to local controller for transcoder adjustments. This feedback should be use-case driven.

The decision made in reference point “B” is passed to the transcoder (TR) to adjust stream delivery parameters in order to maintain or increase required QoE statistics in response to the channel dynamics or mobility:

- *Bit rate of a video*
- *Frames Per Second (FPS)*
- *Screen resolution.*

Having in mind all the assumptions defined in section 1, results of real life measurements reported in section 5.5 as well as the architectural approach to deal with “quality feedback” - below author presents a prototype for designing and evaluating the “congestion control” loop for security scenario (i.e. in the uplink direction).

5.3 CONGESTION CONTROL ALGORITHMS

This section briefly elaborates on the available congestion control algorithms, which were identified in Chapter 2 as most plausible to address optimizations maximizing video quality in the client side (e.g. crisis management control room). These algorithms were meant to be compared through a series of tests in order to specify the most promising ones. The solutions which were considered as the most promising include:

- *Adaptive Polling Service (aPS)* – this algorithm has been originally considered as most interesting but actually its value has been declined after discussions with end-users about relation to actual business scenario.
- *TCP based transmission (with built-in congestion control)* – analyses performed in chapter 2 have shown that the most developed multipath protocol with a very active community of developers is the MP-TCP [282]. It has recently become default part of popular Linux distribution. This solution is however considered relevant for a security scenario architecture. It improves reliability by applying redundancy (two parallel transmission paths or more) and also improves performance, but level of improvement depends in the appropriate path manager and scheduler configurations of the mechanism. The simplified version is the use of TCP video delivery (e.g. based on RTMP protocol), where TCP senses the channel by its built-in congestion control detection and simple mitigation.
- *Smoothing buffer for video transmission* - this solution [115] focuses on the optimal control of traffic variability in the downlink direction of LTE radio access by the use of smoothing buffer. Authors provide algorithms and analytical solution to find optimal state switching strategy. Control of the smoothing buffer should be linked to the channel state variability.
- *Own prototype congestion control agent (MCATS)* - author has designed and developed a QoE based controller to decide the video transcoding based on the instantaneous channel dynamics. It has been combined with reading of channel statistics but also the remote feedback from QoE Probe located in the control room (i.e. at Client side in Figure 55). This controller was successfully developed and integrated into a full-fledged prototype with video streamed from a camera in the field towards an emergency control room.

Given the list of most plausible algorithms picked-up above we will focus on evaluations including the aPS and approach the implementation of the MCATS controller and some variant of the TCP on the Server side in the further work. More detailed analysis of the way to approach simulating above solutions is given in the following sub-sections.

5.4 DISCUSSION OF THE ALGORITHM CHOICES

Below the most plausible choices of algorithms for congestion control relevant for this chapter are shortly discussed. Their role will be on one hand to enhance the reliability of the emulator (e.g. by including certain overhead calculations in the emulator), but on the other hand some will serve to validate the emulator in section 5.10.

5.4.1 Adaptive polling service (aPS)

5.4.1.1 Utility for the thesis

Main rationale behind considering this algorithm is for a video stream which is assigned a rtPS class of service (i.e. the one that reserves certain amount of resources – i.e. OFDM slots for video transmission) and periodically stops sending data while connection is still active from the session management point of view. The aPS algorithm provides the ability to save bandwidth by adjusting polling interval dynamically based on user's activity. Thus if a user was not active for a period of time, his polling interval increases in order to reduce the overhead. The overall formula for this is presented below:

$$Overhead(t^{inactive})(i) = Overhead^{initial}(t)/T_n \quad (5-1)$$

Where T_n determines the time interval in which requests are sent. In this case, the overhead for different inactive terminals may be different and is dependent on inactivity time ($t^{inactive}$), where overhead is decreasing - the longer terminal is inactive. T_n is defined by the following formula:

$$T_n = \begin{cases} T_{min} & n = 1, 2, 3, \dots \\ 2^{n-N} \times T_{min} & n = N + 1, \dots, N + M \\ T_{max} = 2^M \times T_{min} & n > N + M \end{cases} \quad (5-2)$$

This algorithm that is mainly reasonable for networks containing multiple users sending limited data, or do not send any data at all, but are connected to the network. Therefore for scenarios with dynamically changing number of users or dynamically changing behaviour of users it can be valuable to adjust polling interval dynamically in order to save important resources and optimise overhead BW consumption. However, as this chapter focuses mainly on security scenarios, in which it is important to stream video constantly without any breaks or frequent changes. Thus as the data is being sent constantly over the time with presumed throughput, and there won't be any major changes in terms of bandwidth consumption of users within a network, the aPS algorithm might not be always suitable for such scenarios. But it is definitely an interesting approach for improving bandwidth utilisation and thus may be further improved and examined (especially considering the green network targets).

5.4.1.2 An approach to emulation

Although it is difficult to implement and test the aPS algorithm in real environment, it is possible to carry out appropriate tests within the proposed emulator described later in this chapter. The proposed element contains scheduler that is able to imitate real network behaviour. A dedicated, but simplified overhead calculation mechanism was implemented, which subtracts fixed number of symbols every frame for a given user within scheduler. It is important to mention that the overhead mechanism is simplified to easily mimic its behaviour when executing various CC algorithms, therefore the effectiveness of such algorithms may not indicate the most accurate results. However, by adjusting overhead consumption based on rtPS user's variability and intensity in

terms of bandwidth utilisation, the emulator tool allows testing and analysis indirectly any algorithms of this type. Calculations considering the aPS mechanism are presented in the section 5.5.3.

5.4.2 Prototype controller design (MCATS)

5.4.2.1 Relevance for the thesis

The main purpose of this controller module is monitoring the 4G/5G modem and reacting in real time to changes in radio parameters (modulations, RSSI, SINR). Depending on an implemented algorithm detected changes can result in immediate changes in video transcoder parameters. The communication between controller and transcoder is used to set transmission parameters or check current status of the radio metrics. Additionally, it is possible to integrate the controller with the QoE monitor and receive quality measurements from the Client node, therefore possible to use the QoE parameters in processing of the algorithm. The block diagram of the controller and interfacing modules is presented in Figure 56.

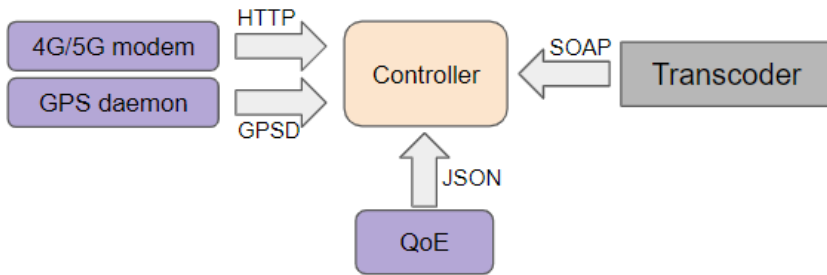


Figure 56 Block diagram of the controller and interfacing modules

5.4.2.2 An approach to emulation

The MCATS is useful mainly in two cases: when (a) preparing an automated (batch) measurement of multiple versions of video streams or (b) designing the feedback loop when real video is being delivered (in transcoded version) and the designer is creating or updating the Controller Agent internal logic prior to its final compilation and deployment. The Python script used there to execute controls is a handful way to work in the interactive mode (trial and error) when delivering optimized algorithms. Furthermore, it is possible to equip the controller with the module reading GPS coordinates from a mobile device and deploy them for predicting and adapting the transcoder parameters based on location.

5.4.3 Channel based traffic policing

As already indicated in chapter 2 [151] - in order to achieve better performance and better accuracy of synchronizing the “channel behaviour” (inside emulator -

in case of real wireless channel) with the “offered rate” of the video stream coming from transcoder, it is required to apply additional processing stage on the side of transcoder – i.e. shaping. Without such element the target congestion control mechanism may become unstable due to hysteresis which is built into the overall variation of delay between quality drop (i.e. identifying such drop) and eventually reacting to such deterioration at the transcoder side. Before the traffic stream will be disturbed by emulator it will be processed in a shaper which will adjust outgoing video rate according to network conditions. The shaper will be controlled by the same set of information about network channel condition as the emulator. The main goal is to relieve emulator from intense and time-consuming processing. The three options considered for implementing the shaper are indicated below:

- Option 1: Introducing additional level for Traffic Control processing on the Server side (Node#1 in Figure 57). The TC architecture will be modified, and the outgoing traffic will be first processed in nodes responsible for shaping where the rate will be roughly adjusted to requested level, and then it will be directed to the Emulator. The potential limitation may be related to TC specifics and its possibility of defining multi-level processing.
- Option 2: Using an external tool installed on the emulator host or on the Server (Node#1). The Emulator host will be equipped with dedicated application that queues incoming traffic and next send them to emulator with desired rate. The other possibility is to perform shaping early on Node#1 - the outgoing traffic rate could be manipulated by the TC tools⁴ in the same way as in emulator.
- Option 3: Using cross-layer approach to TCP protocol. Creating an application that basing on information about network condition could directly manipulate parameters of TCP protocol, especially its timers to adjust it to the channel behaviour.

In order to better visualise the exact architecture of the above mentioned solutions the figures below (Figure 57, Figure 58) present the role of each component, necessary to achieve required functionality. Option 1 can actually be also superimposed with Option 3 to provide hybrid approach but only considering the additional elements of “Cross-layer” and “Resource monitor”.

Option 2 has not been elaborated above as in practice it would mean that such mechanisms would have to be developed inside the node emulating the channel. From obvious reasons (integration perspective) this is technically possible but not rational to have relevant SW components injected into “the channel”. **In order to be able to deal with design, testing and improvement of above mentioned, congestion control solutions, as well as new ones yet to come - it is necessary that an emulator tool of 4G/5G network/channel be provided.** It should be

⁴ Traffic Control (TC) is a popular framework available for Linux based systems that deals with processing of traffic streams

able to use traces from real network measurements in conjunction with 4G/5G signalling traces gathered from a simulator (e.g. ns2/ns3 or 5G Vienna). The latter expectation is important due to limited access to physical layer measurements. There is practically no possibility of tracking a real-time TDMA frame utilisation, neither availability of actual overhead introduced by the RAN signalization mechanisms.

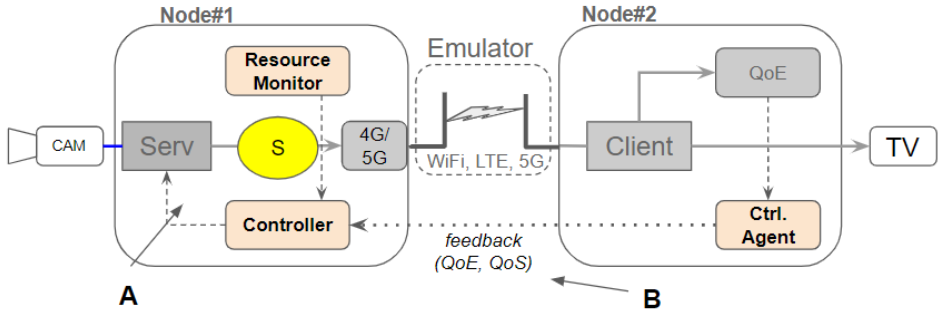


Figure 57 Traffic shaping Option1

Owing to such approach this work enables emulating dynamics of instantaneous channel (and thus bandwidth) that is used by nodes (cameras, traffic generators etc.) in uplink direction. In this approach congestion control mechanisms' behaviour will be based on scenario description (test script) and on this basis overhead for each traffic will be adjusted in order to increase actual bandwidth usage.

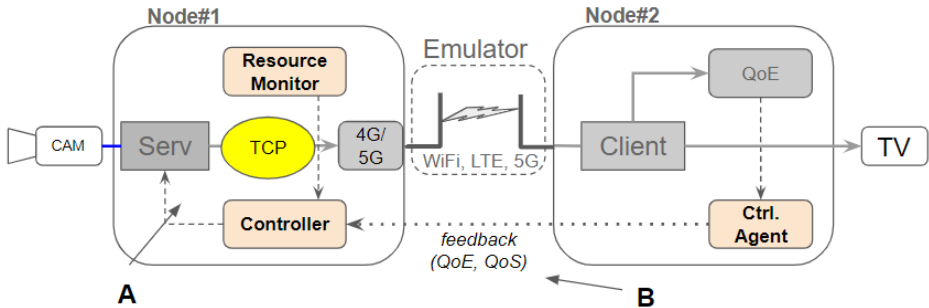


Figure 58 Traffic shaping Option3

5.5 REAL NETWORK MEASUREMENTS – RATIONALE FOR CONGESTION CONTROL

5.5.1 Mobile measurements – first round

During the course of preparing this thesis, multiple tests were performed with different 4G/5G network environments and different radio conditions. Results

achieved so far have been reported in two papers [42], [51]. This approach proved that, because of dynamically changing nature of radio in wireless networks (4G, 5G), the real-life measurement observed may vary significantly even when conditions seem to change only slightly. Table 31 presents results of baseline testing of uplink traffic transmission in two stationary locations. The tests were performed inside network of a polish local 4G operator. Two different modems were used (Teltonika, Greenpacket), with two data rates generated from user terminal towards AP – 128Kb/s and 1024Kb/s. Both locations are in either close proximity with the AP or there is LOS between user and AP. It can be seen that in the tests packet loss ratio is always reasonably lower than 1%, which means very good conditions. The additional proof of the baseline quality is that almost 100% of time the least robust modulations are used (QAM-64).

In the next step the more challenging radio conditions were exercised by driving in a car over the streets of Choszczno Poland, and performing traffic measurements in the uplink. The following traffic characteristics (Table 32) were collected while driving along four streets in a circle (at a speed of 20 km/h). It is evident that when mobility and non-homogeneous coverage are combined together the received traffic is highly degraded (20% of original stream reaches the receiver). The traffic rate sent from the transmitter was 1 Mb/s (CBR) and it was sent to the node representing security operator premises (PC2), located just at the WAN interface of the operator. The case above represents a situation where there were both degradation causes present in parallel: intense NLOS and mobility. That is why the throughput level is largely degraded throughout the plot.

Table 31 Stationary tests in Choszczno, Poland

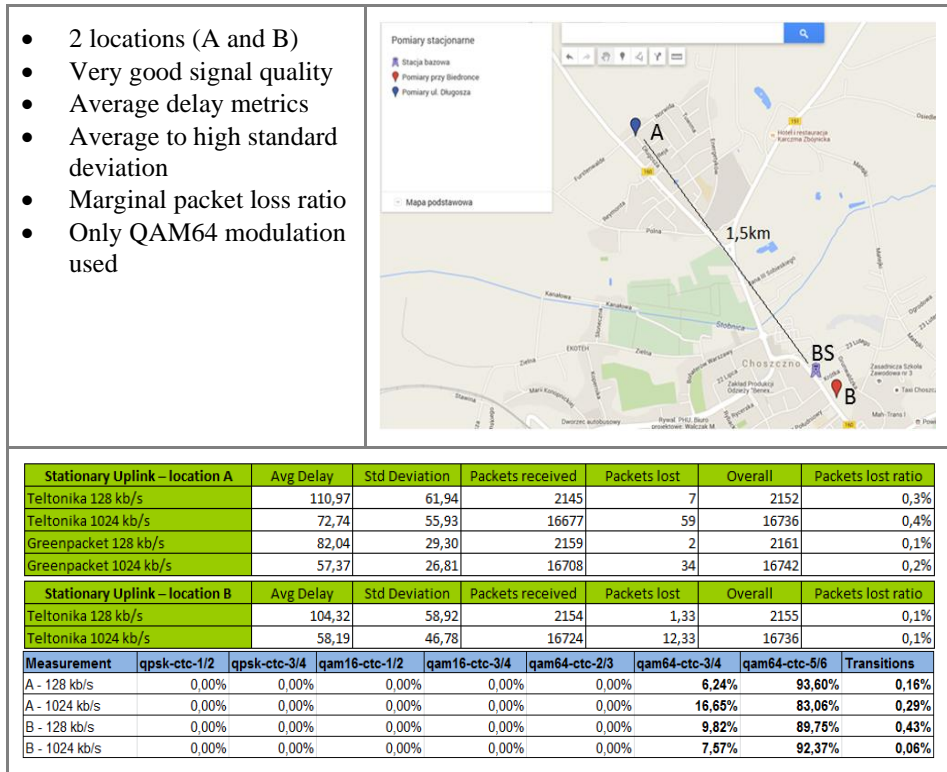
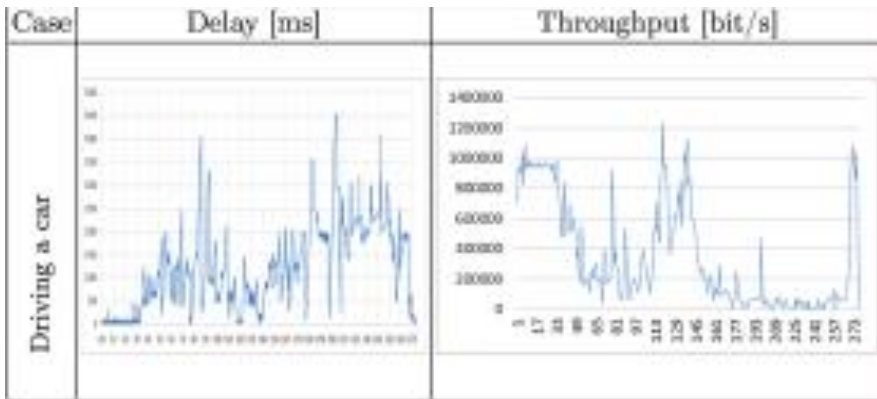


Figure 59 Drive test (Choszczno mobile) - modulations

The plots in Figure 60 show the influence of: a) SISO/MIMO antennas and b) the delay and RSSI metric in the uplink direction. The traffic was transmitted in uplink with two alternative speeds, namely: 128 and 1024 Kbit/s. Moreover, the drive tests were performed in various locations however it is worth highlighting the observed mismatch between fidelity of coverage modelling with popular software package Splat! [267] and utilising the in-house developed signal analyser (RaspberryPi based) with GPS coordinates readings (Figure 59).

Table 32 Sample network traces from drive tests (driving a car)



It is worth noticing that the circular areas of equal signal coverage (the right part of figure) highlights incompatibility of the instantaneous modulation readings from the same area. Results of above-mentioned tests are reported in more detail in the paper [51]. In Figure 61 where the instantaneous throughput change in uplink is presented while driving across the streets in around 1,5km from the serving AP. The numbering of streets (1-2-3-4) is also mapped to the modulation plot to indicate the sequential chain of modulation changes. Moreover from the perspective of channel modelling the bottom table indicates the distribution of modulations for the 4 cases: a) 15km/h with 128 kbps uplink transmission, b) 15km/h with 1024 kbps uplink transmission, c) 30km/h with 128 kbps uplink transmission and d) 30km/h with 1024 kbps uplink transmission.

It can be seen that it is the requested throughput that drives the usage of particular MCS mode and not so much the speed in the range 15-30km/h. Similarly, the tests presented in the Figure 63 of the next section show that the speed of node (10-30km/h) does not have significant influence on the achieved delay values in uplink. It can be seen that it is the requested throughput that drives the usage of particular MCS mode and not so much the speed in the range 15-30km/h. Similarly, the tests presented in the Figure 63 of the next section show that the speed of node (10-30km/h) does not have significant influence on the achieved delay values in uplink.

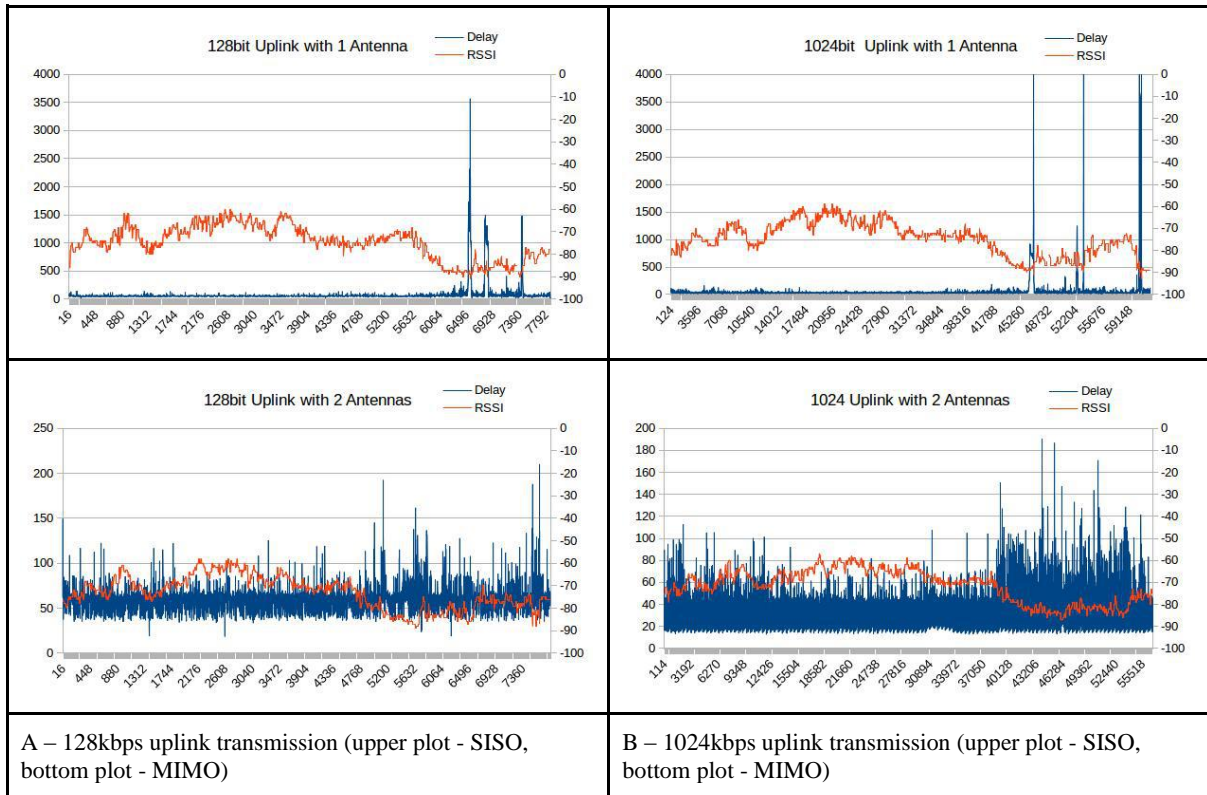


Figure 60 The delay and RSSI of the uplink transmission for different data rates

5.5.2 Mobile measurements – second round

Additional tests performed by author in the preparatory campaign was to assess the connection quality (QoS) while end-user terminal is moving. Traffic flow characteristics used was exactly the same as presented in Table 32. Still the difference as compared to the laboratory (i.e. fixed location) tests was the use of miniPCI version of 4G modem (Teltonika). The modem was used with default settings and with an external antenna, providing additional gain of 5dB and equipped with a fitting magnet for rooftop mounting. Drive tests were performed to identify the geographical locations where the nominal signal quality was at least acceptable. The spectrum analyser Tektronix SA2500 has been used to identify regions with acceptable signal quality. Two locations have been selected in the range of 1km from the base station. The first location (A) was a parking lot in a close vicinity of the base station and the second (B) was the street among blocks.

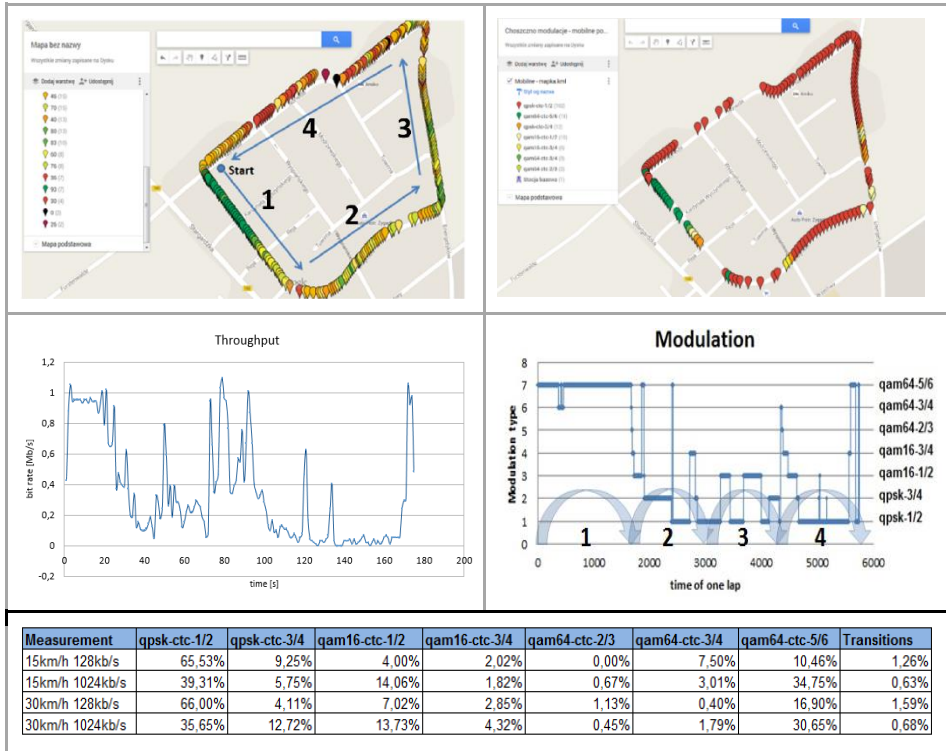


Figure 61 Summary of mobile measurements in Choszczno, Poland

It can be seen from the Figure 63 that with the increase in throughput the OWD decreases. The average delay (OWD) for 1024Kbps equals ca. 30ms and is two times smaller than the delay of the 128Kbps flow (i.e. eight times smaller rate). The same behaviour is observed for both locations and is rather independent within given speed range. The 10ms difference in average delay for lowest throughput (the “Street” location) may result from slight variety in radio conditions (fading) during car movements. It can also be seen from the plots that standard deviation bars are quite high and equal ca. 15ms in average case (i.e. between rates and speeds).

Most probably the reason for such difference is the overhead incurred in transmission of small rates. In case of 128Kbps there is 12 packets sent per second, which means that each packet is using its own dedicated TDD time frame (and thus the additional delay is added). While in case of 1024Kbps there are 93 packets sent per second which means that at least two packets can be transmitted using single TDD frame (as each packet is sent every 10ms). A clear trend can be identified that with the decrease in traffic rate (from 1024 to 128 Kbps) the OWD increases (from ca. 40ms to ca. 60ms).



Legend:
 “Location A” – parking lot (on the left)
 “Location B” – street between houses (on the right)
 “BS” – location of base station (eclipse at the bottom of the map)

Figure 62 Measurement locations

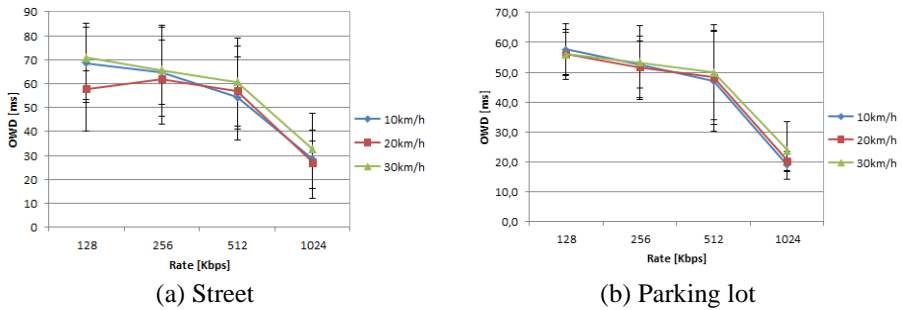
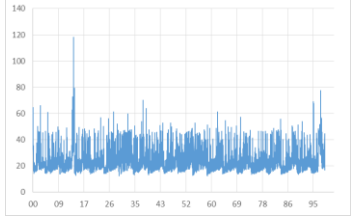
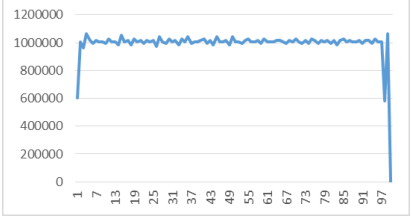

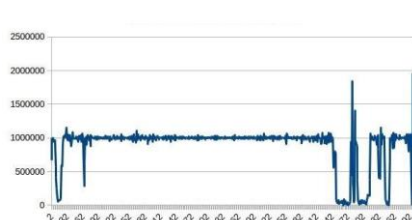
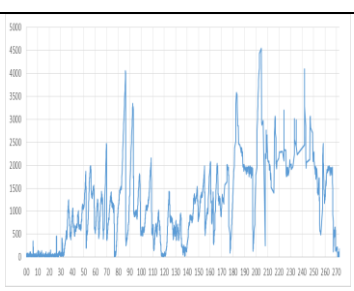
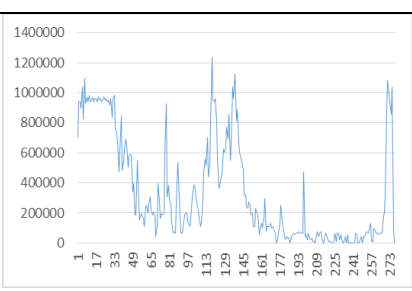


Figure 63 Overview of mobile results

The value of delay measured using ICMP traffic (PING) would actually need to be divided by two (as this is round trip delay and not OWD) and it has to be noted that it was made for a small packets of 32Bytes. It has to be underlined that the measurements were repeated with 4G modem of another vendor (GreenPacket DX350) and the trend of increasing delay with smaller traffic rates has also been observed. The same behaviour of OWD characteristics has been confirmed in numerous mobile tests (see section 5). It has to be noted that average delay measured in downlink direction is on the opposite - stable independent on the rates/speeds and equals ca. 20ms. This way it seems that there is either a delay introduced by sleep modes of the 4G modem or there is some delay due to TDD scheduling at the AP. Based on the observations performed, example network extremes has been classified into three types: Good, Medium and Extreme. This approach is also mentioned in the previous section where traces are discussed. Table 33 shows extreme (representative) network behaviours mentioned above with results attached. In all cases presented the video traffic sent from the transmitting CPE was 1Mb/s and it was send to the node located nearby of the

operator’s base station (i.e. right after the AP).

Table 33 Sample of extreme network characteristics from the field tests performed

Behaviour type	Delay	Throughput
Good		
Medium		
Extreme		

As can be seen from Table 33, so called “good” network conditions provide stable and high quality connection. It does not seem to require any adaptations for congestion control, it does seem necessary to develop specific mechanisms to support or adjust to extreme network traffic transmission. This problem applies especially to real-time (uplink) video traffic transmission, because of the fact that video streaming is particularly vulnerable for both packet losses and bandwidth limitations.

5.5.3 Calculating user-plane throughput with overheads

By utilising traces gathered from real life measurements where all MCS (Modulation Coding Scheme) were known, the scheduler developed for the emulator is able to calculate throughput according to these statistics. Therefore, in order to calculate available bandwidth for specific user efficiently, the formula

below needs to be implemented into the scheduler model.

$$tput(x)(t) = \left[OFDM - overhead(t) - \sum [OFDM(OtherSS)(T)] \right] * MCS(x)(t) * frames/sec * subcarriers \quad (5-3)$$

As it is important to clarify all basic variables which are taken into calculations, table below provides the explanation of all of the values presented in the aforementioned formula.

Table 34 Explanation of terms

Symbol	Description
t	time
x	Terminal identifier
tput (x) (t)	Uplink throughput of SS station of (x) ID in moment (t) [bps]
overhead^{uplink} (t)	Percentage of symbols that are used for ACK, initial ranging etc. For our scenarios the size of overhead varies between different QoS classes: <ul style="list-style-type: none"> • Sending rtPS: approximately 3,2 symbols per frame • Active rtPS: approximately 1,6 symbols per frame • BE user: approximately 0,3 symbol per frame Those values represent results that are the most probable and close to real network overhead usage. The values were averaged based on traces gathered from separately generated simulations and according literature ^{5,6} . Such approach allows for further adjustments and testing scenarios which include aPS or other congestion control algorithms.
OFDM	Number of all OFDM symbols assigned to Uplink part of TDD frame
SUM [OFDM (other SS) (t)]]	Number of symbols scheduled for another terminals (SS) in moment (t) [not counted to Overhead]
MCS(x)(t)	Number of bits per symbol resulting from modulation coding scheme of terminal (x) in moment (t) gathered from each packet
frames/sec	Number of OFDM frames in one second
subcarriers	The overall number of subcarriers that are used for data transmission on a standard 4G network

Methodology of calculating throughput considered for the emulation is (bps):

$$tput^{uplink}(x)(t) =$$

⁵ http://www.cse.wustl.edu/~jain/books/ftp/wimax_ra.pdf

⁶ <https://pdfs.semanticscholar.org/cfd8/5a0f54bf98652f99bd9ea703e82de76f08e4.pdf>

$$\left[OFDM^{uplink} - overhead^{uplink}(t) \sum [OFDM(OtherSS)(t)] \right] * MODULATION(x)(t) * frames/sec \quad (5-4)$$

Calculations of mentioned values:

$OFDM^{uplink} = [OFDM^{total} - OFDM^{downlink}]$ – to calculate or gather system parameters (Fixed attribute for concrete 4G system and DL/UL ratio)

$frames/sec = 1/(Duration_Time_Of_TDD_frame)$ – for 5ms.
we have $1/0.005 = 200 frames/s$

$\sum [OFDM(OtherSS)]$ – for scenario nr 1 (only 1 station is sending data) value is equal to 0 (for data) and is fixed (CONSTANS)

$$Overhead^{uplink} = (initial_{ranging}) + (contetion_{BW}) + rtPS_{BWreq}(n^{inactive_{rtps}}(t)) \quad (5-5)$$

- For $rtPS - rtPS_{BWreq}(n^{inactive_{rtps}}(t)) = n^{inactive_{rtps}}(t) *$

$overhead^{single_{rtps}}$

- For $aPS - rtPS_{BWreq}(n^{inactive_{rtps}}(t)) = \sum^{inactive_{rtps}} (overhead(t^{inactive})(i))$

Thus if a user would like to adjust the overall available bandwidth for selected traffic, it can be done by manipulating the overall number of OFDM symbols while ignoring overhead and OFDM (OtherSS) values. For example, for 11 OFDM symbols, the overall bandwidth will be equal to 2,112Mb/s, and the formula will look like the one below:

$$tput(x)(t) = OFDM * MCS * frames/sec * subcarriers \quad (5-6)$$

$$tput = 11 * 5 * 200 * 192 = 2,112 Mb/s \quad (5-7)$$

The remaining work aimed to define sample video controller (see section 5.9) with the following features: i) traffic shaping - introduced by properly adjusting transcoder rate based on information from a 4G/5G modem at the transmitter (e.g. car), ii) overhead injection - the implemented scheduler subtracts fixed number of symbols every frame for a given user to mimic various congestion control algorithms. The implemented scheduler delivers treatment compliant with the rtPS 4G/5G class of service with its crucial parameters (maximum sustained traffic rate and minimum reserved traffic rate) - as it needs to be considered when

dealing with congestion control mechanisms. However, before being able to design the emulator framework, and perform validating tests, author will now define the statistical framework that will be able to mimic the channel variation and generate the packet losses and delays which are statistically equivalent to the above mentioned measurements.

5.6 STATISTICAL ANALYSIS OF DELAYS AND LOSSES FROM FIELD TESTS

Statistical research was carried out based on the measurement results presented in previous sections. In particular, statistical tests were carried out regarding the inclusion of the HARQ mechanism and the impact of throughput on delays and losses.

- Impact of the HARQ mechanism - the study concerns the frequency of data loss and the size of delays in both scenarios. The analysis of the dependence of the HARQ packet and the occurrence of losses was carried out using the chi-square test of independence (χ^2) and the test for two proportions. The research on delays was carried out using the Mann-Whitney U test and the Student's t-test, and an analysis of the frequency and parameters was carried out using the methods of descriptive statistics.
- Impact of throughput on delays and packet losses - the study concerns the amount of delays in both scenarios and the frequency of data loss. The study in the field of delays was carried out using the Mann-Whitney U test and the Student's t-test, and an analysis of the frequency and parameters was carried out using the methods of descriptive statistics. The analysis of the relationship between the packet size and the occurrence of losses was carried out using the chi-square test of independence (χ^2).

5.6.1 Generic flow of the estimation

The simulation should take place in two steps:

- Latency simulation - the simulation is performed in several steps. First, we determine the first observation (value) and the initial phase (phase i.e. decrease or increase). Then we randomize the length of the phase. After drawing the length of the phase, we determine the amount of delay (based on the change in the previous delay) - in accordance with the previously established distribution. Then we add random noise.
- Packet loss simulation - the simulation is performed in two steps and is based on delay values. The first step is to make a binary decision as to whether a loss will occur in a given simulation (this step is based on the lag values). The second step is to determine the value of the loss - i.e. the

number of packets lost. This step applies if and only if a decision was made in the first step that packet loss has occurred.

5.6.2 General flow of simulation

The entire procedure has been presented in block diagram on the Figure 69.

5.6.2.1 Simulating delays

Delay simulation will be performed in 4 steps:

- Determination of the first element and phase
- Determination of the phase length
- Determining the size of delays (based on the size of the change)
- Adding random noise.

The analysis of observation data was preceded by smoothing (the moving average method). Thanks to this, it was possible to isolate the essential cycles from the noise. However, in order for the simulated observations to be similar in nature to the real ones, noise must be added later.

5.6.2.2 Model validation

Model validation is carried out separately for each scenario (loss simulation, delay simulation), its implementation is aimed at comparing the distribution of field measurement and simulation results. The general course of the simulation is shown in the Figure 69. The flow seems simple enough, but the re-adjustment and compliance loop can be repeated over and over again. All validation steps are described below.

Step1 - Preparation of measurement data

Prepare a measurement data package. Measurements should be carried out under different conditions. The more data, the better.

Step2 - Adoption of original parameters

The parameters for the first trial can be taken from another, similar distribution and modified to best fit our scenario. For this purpose, it is best to prepare a comparative frequency chart of both distributions and adjust the appropriate parameters. Instructions on how to make a comparative frequency plot can be found in the appendix. Examples of such charts are provided in the "frequency" tabs in the validation sheets. Ways to modify parameters, depending on the symptoms, are described in point - "5) Distribution parameter tuning".

Step3 - Running a simulation

After entering the parameters, we run a simulation, the number of observations generated should be similar (at least in terms of scale) to the number of simulations from the measurement. For example, if we have 100,000 measurement observations, then 45,000 simulations can be prepared (preferably in several series, e.g. in three series of 15,000 each).

Step4 - Checking the consistency of distributions using statistical methods

(decision)

The statistical methods that are used for the study deal with frequencies and distributions – they abstract from the order of measurements. This means that we check whether the measurement and simulation are of the same nature (e.g. they could be carried out in the same environment, conditions), and not whether the course is similar. For example, if there are two short-term disturbances during the measurement involving increased delays and resulting from walking behind buildings or other objects, the number of such incidents in the simulation may be different. The check is carried out by:

- a. Analysis of frequency plots
- b. Parameter comparison (mean, standard deviations, median, 1st quartile, 3rd quartile)
- c. Performing a parametric t-Student test for independent samples
- d. Performing a non-parametric U-Mann-Whitney test (since this test is lengthy, it is best to do it only for those simulations that perform well in other methods)

Recognition of the simulation as a good fit completes the validation process. Then, the assumed model and parameter values are assumed.

Step5 - Distribution parameter tuning

The most difficult step in the validation process is modifying the parameters. After determining that the measurements and simulations are not from the same population (i.e. they do not match), the parameters are modified. Below are some ways to modify the parameters for different symptoms of distribution mismatch:

- too many extremes on one side (too many highs, not enough lows):
 - reducing the level of “ramp-ups”,
 - increasing the probability of decreases in relation to increases,
 - shortening the phases (phase lengths),
- too "calm" simulation:
 - increase the level of growth,
 - increasing the probability of increases in relation to decreases,
 - phase extensions (phase lengths),
- too high values:
 - setting a lower limit (max)
- too thick tail (see Figure 66):
 - adding a second reflection

Step6 - End of validation

The validation process ends when simulations and measurements are considered to be a good fit. Then, the assumed model (size of parameters) is assumed. The detailed flow of parameter tuning has been presented in the Figure 67.

5.6.2.3 Consistency checking activities (Step4)

In this section the four steps of consistency checking performed in step4 in previous section is presented in separate sub-section below.

Analysis of frequency plots

The analysis is basically done manually and the general idea is that the distributions should be similar to each other. Below are some examples of distributions along with their evaluation.

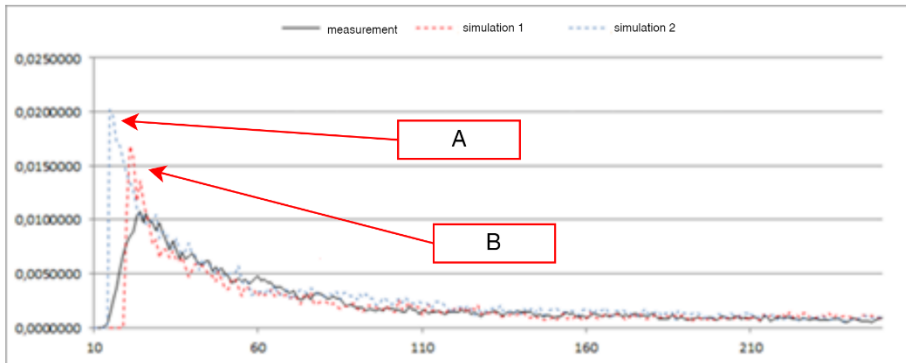


Figure 64 Example – mismatched distributions

In the Figure 64 the two pointed plots represent the distributions that were not providing satisfactory of the model achieved, based on the provided data. The point “B” indicates sample distribution settings where the distribution is not well matched with the actual plot. Whereas the “A” distribution is even worse due to prevailing large number of small values.

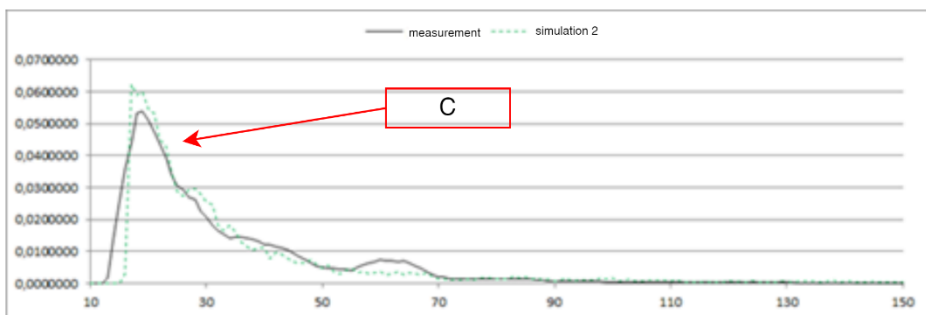


Figure 65 Example – well matched distribution

In the above example (Figure 65) there is very high matching between the original data and the value modelled by the distribution “C”.

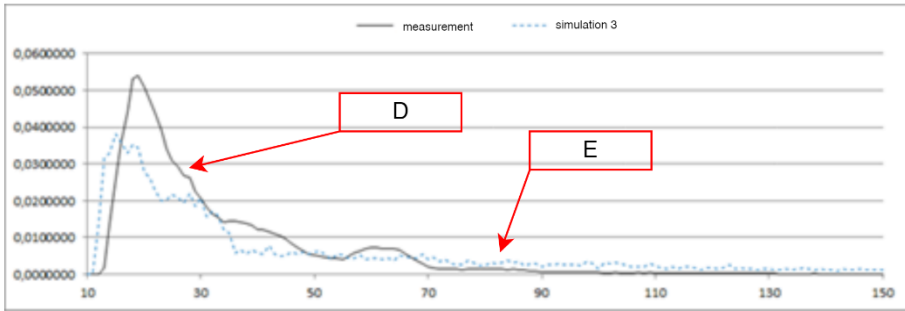


Figure 66 Example – mismatched distribution

As it can be seen (Figure 64) the simulated distribution has too low number of observations at the low values (point “D”) but also the “E” label indicates the problem of too “thick” tail, which is not suitable for the model.

Parameter comparison (mean, standard deviations, median, 1st quartile, 3rd quartile)

The most important parameters in this case are: median 1st quartile and 3rd quartile. It can be compared on the chart.

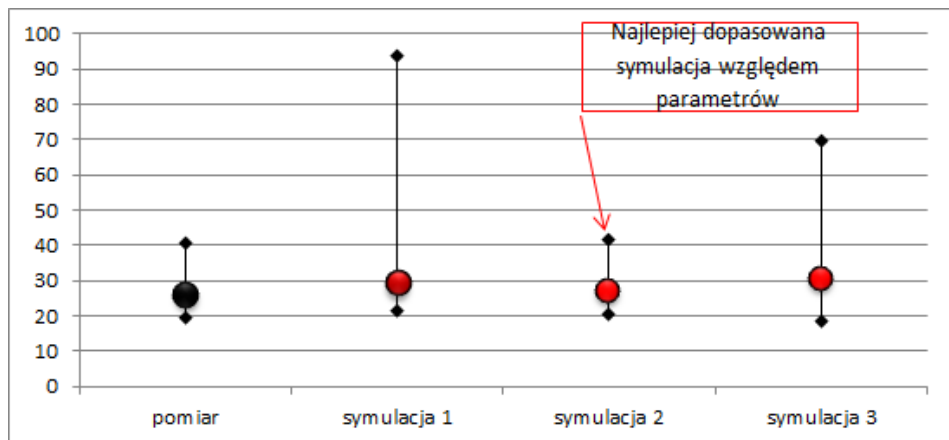


Figure 67 Example – the importance of parameter tuning

Performing a parametric t-Student test for independent samples

The test is carried out in accordance with the formulas, examples of the tests carried out can be found in the "results" tabs of the validation tool developed. The t-Student's test must indicate the agreement of the two samples, otherwise we reject the simulation and modify the parameters again.

Conducting the non-parametric U-Mann-Whitney test

Since the Mann-Whitney U test is time consuming (it cannot be performed automatically), it is best to perform it only for those simulations that perform well in other methods. The UMW test is very strong, which means that it often rejects distributions that are acceptable to us. If the UWM test shows the compatibility of the distributions, then we accept the distribution without any doubts (even if it failed other tests). If it does not agree, then the distribution can be accepted only if the other methods show good agreement and if the value of the test statistic is quite low.

5.6.3 Validation of 4G/5G delays and losses (simulation)

The ultimate result of the delay and loss simulator is provided as the MS Excel file with suitable macros which implement above mentioned distributions, tuned based on the results of field measurements in the 4G networks. The resulting simulation tool can be utilized in order to inject delay and packet loss into the TBONEX emulator. After performing the new field tests it is possible to apply the methodology presented in the current section and tune parameter values, so that they can be than used inside the developed simulator.

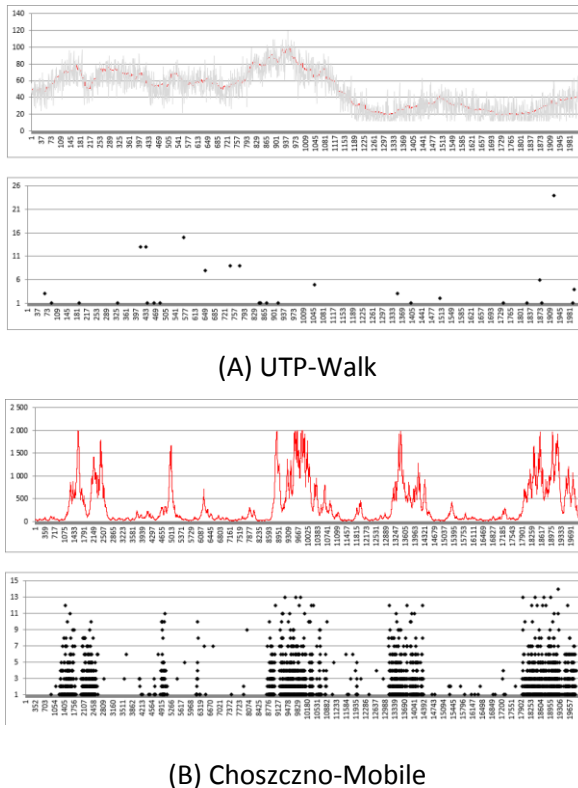


Figure 68 Simulation of delays (top) and losses (bottom) based on a simulation model

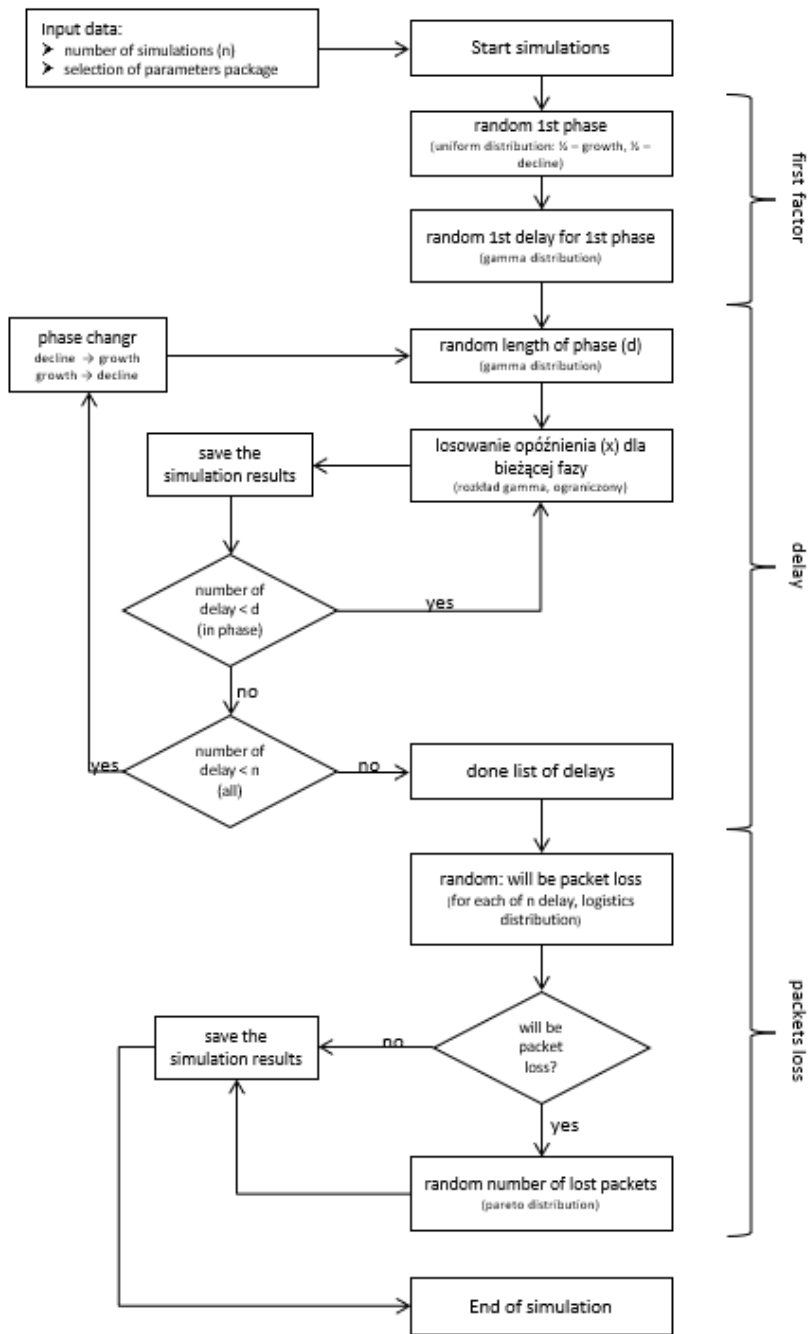


Figure 69 Steps of simulating delays/losses based on the proposed method

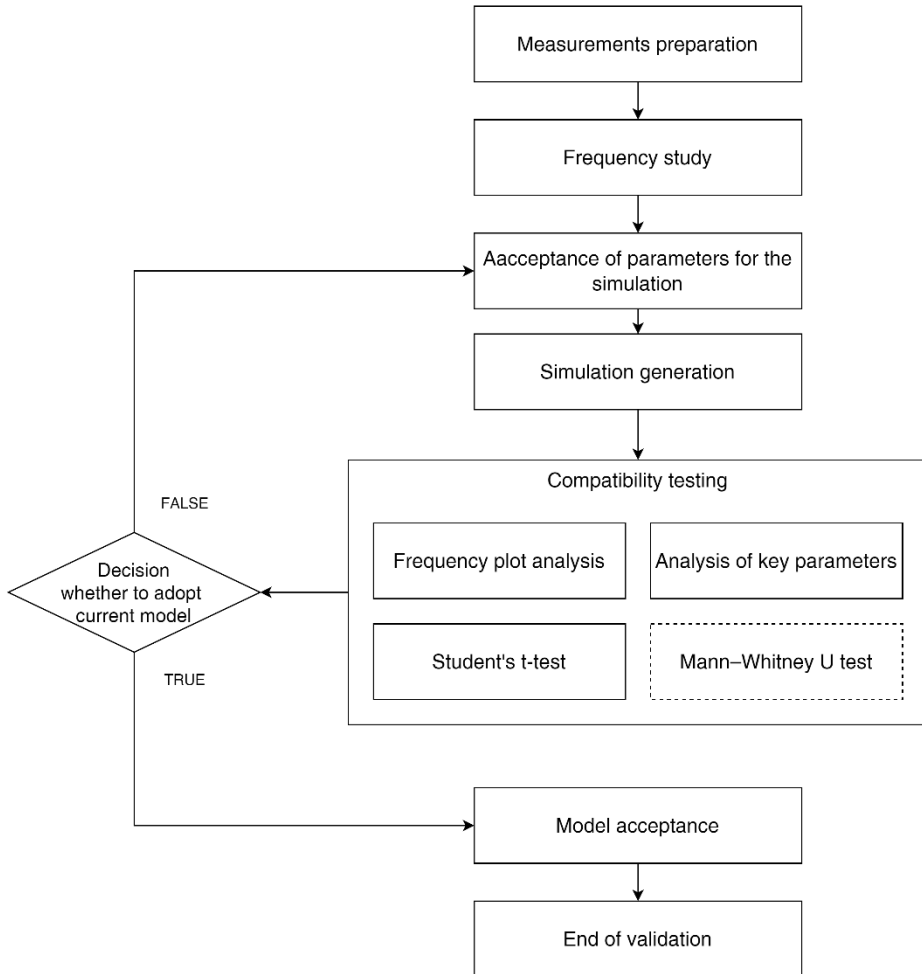


Figure 70 The flow of the statistical model validation process

The example simulation run is presented in Figure 68 and relevant statistics are also displayed in the Table 35. It can be seen that the latency plot is very well resembling the field data presented in the Table 33. Another sample of the simulation is presented in part B of the Figure 68, this time it is using the “Choszczno mobile” field data. The process of determining the exact value of parameters enabling the distribution-based modelling of delay and losses in wireless 4G/5G channel, requires manual adjustment to the conditions (mobile/stationary, high/low variability). Moreover, changing one parameter often involves changing others (this is how it is with these distributions). Therefore, prepared set of parameters has been utilized in the simulator to enable simulations based on these particular channel models. Sometimes a small change in one parameter without adjusting the others will change the results by a scale of

values. Packet loss simulations are based on the trace files "Choszczno-mobile" and "UTP-spacer" cases only. In the other measurements, the number of losses is so small that it was pointless to study it (see Figure 61). It is best to assume that such situations do not occur in measurements at fixed locations, or to assume some distribution with a very low probability of occurrence.

5.7 EMULATION FRAMEWORK DESIGN

The overall, high-level description of the proposed network emulation framework is presented in the figure (Figure 71). The system allows the user to execute customizable emulation scenarios related particularly to video streaming in 4G network. Such scenarios comprise many variables including network properties and conditions as well as various users' activities. The proposed software includes scheduler implementation for real-time traffic class, compliant with tuneable requirements of the 4G class of service (extendable to 5G).

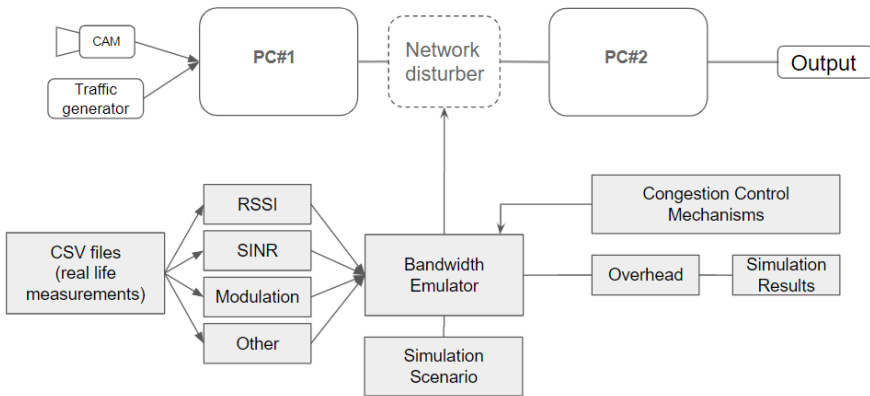


Figure 71 Network emulation framework for 4G/5G

Table 35 Sample statistics for the Figure 70 plots

SIMULATION REPORT		SIMULATION REPORT	
General information		General information	
Time:	14:06:26	Time:	23:22:07
Scenario:	UTP-walk	Scenario:	Choszczno-mobilne
Type (1):	with noise	Type (1):	with noise
Type (2):	with losses	Type (2):	with losses
Number of observations	2012	Number of observations	20 003
Latency (without noise)		Latency (without noise)	
smallest value:	18,91	smallest value:	19,40
biggest value:	98,36	biggest value:	1998,99
Average:	49,66	Average:	339,87

Median:	52,90	Median:	117,02
standard deviation:	20,88	standard deviation:	435,52
Latency (with noise)		Latency (with noise)	
smallest value:	12,3	smallest value:	N/A
biggest value:	119,44	biggest value:	N/A
Average:	47,59	Average:	N/A
standard deviation	23,11	standard deviation	N/A
Avg. noise	-2,07	Avg. noise	N/A
Packets loss		Packets loss	
number of observations with loss	28	number of observations with loss	2401
percentage of observations with a loss	1,39%	percentage of observations with a loss	12,00%
Avg. loss	4,61	Avg. loss	2,86
Max loss	24	Max loss	14
Avg. latency for losses	52,29	Avg. latency for losses	976,29
Avg. latency (with noise)	53,54	Avg. latency (with noise)	N/A

To run it, it is required to specify and customise the overall network properties, which include the following: overall bandwidth available within a network in the uplink direction based on Downlink/Uplink ratio and number of symbols, TDD frame duration, overhead behaviour for rtPS terminals, scheduler's efficiency (variable deciding how efficiently scheduler allocates bits in symbols) and network delay distribution scheme (for efficient delay emulation). The above emulator architecture enables definition of real-time (rtPS) flows, together with its key parameters (GBR, MBR, priority) whether CBR or VBR/adaptive traffic. Based on the configured rtPS parameters the built-in scheduler is able to perform scheduling compliant with the rtPS scheduler. The PC1 plays a role of video source with transcoder in order to mimic the camera mounted inside a car or on a drone (whether front, rear or sideways). On the side of PC1 it is possible to utilize (i) real camera attached via network interface, (ii) pre-recorded video from files or (iii) MGEN traffic generator. The PC2 mimics a traffic receiver "inside the control room" - where e.g. the Uber application employees are triggered to provide remote guidance to a member of the fleet of autonomous cars upon request from the particular car's controller [13]. Technically during any tests and evaluations with the proposed architecture the PC2 requests video stream from the "car node" (PC1) via RTSP request to initiate the transmission. For this purpose, a video client like e.g. ffplay with dedicated shell scripts that initiate

video delivery from PC1 are used at the side of PC2.

The “Network Disturber” block is realized by a Linux software router, and plays the role of:

- Access Point - with scheduling of traffic in the uplink, modulation adjustment (AMC), as well as introducing signalling overhead (rtPS signalling, BE signalling).
- Uplink radio channel - by introducing modulation driven rate adaptations caused by mobility (slow-fading, multi-path), NLOS, packet drops and delays.

This solution uses real values of modulation, rate and delay collected from the trace file or from the simulator presented in section 5.6 of this chapter. The Bandwidth Emulator is an external application that based on traces from drive tests performed by the author in multiple (selected as representative) locations produces two configuration scripts that have to be deployed on the Network disturber. The scripts are used to adjust in real-time the QoS classes within Linux node that acts as a “Network disturber” dynamically configured by an emulator. The proposed solution allows defining all crucial parameters such as OFDM symbols count, frame duration and overhead. For target scenarios different users relate to different cameras. To prepare a test one needs to specify the overall number of cameras in a scenario, provide their operation status updates within given timespans, and their respective class parameters including: priority, GBR and MBR. For each camera, the user also defines the radio quality behaviour by assigning CSV trace from a particular drive/field test in real 4G/5G network. Each transmitter that is sending video footage to the receiver (i.e. each data flow) is experiencing network conditions according to the CSV file assigned for this specific transmitter in the process of preparing tests. In order to provide representative samples of channel variability author has performed drive tests (or walking tests) in different networks. Author originally was considering several approaches regarding emulation of 4G/5G connection, they are summarized in 19. From the approaches considered above, the original contribution of delivering the „TBONEX” emulator facility approach is the most appropriate one. This approach is sort of a middle ground between realistic amount of work needed to implement such mechanism and expected quality (fidelity) and thus validity of final results.

5.7.1 Network disturber details

This section provides information about all the calculations that are done inside scheduler loop. Figure below shows a high-level depiction of TBONEX processes in order to emulate wireless network conditions and then allow streaming video atop the specified network conditions.

At the beginning, the tool requires to define initial scenario. First process is related to configuration of network settings which include specifying network’s overall bandwidth resources, downlink/uplink ratio or frame’s duration. The

second step include configuring all the traffic that will be emulated for specific scenario, which include the number of users, their behaviour and traffic types (but only for artificial traffic).

After setting up network and filling all required scenario options, the third stage begins namely “Video Streaming Emulation”. The whole process starts executing two simultaneous stages in parallel. The first, is the “Video transmission” where also QoE statistics are gathered. The second is the Scenario Execution stage, where all the calculation of network conditions are specified.

The Bandwidth Emulator is an external application that based on traces from custom drive tests produces two configuration scripts that have to be deployed on the Network disturber (Figure 72). The first output script (TC script) contains a set of commands for the traffic control framework installed on the Linux router. It creates hierarchy of traffic control nodes such as qdisc, class and filters, and attaches such structure to the outgoing network interface of the Network Disturber. The NETEM qdisc is used for defining delay and loss for the user’s flow. Additionally, for isolating any other traffic from the simulated transmission the script defines qdisc and filter for all remaining traffic flowing through the interface and passes further it without any disturbing.

Scenario execution is the process where for each frame a dedicated scheduler allocates bandwidth to every user by specified rules and calculations. After initial assignment of all symbols for rtPS users, the scheduler checks their current network conditions and takes away symbols related to overhead (initial ranging etc.). **In order to ensure the ability to implement various CC algorithms/mechanisms, the overhead implementation was simplified into the process that is widely configurable and customizable.** This way it is possible to easily test and analyse algorithms that can optimize overhead such as e.g. the aPS algorithm described in the section 5.4.1.

After subtracting overhead from all available symbols of a user, the rest of effective bandwidth can be processed into further allocation mechanisms. If network conditions in which currently the user is located allow for it, he receives requested bandwidth. Otherwise, the user gets as many bandwidth resources as possible according to current network conditions (e.g. modulations and signal strength values).

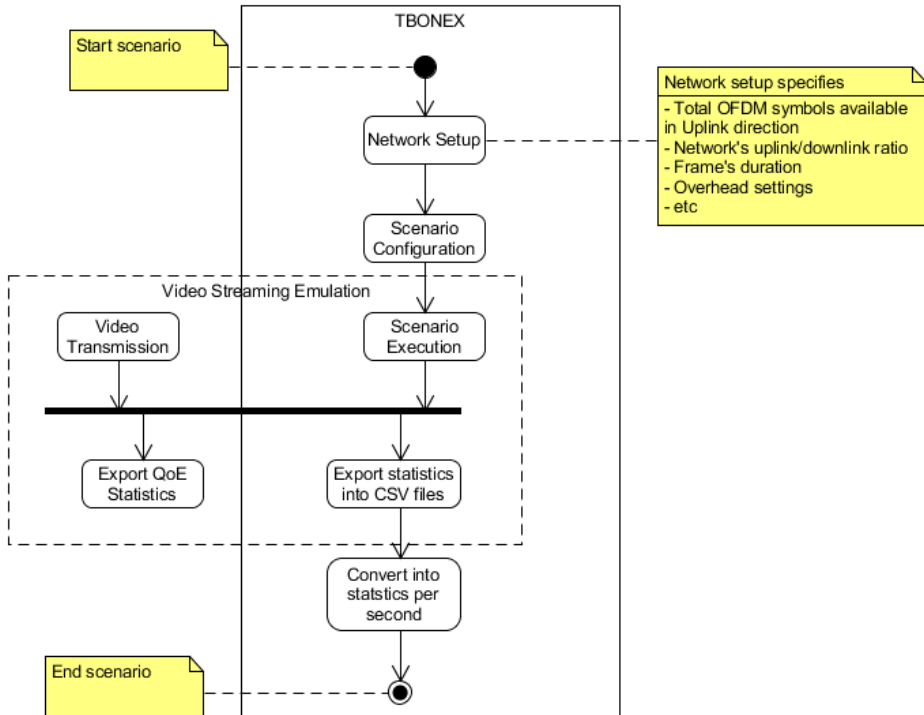


Figure 72 TBONEX scenario flow

In the end if there is any bandwidth available left, the BE users will get the rest divided equally for all of the users. Scheduler allows for dynamically changing number of connections meaning that all the resources reserved for specific user become available for other connections if this user disconnects. Scheduler works until the last packets are being sent which is specified in previous stage called Scenario Configuration. Figure 73 presents the overall, step-by-step scheme of the whole process highlighting the main processes, where all allocation mechanisms take place. By utilizing traces gathered from real life measurements where all MCS (Modulation Coding Scheme) were known, the scheduler is able to calculate throughput according to these statistics.

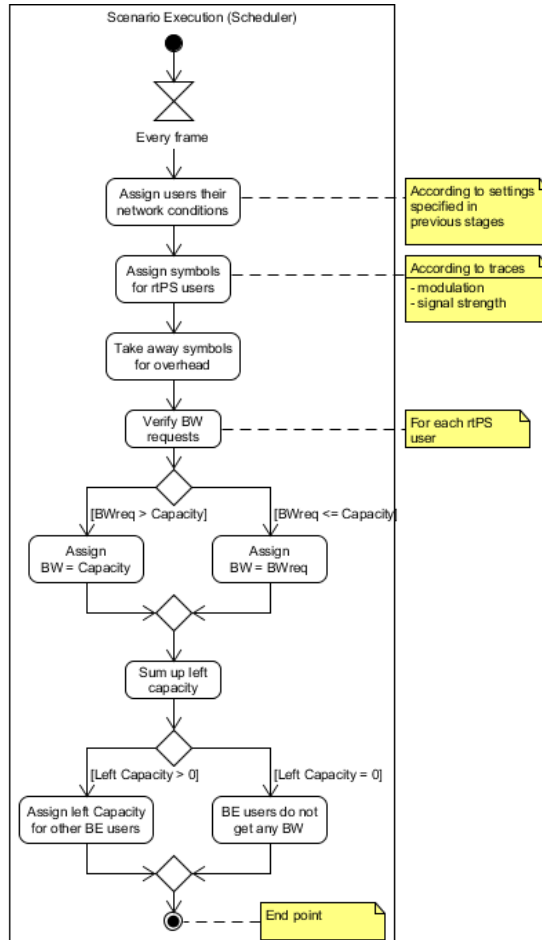


Figure 73 OFDM scheduler inside TBONEX emulator – activity diagram

5.7.2 Baseline trace profiles

Thanks to the metrics captured which were described in the “system model” section, TBONEX is able to estimate delay and available bandwidth of network. Additionally enhanced by settings and packet losses specified by user, tool is creating a traces that are used as a base for network emulation. There can be distinguished 3 types of traces:

- Extreme – mostly NLOS condition, with huge variation in delay and throughput, with high packet loss ratio (exceeding 20%)
- Medium – mostly LOS condition, with medium variation in delay and throughput, with medium packet loss values (exceeding 5% but less than 20%).

- Good – mostly LOS condition, with low variation in delay and throughput, with low (less than 5%) or without packet losses.

5.8 VALIDATING NETWORK DISTURBER

At the stage of framework development and configuration authors have been connecting the proposed Network Disturber and the transmitting and receiving endpoints using wireless (WiFi) but eventually switched to wired network (Ethernet) due to identified unstable behaviour of the wireless card driver in connection with Netem, manifested in anomalous delays. After removing this problem of special interest was to validate the fidelity of replaying the radio conditions based on baseline traces. The tests have confirmed that both instantaneous bandwidth and modulations are properly recovered.

However already at this stage we have found that the delays resulting from sequentially adjusting channel (i.e. available rate) can exhibit large variations. On one hand the fact of delays building, simply indicates that the channel capabilities of a user (car, UAV) have degraded, on the other this is the clear reason for triggering adaptations. In order to minimize extra delays resulting from the internal queues of the Netem tool, it is essential that the source traffic is properly adapted to accommodate to the instant values of the artificial channel. The excessive delays appear when arrival rate on the ingress of the emulator queues exceeds the service rate emulated on the egress of the Network Disturber. In order to mitigate such mismatch an option was introduced in the Bandwidth emulator that configures traffic generation script timing in exact synchronization with the channel changes of the emulator. The resulting data flow (UDP packets) mimics the video source with “ideal feedback”. However, this latter option is used just for controlling QoS of the data stream under particular channel variability emulation (i.e. when it is enabled no video is sent so one cannot evaluate video QoE). The next figure (Figure 74) shows the two plots which represent the effective instantaneous rate of “adaptive flow” at the transmitter (orange plot) and the rate after the packets have been received at the receiver (PC2). It is clearly seen that there is a mismatch between the timing of various “spikes” at both plots. This is caused by “desynchronization” between the flow that starts to grow whenever emulated bandwidth experiences “rate drops”. Such drops cause packets in the delay queues of Netem at the Network Disturber to face the head of line blocking due to slower channel.

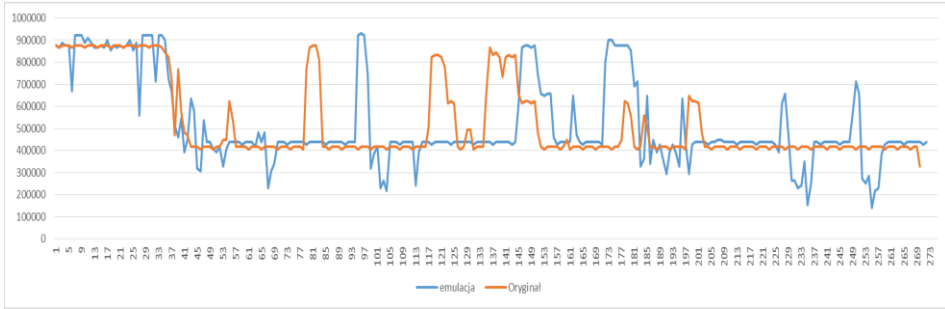


Figure 74 Emulated rate at the receiver (channel - Choszczno mobile)

Following plots demonstrate the delay of the “adaptive traffic” represented by the orange plot on the Figure 74. The first of them (Figure 75) represents exactly the case presented above. The Figure 76 has been plotted after shifting the “orange peaks” to the right, to cover respective “blue peaks”. We can see that delay in the Figure 77 seems to follow the shape of delay from Figure 75 but the “spikes” seen in the latter in the intervals of “80–98 s” and “120–170 s” have been reasonably decreased. **It shows the influence of properly adapting the sending rate at the sender (e.g. car, UAV).** Only if rate adaptation of the source is properly tuned to match the emulated channel, the Netem-based emulation can deliver more realistic delays. However, it can be seen that even though the macro adaptation is applied some smaller variations (mismatches) still cause delay spikes in the Figure 77. This is caused by the temporary behaviour of emulator where “packets delayed by X seconds” are being processed in parallel with “packets delayed by Y seconds”, as the delay manipulations are introduced sequentially by the TC scripts in the Network Emulator.

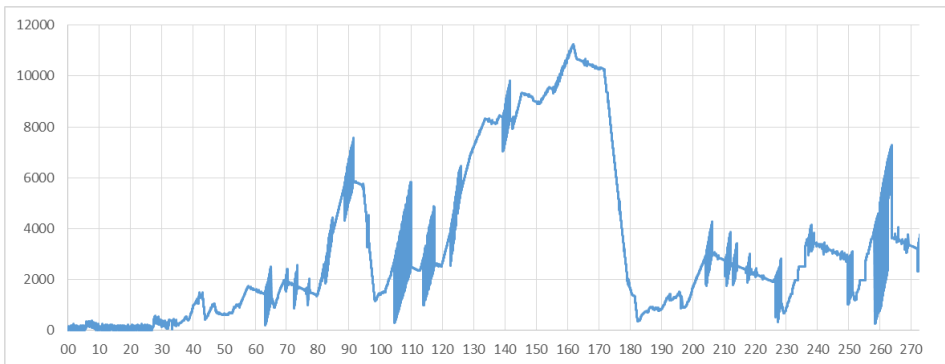


Figure 75 Delay at the receiver - desynced case (trace: Choszczno mobile)

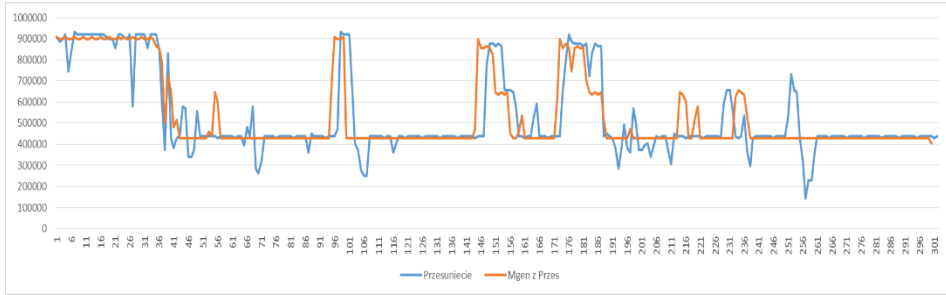


Figure 76 Emulated rate at the receiver (after tuning)

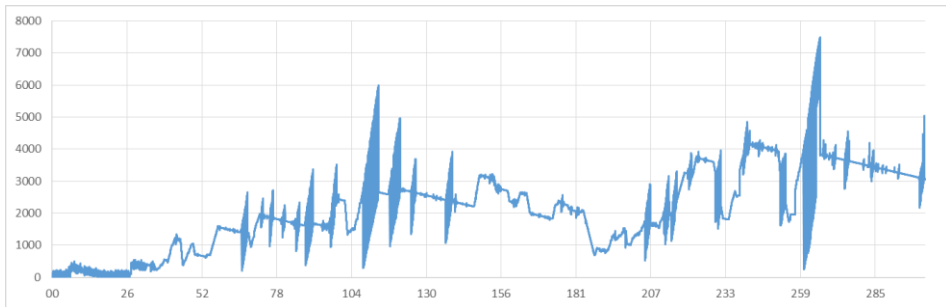


Figure 77 Delay at the receiver - after synchronization was made between source and emulated channel (trace: Choszczno mobile)

The measurements with the real traffic from the camera will be presented in section 5.10.

5.9 E2E CONGESTION CONTROL MECHANISMS FOR SECURITY SCENARIO

Main goal of this section is to apply the complete wireless link emulation, described above and apply them to the design and conceptualization stage for an adaptive video application “MCATS Controller” logic. Such module should be able to decide about the necessary adjustments of a transcoder module (TR) installed inside the “Server” box in order to respond to temporal variability of wireless 4G/5G network link, and support maximizing the E2E QoE.

5.9.1 Congestion control for security scenario

Having in mind all the assumptions defined in section 5.2, the results of real life measurements reported in the introductory section of 5.5, as well as the architectural approach to deal with “quality feedback” inside the generic architecture for congestion control (Figure 54) - below we present a prototype of “congestion control” loop for security scenario (i.e. in the uplink direction).

5.9.2 Remote loop – prototype

The quasi-deployment diagram for enabling the prototypical remote feedback loop is presented in the Figure 78 below. It describes an example solution of the Controller Agent for evaluating congestion control logic which interacts with the transcoder through Controller deployed at the Video Server side – see the left side in the figure. The latter is responsible for adjusting parameters of streamed video. Same time on the Client side, which is composed of the “Receiver” and “Analyzer” (representing the receiver node) the QoE_Probe box denotes the QoE monitoring software used for measurements after [233]. Some more details on the interfaces and methods used to control the transcoder are available in Annex

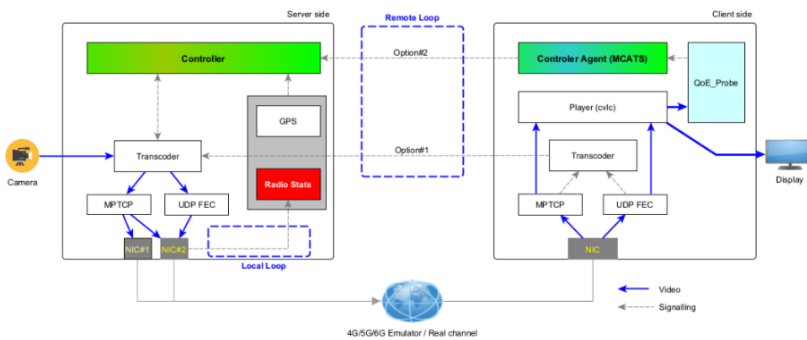


Figure 78 Prototypical remote feedback loop architecture for evaluating congestion control

The above mentioned QoE evaluation models were used to evaluate various settings of video streams in order to identify optimal decisions of the Controller that optimizes video to deliver optimal QoE parameters. Dedicated API can be utilized on the side of transcoder to realize the needed controls. In the case of “remote feedback” from Controller Agent the decision about transcoder parameters for adjusting to a channel quality metrics will be made. The Controller is present near to the real-time streaming video source (camera), the decision about transcoder parameters (bitrate, frames per second, quality profile) for adjusting to a channel quality metrics should be made. However, from practical point of view the feedback about quality (QoE) perceived on the reception side can be acquired by (a) use the out-bound signalling by piggybacking the channel probing packets to IP packets transporting RTP frames or (b) use the in-bound signalling and rely on the QoS statistics provided by the control protocol (e.g. RTCP) between streaming server (Server) node and the receiver side node (Client). Although it is feasible to exchange QoS parameters (e.g. r_i , $d_{E2E,k}$, P_{loss} , P_{DV} , together with its min/max/avg values) between receiver and transmitter e.g. by following the option “b” above, it would not be possible to

achieve in real-time even if some additional elements would be available on the receiving end. That is why in case of evaluating the controller for mobile networks it was decided to focus on the Server side where Controller will be interfaced with the resource monitor („Radio stats”) and be able to read wireless link quality as well as geographical location of the Server node. The Controller at this end is capable of acquiring the real-time statistics about channel/location by probing the modem.

5.9.3 Local loop – prototype

The local loop SW prototype was developed in order to provide test and evaluation capabilities. The details of the cooperation between Controller and: Radio Stats, Transcoder and Controller Agent are described in section 5.3.

5.10 RESULTS OF EMULATING WIRELESS SCENARIOS

We have focused mainly on tests based on real-time video transmission (usually), utilizing wireless channel and the resulting observation of the degradation of video quality and the impact of various scenarios and events on the results. The TBONEX tool has been thoroughly evaluated and validated. In addition, the schedulers implemented were compared with the OPNET simulations as well as with the real life data from the field tests. The following group of tests were performed:

- Validation of the emulator (TBONEX) with IP camera transmitter
- Validation of the emulator (TBONEX) with the use of Server and Client components
- Validation of the emulator, with local video streamed using TCP and Server
- Validating emulation with both rate and delay enabled
- Validating the resource consumption of video processing at the Server
- Validating the influence of TCP use (instead of UDP) with full emulation
- Validating MCATS automation tool.

Each test is described in details together with short summary of findings in the section 20. Thorough discussion of results is provided in the chapter 9.1

5.11 SUMMARY OF EMULATOR FRAMEWORK

In this section it has been shown architecture and application of trace based OFDM network emulator for 4G/5G and future networks. The emulator has been thoroughly validated considering real and artificial traffic (camera, emulator respectively), real channel behaviour (from traces), multiple transport protocols (UDP, TCP), with adaptive traffic (idealistic feedback) etc. The framework can be utilized to ease the costs and effort required to perform real tests at low cost of the environment preparation (it is based on open-source tools). Thanks to the

proposed architecture repeatability of results when testing new versions of video control algorithms is gained as it is possible to replay sessions from real life measurements. Such algorithms (congestion controller) can be placed on the Server node (PC1) and connected with the *Bandwidth emulation* block directly so that it replays channel behaviour in real-time. This way controller installed on PC1 can be tuned as it would react also in real-time to the perceived channel (modulation) changes.

6 QOE CONTROL ALGORITHMS FOR MULTI-RAT

6.1 INTRODUCTION

5G services are envisioned to utilise different RATs (as well as component carriers in case of carrier aggregation) to fulfil ambitious requirements with respect to data rate, latency, reliability, co-existence, coverage, etc. To realise this vision, a base station of the future needs to: integrate different RATs or (ii) enable interworking of stacks at various layers.

The previous chapter 5 has shown, that in single-RAT case for uplink, limiting the quality evaluation of a real-time video stream to a set of baseline QoS metrics only, can lead to incomplete and inadequate understanding of the objective measures of the perceived quality at an end user side. Monitoring of throughput and delay levels, without evaluating other metrics like e.g. packets reordering and jitter at same time can lead to unacceptable QoE. The UL direction has been considered there, as it is the main case for video delivered by cameras present in autonomous cars or drones. While the latter two have recently become very important use-cases for applications in 4G, 5G and next generations of mobile networks. Efficient remote control based on so called “recognition tasks” requires an appropriate level of quality [276] in order to be able to perform operator tasks in a reliable way. Such scenarios may benefit when executed in networks equipped in any of the following features and capabilities:

- MEC/LBO servers introduced close to data-plane/control-plane outputs from RAN, in order to bring video stream processing closer to the source of video feed, as well as deal with challenges of service mobility between access points (e.g. by an application context switching in MEC)
- The use of multi-path transmission in UL, where alternative RATs (multi-RAT, dual-connectivity) or operators can smoothly be activated to convey traffic captured from cameras, especially to mitigate risks of coverage or quality degradation
- Dynamic management of workloads representing 5G (and beyond) network virtualized functions inside of edge servers, or the workloads dealing with the processing of traffic flows in order to assure optimal processing within a network slice
- The use of task offloading between traffic sources and roadside units in order to minimise energy consumed by the node that originates the video stream or improve quality of image recognition and object tracking tasks.

From the point of view of current trends present in wireless networks due to exponential growth in network densification, the availability of programmable network aggregation in the radio access, can be helpful in offloading non-priority data to non-cellular network. Although the LWA/LWIP mechanisms can only be utilised in downlink direction (due to lack of support for the UL), they offer standard-based solutions to let RRM mechanisms decide to decrease cellular

network load by offloading selected data bearers.

The main challenge lies in designing appropriate decision-making agent that will trigger activation of so called split bearer when WiFi is available and difference between delay's in both networks is in an acceptable region. Dealing with mobility of terminal under WiFi AP set is transparent to the controlling cellular AP, and same way when UE is under cellular network coverage, WiFi does not interfere with the decision. As the delay levels between the two networks should not be too high, author in this section focuses on the use of metrics that (i) either directly influence level of delays in the radio interface (SINR) or (ii) are heavily impacted by level of delays (QoE).

6.2 ASSUMPTIONS FOR EXPERIMENT

In this section the author focuses on the delivery of non-priority multimedia traffic in the downlink direction to users with multiple radio stacks in order to exploit optimal strategies for steering the traffic between access networks. The existing LTE and WiFi aggregation (LWA) mechanisms as defined by the 3GPP is supported only in the DL direction. The cellular AP can select to deliver data over LTE, LTE and WiFi or just WiFi or even dual connectivity (DC), which exploits synergy between 4G and 5G. The LWA is mainly designed to provide capacity to offload non-priority data over WiFi, and will be the more effective the lower the delays between LTE and WiFi AP. If this rule is not obeyed it can lead to degradation of performance due to “head of line blocking” of packets from the slower network segment (e.g. LTE or WiFi). Alternatively, the Multipath-TCP (MPTCP) solution based on IP protocol can be especially efficient in the uplink direction, with the limitation of the TCP being the main protocol utilised in delivery of packets from UE to the far-end location. Typically, such a location resides somewhere in the wired network e.g. at the crisis management control room where various video streams are collected for improved situational awareness. All the above-mentioned options are considered in the diagram below. It is critical how to optimally exploit the RAN controller prototype to demonstrate how the enablement of such local control with a certain feedback loop (SINR or QoE) can improve the simple traffic steering based on measuring the instantaneous data rate. The main aspect of this section is to show the involvement towards coordination strategies amongst multiple RATs and integration of parametric control of higher MAC and upper layer network protocols. **Activation of such option can provide significant improvement to admission control actions available to decision making algorithms.**

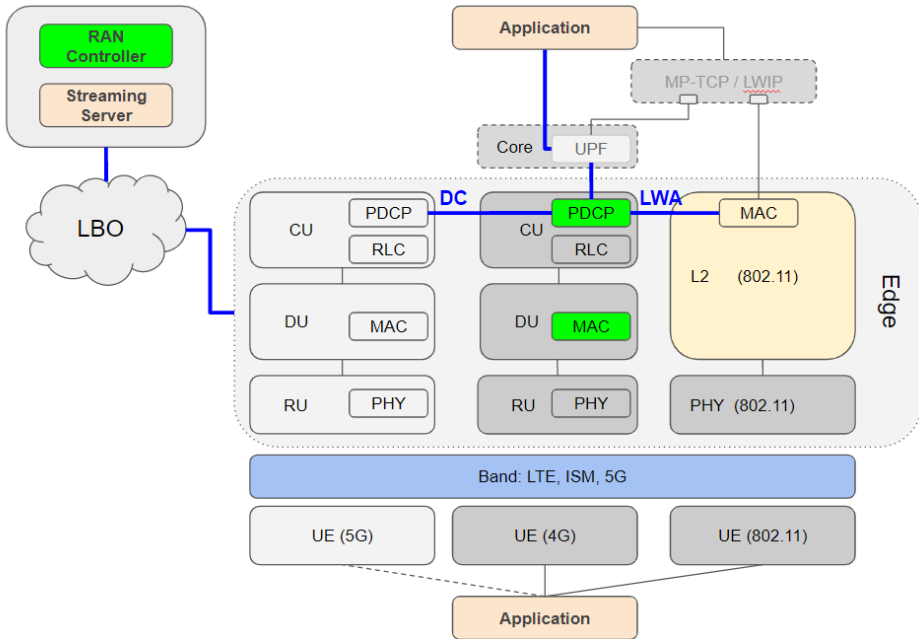


Figure 79 The switching decision is executed within the radio stack i.e. at the PDCP layer, where traffic routing to a second stack (WiFi) is activated based on either the QoE-basis or based on the channel quality (SINR).

Moreover, this chapter also evaluates and examines the possibility to take advantage of the fact of using interworking concepts such as lightweight Internet protocol (LWIP) and LTE-WLAN aggregation (LWA) while making the scheduling decisions. Such a solution would contribute to the software-defined networking (SDN) approach, where a multi-RAT aware scheduler adapts to dynamic channel conditions to provide robustness against severe real-time channel conditions. Finally, we provide comparative analysis of multi-RAT scenarios and evaluate the QoE performance of different scheduling algorithms with SINR based information centric LWA switching and QoE-aware LWA switching by using RANC. The conceptual input/output model of this section is presented in Figure 80.

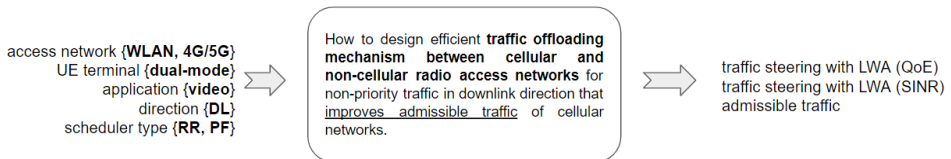


Figure 80 High level diagram of the chapter goal

As it can be seen this chapter defines traffic-steering decisions with the utilization of both a) radio and b) QoE metrics to give more flexibility to the decision making strategy. The offloading to WiFi can be attributed to an admission decision where smart agent is capable of observing and learning the optimal decision upfront. Alternatively, the action of shifting certain connections to WiFi could as well be treated as reactive action in order to perform congestion control in case the current system load or level of selected connections quality is deteriorating.

6.3 APPROACH

This chapter firstly explores a prototype of a RAN controller (RANC) for a multi-RAT environment and evaluates the performance of predicted QoE. The key contributions of the section are summarised as follows:

- Firstly, a prototype of RANC in a multi-RAT environment was implemented and tested within the scope of an external testbed. The prototype of RANC can take over the decision by enabling LWA and LWIP in a multi-RAT environment that can be further communicated with the LTE MAC scheduler to enhance the efficiency of scheduling decisions. Additionally, possibilities of adopting a RANC from a 4G scheduler perspective to support single (LTE) or multiple (LTE and Wi-Fi) technologies were also evaluated.
- Secondly, potential migration of the MAC scheduler from single RAT to multi-RAT technologies was verified and analysed together with the RANC requirements to understand the LWA and LWIP mechanism while identifying the potential issues for multi-RAT deployment for future wireless networks.
- Finally, a comparative analysis of multi-RAT scenarios based on LWA and LWIP concepts are provided that will evaluate the performance of Round Robin scheduler in only LTE access network without RAN Controller (denoted as RR-LTE), the LTE access network only with Proportional Fair scheduler without RAN controller (denoted as PF-LTE), the SINR based information centric LWA switching from LTE to Wi-Fi using a RANC, and the QoE-aware LWA switching from LTE to Wi-Fi via RAN controller, respectively.

6.4 SYSTEM MODEL

This section considers a multi-RAT activation framework as a system model where a user coexist with two different RATs. As shown in Figure 81, the UE will collect the Quality of Service/Quality of Experience (QoS/QoE) Key Performance Indicators (KPIs). During experiments, signal to interference plus noise ratio (SINR) is considered as a QoS-like KPI and the Internet Protocol Television (IPTV) streaming quality parameters (video resolution, playback bit rate, frame rate, the packet loss frequency, and frame loss frequency) as a QoE based KPI. The aforementioned information is collected by the UE and sent to

the so-called specific monitoring agent called “testman server” (a dedicated instrumentation solution) through the evolved Node B (eNB) and it is stored in the database of the RANC entity.

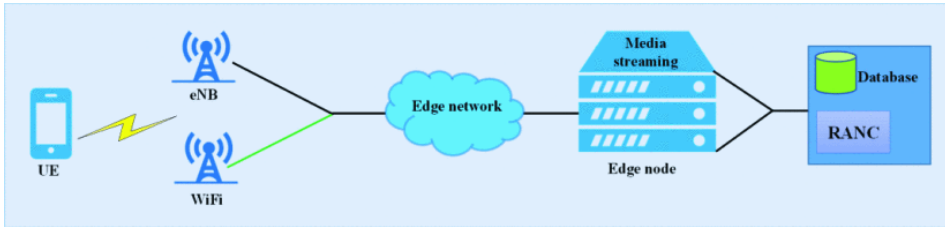


Figure 81 System Model

The RANC retrieves this information from the database and evaluates according to the predicted objective function (predicted QoE). If the predicted QoE criteria is not satisfied, the RANC performs a LWA and LWIP mechanism between LTE and Wi-Fi to meet the predicted QoE. In the experiments, the criteria of activating LWA and LWIP mechanism is based on the SINR or the QoE measurements depending on the scenarios. The RANC for multi-RAT activation is implemented in Python, which performs the LWA and LWIP mechanisms between the LTE and Wi-Fi links. The RANC is composed of multiple applications for the network monitoring and management. The implemented RANC module is composed of several different RAN controller applications which will perform the functionalities of monitoring of SINR via testman client, CSV log of the monitored SINR for further data analysis and network management. In the proposed experimental validation, the LWA and LWIP mechanisms are executed based on the monitored SINR, monitoring of QoE influence factors of the IPTV transmission via MongoDB, QoE measurements of the IPTV transmission using QoE model presented in [283]. The traffic is steered between the LTE and Wi-Fi based on the monitored QoE KPIs of TUD experimental platform.

6.5 MULTI-RAT ACTIVATION ALGORITHM THROUGH RAN CONTROLLER (RANC)

In the section, it is assumed that a single slice is already accepted but its corresponding RAT is switched (tweaked), based on the collected QoS/QoE measurements information from the user and the traffic perspectives. The multi-RAT activation through RANC should allow the user to receive services from different RATs based on the reported QoS/QoE measurement information to the RANC. At the beginning of the multi-RAT activation procedure, the application at the RANC will be collecting the QoS/QoE measurements information from the user. The RANC application will perform the SINR monitoring in real-time for the LTE link and it will also create a CSV log file of the measured SINR. Once the SINR information is available at the RANC it will perform LWA and LWIP

operation-based switching/splitting based on the real-time value of the SINR, e.g., if the SINR value is less than 10 dB, the RANC will perform an action to steer traffic to Wi-Fi link. On the other hand, the video streaming client at the UE is also monitoring the QoE KPI and sending the information to the MongoDB database and it is assumed that the database is available at RANC for collecting the measurement information. The RANC application of multi-RAT activation can perform the QoE monitoring in real-time by collecting updated QoE KPI's from the MongoDB database. Similarly, the application can also generate local CSV log files of the measured QoE.

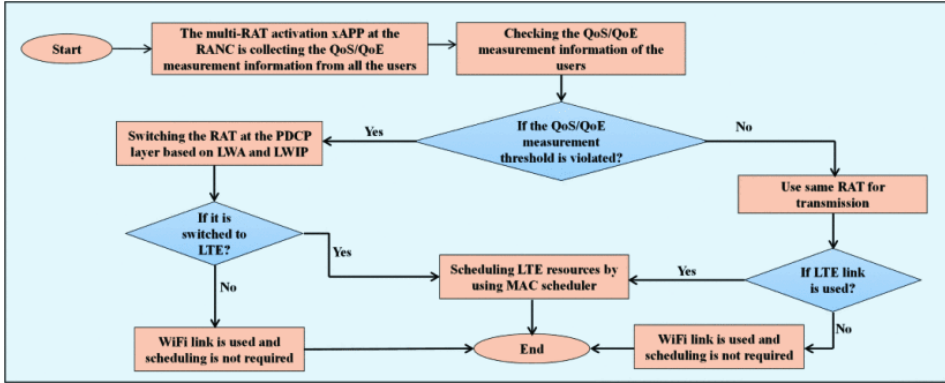


Figure 82 The multi-RAT activation algorithm through RANC

As shown in Figure 82, the switching of the RAT will only happen, if the measured QoS/QoE violated the given threshold. If the threshold of QoS, i.e., the measured SINR of the LTE link is violating a given threshold, it will switch to Wi-Fi link. However, if it is not violating the threshold, it will use the LTE link, and the MAC scheduler at eNB will schedule resources to the user. Similarly, in the case of QoE KPI of IPTV video streaming, if the actual QoE measurement is also violating a given threshold, it will switch from one RAT to another. In the traffic switching stage, if the traffic is steered to LTE link, then the eNB MAC scheduler will schedule resources to the user, otherwise Wi-Fi link will be used and the video streaming will be delivered to the user.

To compute the QoE of IPTV streaming, we utilised the QoE Model proposed in [283] defined as follows:

$$QoE = 1 + \left(v_1 - \frac{v_1}{1 + \left(\frac{BR}{v_2}\right)^{v_3}} \right) \exp\left(-\frac{PLF}{v_4}\right) \quad (6-1)$$

where $v_1 = 3.8$, $v_2 = 4.9$, $v_3 = 3.6$, and $v_4 = 3.5$ are the model coefficients while BR and PLF are the source coding rate of the video and the packet loss rate of the network, respectively. To validate the multi-RAT activation through RANC, we

have carried out an experiment with real HW testbed, by taking an advantage of link diversity (LTE or Wi-Fi) in the system. The system initiates the transmission using LTE under the continuous control from RAN. The RANC takes the responsibility of switching traffic between the links based on the SINR. In addition, one of the customized scheduling algorithms is configured in LTE to exploit fading conditions.

The detailed implementation of the RANC module is composed of several applications and execute following functionalities: i) e.g., monitoring of SINR via dedicated “testman client” log file of the monitored SINR for further data analysis, ii) network management by enabling LWA and LWIP and switching based on the monitored SINR, iii) monitoring of QoE influence factors of the IPTV transmission via Mongo DB, iv) QoE measurements of the IPTV transmission using QoE model presented in [283], and the QoE-aware network management using LWA and LWIP enabling and steer traffic based on monitored QoE KPIs, respectively. The experimental validation of multi-RAT activation through RANC will be discussed in detail in the reminder of this chapter.

6.6 FUNCTIONAL ARCHITECTURE OF MULTI-RAT ACTIVATION

The proposed architecture follows the information-centric functional architecture for the network monitoring and management in the multi-RAT environment. The architecture is inspired by the work proposed in [284] where the probe installed at the UE terminal provides QoS/QoE KPIs to the controller to perform network management operations. The proposed architecture as shown in Figure 83, assumes that the probe on the UE will collect the QoS/QoE KPIs and deliver it to the RAN controller for the RAT activation decision. In the section, we consider SINR as QoS and IPTV streaming as a QoE KPI-s that will include video resolution, playback bitrate, frame rate and packet loss frequency/frame loss frequency. The aforementioned information is collected by the UE probe and stored in the database that is accessible by the RANC on regular time intervals e.g., the UE sends the collected information after every 1 second to the database. The RANC accesses this information from the database and evaluates according to the objective function. If the objective function criteria are not satisfied, the controller performs LWA and LWIP procedures between the LTE and Wi-Fi to meet the objective criteria. Moreover, the customized LTE MAC scheduler is a part of the experiments which run in all the scenarios. The following information is being exchanged among the modules which is shown in Figure 83 and the functionality of each step is described as follows:

1. Feedback from the UE: The UE probe sends the QoE related KPIs to the database.

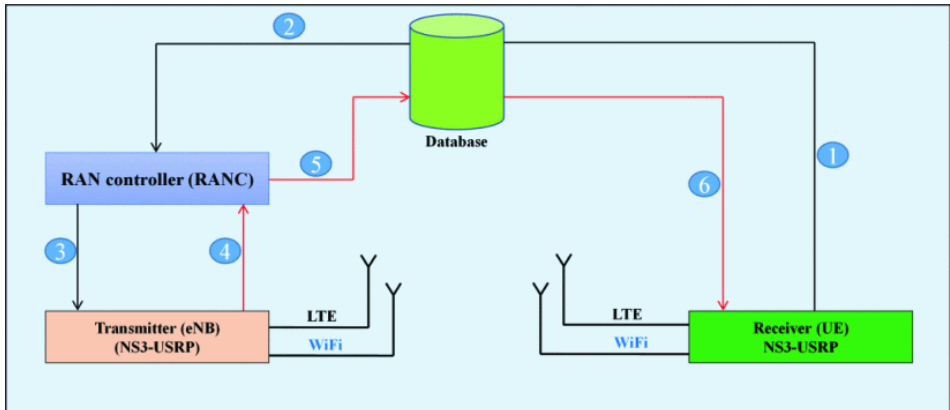


Figure 83 Functional Architecture of Multi-RAT Activation Through RANC.

2. Retrieval of KPIs: The RAN controller retrieves the QoE related KPIs from the database.
3. Control Operation: The RAN controller performs the control action based on the measurement criteria.
4. Acknowledgement from the base station after performing control action.
5. Acknowledgment from controller to the database that control action has been performed.
6. Acknowledgment from the database to the UE that control action has been performed.

The flow diagram of exchanging the messages is visualized in the Figure 84

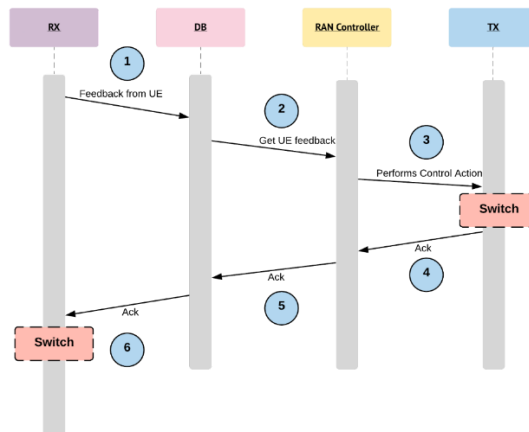


Figure 84 Sequence flow diagram for RAN controller based LWA/LWIP switching

The service considered here is the IPTV video transmission as a use case and is validated in the later sections.

6.7 CUSTOMISED LTE SCHEDULING FOR MULTI-RAT TESTBED

Scheduling in LTE mainly revolves around resource allocation type. Resource allocation type defines the pattern in which resources should be allocated for each transmission. Moreover, it also specifies the flexibility of resource block allocation. In LTE, different resource allocation types have been defined where each of the resource allocation type uses a predefined procedure. In this regard, three different allocation types in LTE are proposed, namely resource allocation type 0, 1, 2. In this chapter, we use Resource allocation type 0 as it is the simplest of all and generalises most of the use-cases in LTE. Resource Allocation Type 0: Resource allocation type 0 divides the resource blocks into multiple resource groups, called resource block group (RBG). The number of resource blocks in a RBG varies depending on the system bandwidth. The Table 36 shows the relationship between resource block group size (RBS) and system bandwidth.

Table 36 System bandwidth vs RBG size for LTE systems

System Bandwidth (MHz)	RBG size
1.4	1
3	2
5	2
10	3
20	4

Resource Allocation type 0 allocates the resources using a bitmap and each bit represents one RBG. The allocation granularity is RBG, i.e., the minimum resource that is allocated to any user is one RBG. In this regard, we propose two customised functions in the scheduling block in LTE-MAC, namely interleaving and localized. In the conventional Round Robin schedulers, as indicated in the Figure 85, static allocation is considered and so the localised RBGs are allocated.

```

NI.CLIENT: sent 1000 bytes to 7.0.0.2 Uid: 89 Sequence number: 3
RrFfMacScheduler::DoSchedDLRlcBufferReq line 447
11111111110000000000000000000000---->we have just printed allocation rbgs in system
NI.SERVER: received 988 bytes From 10.1.1.2 Uid: 89 Sequence Number: 3
NI.CLIENT: sent 1000 bytes to 7.0.0.2 Uid: 96 Sequence number: 4
RrFfMacScheduler::DoSchedDLRlcBufferReq line 447
11111111110000000000000000000000---->we have just printed allocation rbgs in system
NI.SERVER: received 988 bytes From 10.1.1.2 Uid: 96 Sequence Number: 4
NI.CLIENT: sent 1000 bytes to 7.0.0.2 Uid: 103 Sequence number: 5
RrFfMacScheduler::DoSchedDLRlcBufferReq line 447
11111111110000000000000000000000---->we have just printed allocation rbgs in system

```

Figure 85 Conventional Round Robin Scheduling Approach

In Figure 86, the localized RBG allocation is about scheduling contiguous RBGs in a sub-band. Unlike interleaving RBG allocation, the localized RBG allocation spans only a portion of the system bandwidth. The adjacent RBG allocation size depends on system designer, and it can be varied depending on the requirements.

```

RrFfMacScheduler::DoSchedDLRlcBufferReq line 447
11111111111100000000000000000000----->we have just printed allocation rbgs in system
NI.SERVER: received 988 bytes from 10.1.1.2 Uid: 123 Sequence Number: 8
NI.CLIENT: sent 1000 bytes to 7.0.0.2 Uid: 130 Sequence number: 9
RrFfMacScheduler::DoSchedDLRlcBufferReq line 447
0000000000001111111111110000----->we have just printed allocation rbgs in system
NI.SERVER: received 988 bytes from 10.1.1.2 Uid: 130 Sequence Number: 9
NI.CLIENT: sent 1000 bytes to 7.0.0.2 Uid: 137 Sequence number: 10
RrFfMacScheduler::DoSchedDLRlcBufferReq line 447
11111111111100000000000000000000----->we have just printed allocation rbgs in system
NI.SERVER: received 988 bytes from 10.1.1.2 Uid: 137 Sequence Number: 10
NI.CLIENT: sent 1000 bytes to 7.0.0.2 Uid: 144 Sequence number: 11

```

Figure 86 Customized Scheduling with Localized Approach

The interleaving RBG allocation, as shown in Figure 87, allows the use of entire bandwidth by allocating RBGs across the entire system bandwidth. Through interleaving RBG, frequency diversity can be achieved in highly frequency selective channels. The utilization of resources is considered equal in all allocation strategies, thus, making it a fair comparison.

```

RrFfMacScheduler::DoSchedDLRlcBufferReq line 447
0101010101010101010101010000----->we have just printed allocation rbgs in system
NI.SERVER: received 988 bytes from 10.1.1.2 Uid: 337 Sequence Number: 39
NI.CLIENT: sent 1000 bytes to 7.0.0.2 Uid: 344 Sequence number: 40
RrFfMacScheduler::DoSchedDLRlcBufferReq line 447
101010101010101010101010100000----->we have just printed allocation rbgs in system
NI.SERVER: received 988 bytes from 10.1.1.2 Uid: 344 Sequence Number: 40
NI.CLIENT: sent 1000 bytes to 7.0.0.2 Uid: 351 Sequence number: 41
RrFfMacScheduler::DoSchedDLRlcBufferReq line 447
0101010101010101010101010000----->we have just printed allocation rbgs in system

```

Figure 87 Customized Scheduling Approach with Interleaving Approach

6.8 EXPERIMENTAL VALIDATION

In this section, the details of the validation of multi-RAT activation through RANC are presented.

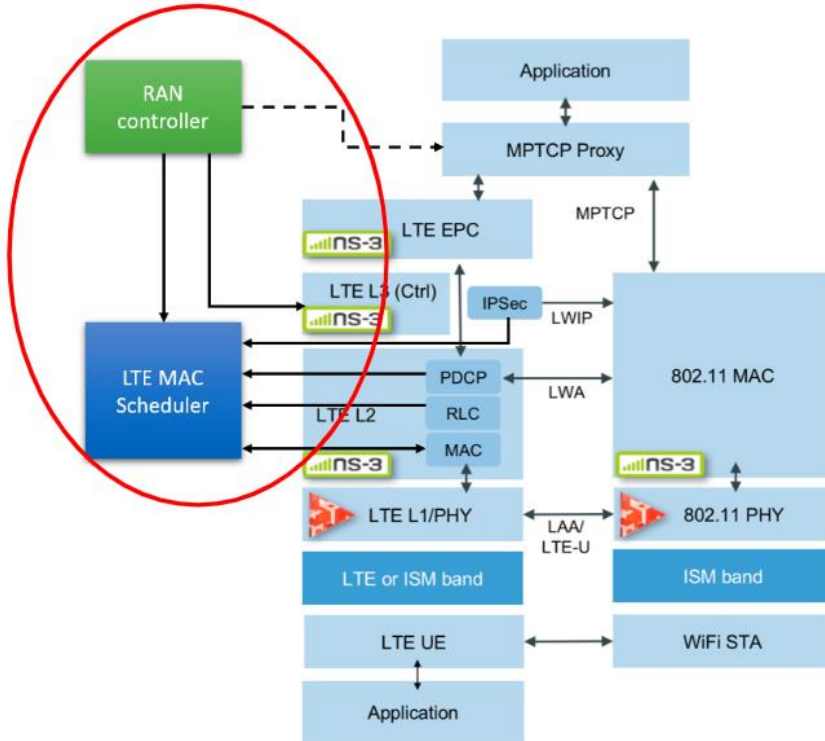


Figure 88 Experiment design of intelligent switching in multi-RAT

The experiments are performed by setting a number of runs to 5 per scenario. In the experiments, we use a two-minute video sequence of Big Buck Bunny which is encoded in H.264 having HD 1080p resolution and 30 frames per second frame rate. The IPTV streaming based on UDP protocol is started from the media server to the media client over the air interface – which means that video is delivered in downlink. For all the experiments, we have utilized Additive White Gaussian Noise (AWGN) at frequency 2 GHz with the 2 dBm output power and 10 M samples/s generated by the interferer setup. The generated noise by the interferer is coupled with the LTE link in the testbed setup by combiner in the downlink direction. The generated noise by the interferer varies the SINR in the range 12-14 dBs at the LTE link. For the RAN controller-based scenarios to monitor the QoE and SINR, the monitoring frequency of the probe is kept to 1 Hz (1 second sampling interval) in all the runs. For the LTE link, we consider Round Robin and Proportional Fair schedulers at the MAC layer.

The SINR is made available at the RANC through a testmen server that enables sending the SINR value to testmen client, as shown in Figure 88. In order to leverage the multi-RAT testbed, intelligent RAN switching can be a viable design for evaluation in the testbed. Because of the reconfigurable hardware used in the experiment, such implementations can be performed in a software

environment where the RAN controller performs actions based on information sharing with the scheduler. Moreover, we have also conducted another experiment in a multi-RAT environment based on the selecting the modulation and coding scheme (MCS) based on the channel quality indicator (CQI) of LTE links.

6.9 SCENARIO1: INTELLIGENT RAN SWITCHING BETWEEN LTE AND WIFI

6.9.1 Experiment Aim

In this experiment, we aim to design a setup where RAN controller switches between LTE and WiFi based on the radio conditions in operating frequency bands.

6.9.2 Experiment Details

The experiment is about taking advantage of link diversity (LTE or WiFi) in the system. The system initiates transmission using LTE under the continuous control from RAN. RAN controller takes the responsibility of switching between the links based on the signal to interference plus noise ratio (SINR). In addition, one of the customized scheduling algorithms is configured in LTE to exploit fading conditions. The SINR is made available at RAN controller through testmen server that enables sending the SINR value to testmen client, as shown in Figure 89.

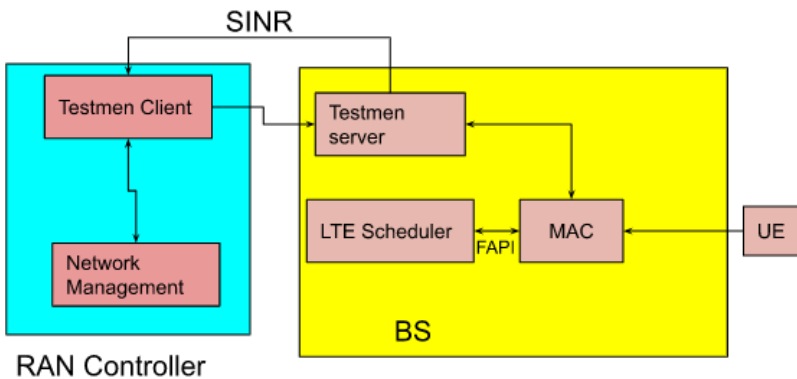


Figure 89 Experiment Design of Intelligent switching in multiRAT framework

The main motivation behind designing this experiment is that the underlying testbed does not support simultaneous activation of both LTE and WiFi links. In order to leverage multi-RAT testbed, intelligent RAN switching can be a viable design to evaluate testbed. Because of reconfigurable hardware used, such implementations can be performed in software environment where RAN controller perform actions based on information sharing with scheduler.

6.10 SCENARIO2: SELECTING MODULATION AND CODING BASED ON LINK QUALITY

6.10.1 Experiment Aim

Another experiment we can conceive is the selection of a modulation coding scheme (MCS) for LTE based on channel quality indicator (CQI).

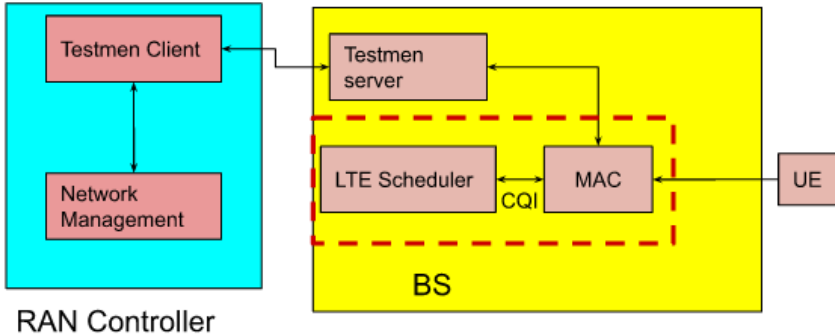


Figure 90 Experiment Design of Selecting MCS based on CQI

6.10.2 Experimental Detail

The main motivation behind this experiment is that there is only one UE available in testbed setup, which rules out the possibility of multiple access. Moreover, scheduling of radio resources is only available over LTE interface. We can see in Figure 90 that red dotted lines encircled the MAC and scheduler involves the core operation in this experiment design. The LTE scheduler decides upon MCS based on channel quality indicator (CQI) sent by the UE. The plan is to create a dynamic setup wherever changing environment conditions enable the system to adapt to real channel conditions. Such dynamic channel conditions can be due to the introduction of interference in experimental setup. As a result, SINR can be made variable. The main essence of the setup is to conceive total control over resource allocation that impacts directly on MCS selection. Such changes are not applicable for the WiFi available in the TUD testbed (i.e. the non-802.11ax access points). Further extension of the TUD testbed should seriously consider its capabilities to also cover the up to date WiFi6 APs. This way a more relevant scheduling of resources in both networks could take place.

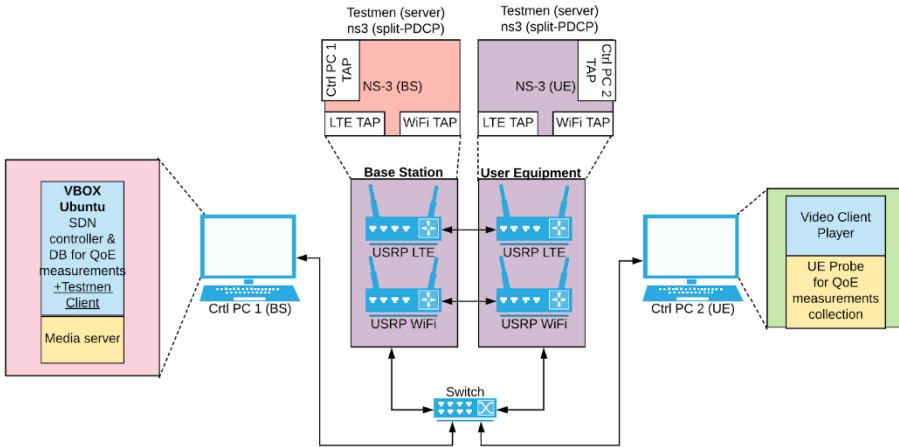


Figure 91 Overview of the experiment Setup

6.11 DEPLOYMENT OF THE IMPLEMENTED MODULES

For the experiments, as shown in Figure 91 the media server is implemented within the Win7 Host PC connected to the eNB PXI controller whereas the media client is made at the Win7 Host PC connected to the UE PXI controller. The RAN controller modules and the database for the QoE related information exchange are implemented inside the Ubuntu PC. In order to exploit the multi-RAT use case of the testbed, we intend to add noise to the LTE link so that RANC can trigger automatic switching to Wi-Fi link based on the SINR and QoE measurements. For the addition of the noise, we implemented the noise generator using LabVIEW. The noise generator output is combined with the transmission of LTE link using the combiner in the cabling setup to add noise in the LTE downlink channel. The experiment utilises the NS3 in the TAP bridge configuration which means that external traffic is forwarded to the NS3. Therefore, two packet forwarding scripts are used at PXI controllers: first at eNB PXI controller for forwarding the media traffic from media server into the NS3 TAP bridge and second at UE PXI controller to forward the media traffic from the NS3 TAP bridge to media client. All the implemented modules are deployed successfully in the testbed except the scheduler module. Though the customised scheduler module works fine in the standalone locally installed NI NS3 (simulation mode of NS3 with SDR) but the deployment of the scheduler module into testbed encountered problems using the real physical layer provided by NI LTE Application Framework. Consequently, we have performed the evaluation of the customised scheduler in the standalone NI NS3 which was locally installed.

6.12 EXPERIMENT EXECUTION PROCESS

This section describes the execution process of the proposed experiments. The execution of the experiment followed nine steps in total including the one offline step of data analysis of the collected QoE KPIs from the experiments to acquire results. The process of the experiment execution has following steps:

- Step 1: For each scenario, the first step of experiment execution involves starting of the 802.11 Application Framework (needed for Wi-Fi/LWA) and the UE NS3 instance using NI LTE and 802.11 Application Frameworks (AFW). This is performed by compiling the NI NS3 configuration code on the UE controller followed by starting the transmitter and receiver at UE side.
- Step 2: The second step in the experiment execution involves the compilation of the NI NS3 configuration code at eNB followed by the initiation of the transmitter and receiver at eNB side
- Step 3: Once the eNB and UE instances start working, the noise generator at the interferer side is configured and initiated to degrade the LTE channel quality of our use case scenario.
- Step 4: The media server implemented via VLC is configured for the IPTV video transmission. In this process, a video file is added in the media server for the UDP legacy IPTV streaming which sends the video stream to the eNB PXI controller.
- Step 5: In order to forward the video streaming in the NS3 from the media server, video streaming forwarding scripts are used at PXI controller. The forwarding script transfers IPTV video packets to the NS3 TAP bridge which then passes through the NS3 and then over the air interface towards the UE side.
- Step 6: The video forwarding script at UE PXI controller is initiated to forward the video traffic from NS3 TAP bridge at UE to Win7 host PC host where the media client is running.
- Step 7: Depending on the scenario, the RAN controller instance is started. In case of SINR based LWA switching, SINR based RAN controller application is started while in case of the QoE based LWA switching, QoE based RAN controller application is started. In case of only LTE link (baseline case with LWA enabled switching) RAN controller is not initiated.
- Step 8: The media client at the Win7 host PC attached to the UE side is started for the IPTV media streaming. Once the media client is started, the UE client side probe implemented in the media client starts sending the QoE KPIs to the database available to the controller.
- Step 9: The data analysis step is performed offline. In this step, the data regarding QoE KPIs for all the scenarios is analysed to acquire the final results.

The IPTV media streaming was implemented via VLC in Python language. In the current implementation, the media client receives the stream from the media server. The user-end probe implemented at the client side collects the QoE KPIs. The IPTV media streaming module is composed of Media Server and Client. In the Media Server the Media files are placed in the media server which will broadcast the IPTV transmission using UDP legacy protocol at the network transport layer. Moreover, the Media Client will perform the IPTV media streaming from the media streaming server over UDP protocol. It will also collect the media streaming session information (QoE KPI) via passive user-end probe for QoE monitoring. Additionally, it will also store the QoE KPI in the database on the regular intervals based on the monitoring frequency. The Media Client generated CSV logs of the monitored QoE KPI's for further analysis.

6.13 RESULTS

This section provides a discussion on the achieved results from the proposed experiments. The results provide comparative analysis of multi-RAT scenarios where the LWA based switching is performed by the RANC in case of LTE channel degradation. We have compared the results for different network scenarios.

- Round Robin scheduler in only LTE access network without RAN Controller which is represented as RR-LTE.
- The LTE access network only with Proportional Fair scheduler without RAN controller. This scenario is represented as PF-LTE. The SINR based information centric LWA switching from LTE to Wi-Fi using a RANC.
- This is a multi-RAT scenario which is represented as SINR-RANC. The QoE-aware LWA switching from LTE to Wi-Fi via RAN controller.
- This scenario is a multi-RAT use case where QoE information is shared from UE to the RAN controller. In the results, this scenario is represented as QoE-RANC.

Figure 92 represents the mean predicted QoE (averaged over the number of runs) over epoch time. For the LTE access network only (RR-LTE and PF-LTE), in the case of the noisy channel, it is visible that IPTV streaming quality varies abruptly. The high level of QoE variations in RR-LTE and PF-LTE are due to the frequent occurrence of the packet loss events in the noisy channel. As the QoE model for QoE prediction highly depends on the packet loss frequency, the delivered QoE in the RR-LTE and PF-LTE scenarios fluctuates throughout the streaming session. The use of the proposed SINR-RANC and QOE-RANC algorithms on top of the RAN controller allows keeping the quality of video as priority by switching the video smoothly at the PDCP layer to the WiFi for offloading.

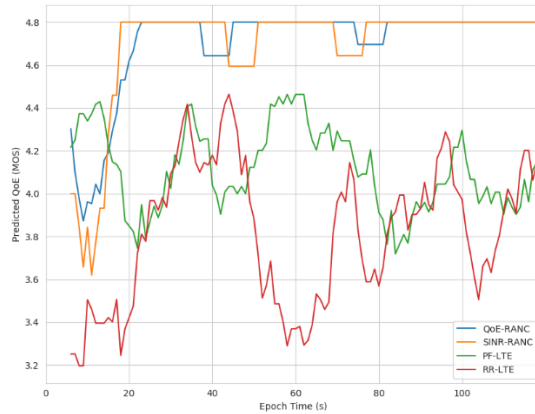


Figure 92 Predicted QoE (MOS) over Epoch Time (s) for each scenario.

While in case of multi-RAT with RAN controller (SINR-RANC and QoE-RANC) once the predicted QoE goes below 4, the RAN controller switches from LTE to WiFi to improve the delivered QoE. In the SINR-RANC and QoE-RANC scenario, the QoE degradation can be noticed until 15s epoch time⁷, after which the RAN controller triggered LWA switch from LTE to WiFi link leading to the significant improvement in the delivered QoE of IPTV streaming. Moreover, the slight degradation in the delivered QoE in SINR-RANC and QoE-RANC scenarios can be observed when switching to WiFi link. This is due to the packet loss events however it can be noticed that the probability of the packet loss event is very low in the SINR-RANC and QoE-RANC scenarios as compared to the RR-LTE and PF-LTE scenarios.

Figure 93 provides a comparison of the four scenarios in terms of accumulated average delivered QoE for all experiments run in all scenarios. Mean Opinion Score (MOS) is a QoE metric for multimedia (video and audio) traffic ranging from 1 (bad) to 5 (excellent). The mean delivered QoE in case of SINR-RANC and QoE-RANC is higher 4.5 MOS while in case of RR-LTE and PF-LTE lower mean QoE is delivered. The higher delivered QoE is the multi-RAT depicts the effectiveness of the information centric RANC, to trigger the LWA based switching between LTE and WiFi. Both for the SINR-RANC and the QoE-RANC it can be observed that it is delivering almost the same mean QoE for IPTV

⁷ Note that this delay includes the total delay from UE reporting to the triggering of LWA which includes delay from UE to controller for sending measurements, delay from controller to the testmen server for sending the control signals, delay for enabling LWA switch from LTE to WiFi and delay for the forwarding scripts

streaming. This is due to the reason that once the traffic is switched to the Wi-Fi link the channel condition remains the same for both scenarios. Moreover, the time taken by the RANC to enable LWA switching from LTE to Wi-Fi in both cases is about the same.

Figure 94 shows the histogram plot of the delivered QoE in all scenarios. The histogram plot for RR-LTE and PF-LTE highlights that the delivered QoE is most frequently in the range of 3.5 and 4.0 MOS respectively.

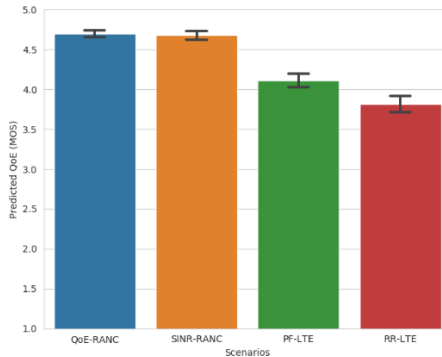


Figure 93 Accumulated average predicted QoE for each scenario.

The spread of the QoE histogram plot for RR-LTE and PF-LTE scenarios shows that the QoE is not ensured in these scenarios and QoE varies from 1.5-4.6 and 3.0-4.75 MOS respectively. Whereas in SINR-RANC and QoE-RANC, the higher QoE is delivered most often which shows that RAN control usage with the multi-RAT enabling technology ensures the delivered QoE to the user in case of the bad channel conditions. Furthermore, the spread of the histogram for SINR-RANC and QoE-RANC varies from 2.5-4.6 and 3.4-4.8 MOS respectively. According to the study in [283], the frequent changes in delivered quality of the video streaming leads to lower user perceived quality. Therefore, among all the scenarios the QoE-RANC approach outperforms others in terms of the QoE maximisation as the QoE-RANC approach delivers higher QoE with less variance.

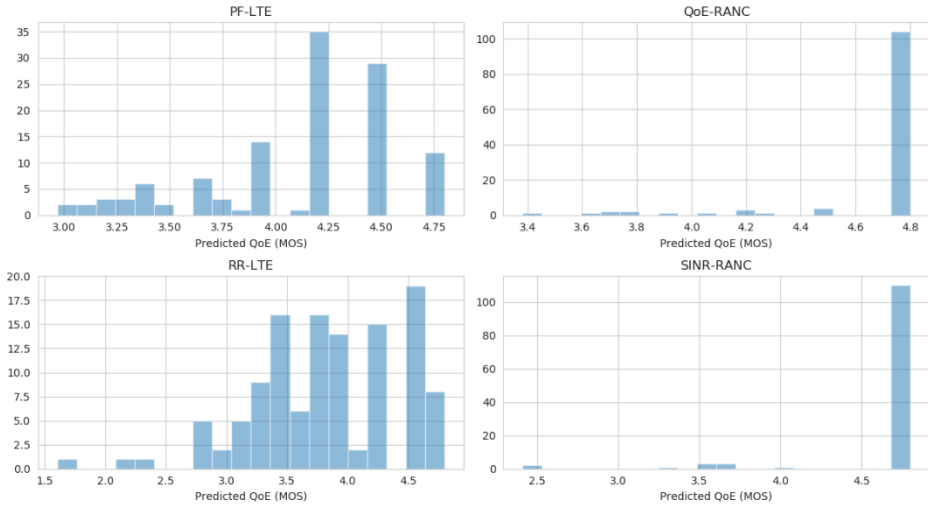


Figure 94 Histogram plot of predicted QoE (MOS) in each scenario.

6.14 SUMMARY

In this chapter a prototype solution has been built to test the advantages of a multi-RAT environment controlled by LWA and LWIP mechanisms and the corresponding experiments have been performed. To implement the switching from the two radio access technologies selected, LTE and Wi-Fi, a RANC controller module has been integrated. These experiments clearly show the advantages of the use of LWA and LWIP.

7 5G ORAN WORKLOAD PREDICTION TO SUPPORT CAC ALGORITHMS

7.1 INTRODUCTION

This chapter introduces the importance of the workload prediction for virtualized radio access network (RAN) deployment in the edge micro data centre (EMDC). To predict the workload, several machine learning algorithms are evaluated in terms of central processing unit (CPU) usage from the Kubernetes cluster while deploying the vRAN components (i.e., radio unit, distributed unit, and centralised unit) in the EMDC. In addition, the prediction results were validated by using data collected from an experimental testbed. As already indicated in the chapter 2, none of the above-mentioned research indicates the workload prediction of vRAN in the EMDC which represent an emerging trend of AI/ML capable edge servers. In this chapter, the author shows the importance of workload prediction of 5G disaggregated vRAN in EMDC. The main contributions are listed as follows:

- Firstly, a novel system model is proposed for the workload prediction and CPU usage forecasting mechanism in an EMDC architecture.
- Secondly, several ML algorithms for CPU usage prediction are proposed in EMDC architecture based on Long Short-Term Memory Neural Network (LSTM), Auto Regressive Integrated Moving Average (ARIMA), and interpretable time series forecasting (N-BEATS) in combination with collecting the data from a real testbed.
- Moreover, a reference comparison of accuracy of prediction is made against legacy regressors (section 7.6).

The ML-based workload placement process, as introduced by the author in Figure 95, will collect the data of interest from a specific application (workload) running over the EMDC platform. The ML-based framework will learn the task computation workloads e.g., CPU cycles per bit which can be available from offline measurement that can be in-prior collected by the telemetry framework. The collected data will be compressed and stored for training the ML model and from the trained model the framework will take an optimal decision to offload the workload, e.g., which tasks should be offloaded through the connectivity of 5G vRAN for the downlink streaming and/or which tasks should be similarly offloaded through the connectivity of 5G vRAN to which EMDC (if multiple EMDC is available) within the architecture. The ML optimization tools and techniques available at the EMDC edge servers, will support training of the learning model based on the application-specific data and provide an optimal decision by maximising or minimising and analysing the KPI's as a reward of the training model. The EMDC HW solutions are robust enough to support the requirements of URLLC applications but whenever such low latency is not required, it might be worth optimising workloads at the cloud level to save local

computing resources and process higher volumes of data in a single location. Here an example could be to port selected workload into could due to lower cost of computing there, or only in certain conditions of the environment.

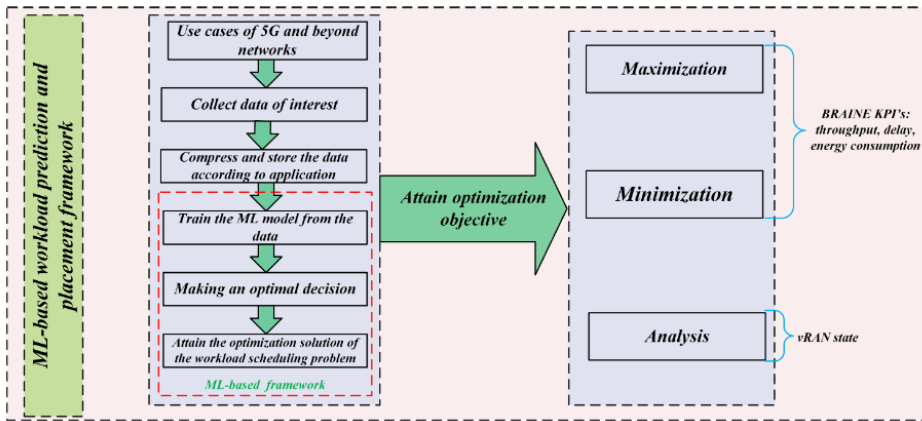


Figure 95 ML learning-based workload prediction and placement process for the EMDC architecture

The common denominators are the data-driven nature, learning-based operation, cross-layered approach (to deconflict optimization directions), etc. The Figure 95 shows the evolution in the way such mechanisms should be designed and cross-dependencies between them should be considered. Nevertheless, the utmost importance for attaining is the high energy efficiency and energy consumption minimization - which should be always be priority for the future proof network designs. The particular AI/ML solutions used to augment (amend) selectively radio protocols of B5G, which have been so far identified in the prior state-of-art, has been highlighted in the chapter 2.

7.2 SYSTEM MODEL AND PROBLEM DEFINITION

Figure 96 shows the proposed system model of workload prediction. At the first step, the metrics which will be used for the prediction algorithm are collected from the 5G Open RAN radio stack gNB network function and they are delivered with internal EMDC messaging to the Resource Manager (RM) entity. A predictive technique is defined as a statistical model that can be applied to known data of a given phenomenon to estimate future metric evolution [285]. In this work, the RM utilises this data to feed arbitrary prediction techniques based on several ML algorithms such as ARIMA [286], LSTM [212], and N-BEATS [287], with proper inputs that allow characterization of the virtualized gNB operation regarding its demand for computing resources of the EMDC.

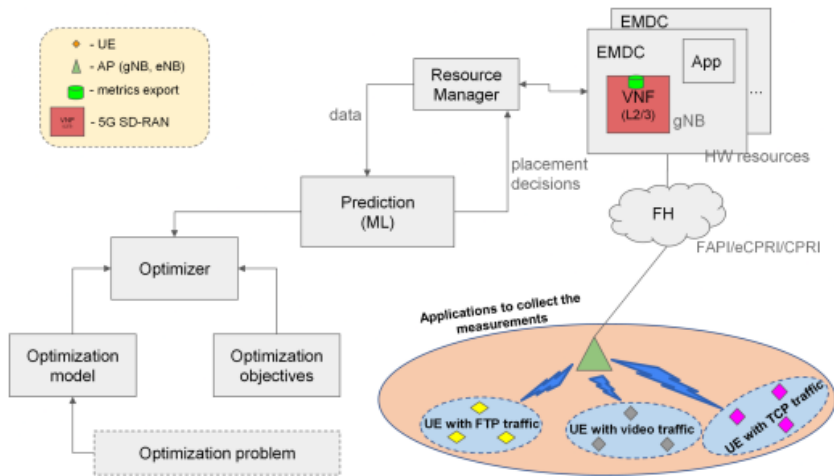


Figure 96 System model for workload prediction

The optimization problem in the figure is considered as a CPU usage forecasting which can predict the future metric of the underlying RAN. However, the optimizer block in the proposed system can be further extended for optimizing the key performance indicators (KPIs). Such KPIs can be considered as an optimization objective building block and prediction of the parameters for defined problems can be solved at the prediction building block and the decision can be delivered at the optimizer for the underlying RAN. The workload forecasting steps are described as follows: firstly, the metrics are collected from the 4G/5G Open RAN radio stack gNB network function and they are delivered using a data bus of EMDC to the Resource Manager (RM) entity. Secondly, the RM utilizes this data to feed arbitrary prediction algorithms (mainly based on model-free approaches) available inside of the prediction block, with proper inputs that allow characterization of the virtualized gNB operation regarding its demand for computing resources of the EMDC. Finally, the CPU usage metric from the Kubernetes cluster is considered for the prediction techniques to forecast the vRAN workload. The detailed procedure of the prediction technique is briefly discussed in the next section.

7.3 PREDICTION AND OPTIMIZATION ALGORITHMS

Time series forecasting is used for predicting the future events based on current and past data value. As shown in the chapter 2 there are many different types of time-series forecasting models, each with its own strengths and weaknesses. Our problem of predicting CPU usage can be described as a univariate time series point forecasting task, which is characterized as a ML problem. Applications of

time-series (TS) forecasting either aim to predict future behaviour such as seasonality and trends or to detect anomalies. Working with a resource workload key focus is on seasonality and anomalies. In this section, both classical statistical time series models and deep learning models are considered in order to evaluate the performance of workload prediction of vRAN. For classical time series models, ARIMA is used. In case of deep learning models, the LSTM, and N-BEATS are considered. In the following subsection, we briefly describe these models with their characteristics.

7.3.1 ARIMA

ARIMA is actually a time series model which predicts future events or points in the series based on its own past values combined from its own lags and lagged forecast errors. For this model, a ML model is implemented which predicts value based on its own past trained value.

The ARIMA model comprises three parameters (p, d, q) which are described in the remained of this sub-section. Autoregressive, AR(p) part of the model describes the linear dependencies between target variable (e.g., CPU usage for this) and its lagged (previous) observations, where lag is set by parameter p and can be written as

$$Y_t = \phi_0 + \sum_{i=1}^p \phi_i Y_{t-i} + \epsilon_t \quad (7-1)$$

where ϕ_0 is a constant term, ϕ_i are autoregressive coefficients, and $\epsilon_t \sim N(0, \sigma^2)$ is the error term (σ is a variance based on the error). Next part of the ARIMA model is the integrated process $I(d)$. Target processes can be affected by cumulative effects of some processes and to eliminate non-constant trend and seasonality differencing is used with parameter d , which describes the order of differentiation of the time series. Moving average $MA(q)$ is part of the ARIMA model which describes dependency between an observation and the residual errors resulting from the application of a moving average model to lagged observations and can be described as

$$Y_t = \theta_0 + \sum_{i=1}^q \theta_i Y_{t-i} + \omega_t \quad (7-2)$$

where θ_0 is constant term, θ_i are moving average parameters and $\omega_t \sim N(0, \sigma^2)$ is the error term. Since ARIMA model comprises $AR(p)$ and $MA(q)$ components, the model can be defined as

$$Y_t = \phi_0 + \sum_{i=1}^p \phi_i Y_{t-i} + \epsilon_t + \theta_0 + \sum_{i=1}^q \theta_i Y_{t-i} + \omega_t \quad (7-3)$$

7.3.2 LSTM

LSTM is a special architecture of recurrent neural network (RNN) designed for modelling long-term dependencies of sequences. In the standard architecture of LSTM networks, there is an input layer, a recurrent LSTM layer and an output layer. The input layer is connected to the LSTM layer. The recurrent connections in the LSTM layer are directly from the cell output units to the cell input units, input gates, output gates and forget gates. The cell output units are connected to the output layer of the network. Using LSTM, time series forecasting models can predict future values (e.g., CPU usage) based on previous, and sequential data. In this section, the LSTM model is used to evaluate the vRAN workload prediction. The mathematical modelling of LSTM is explained in (Flizikowski et al., 2022).

7.3.3 N-BEATS

N-BEATS is a deep neural architecture based on backward and forward residual links and a very deep stack of fully connected layers. The general idea is flattening multivariate time series data to one dimension data (e.g., CPU usage) with a stack of blocks of fully connected layers. The model comprises a sequence of stacks, while each stack is a combination of multiple blocks, which connect feedforward networks via forecast and back cast links.

7.4 EXPERIMENTAL SCENARIOS AND RESULT ANALYSIS

The EMDC architecture allows inclusion of semantic description of use cases. Moreover, various policies can be defined for EMDC node operation under changing context of execution. This way certain steps will be taken for the design of the EMDC to assure the following targets: (i) use case providers together with vRAN developers will co-design a purpose-build, higher level KPIs, aggregating detailed characterization and traffic flow of use cases, (ii) to capture cross-dependencies between space/time dimensions of a use case at the level of networking infrastructure.

Designing vRAN deployment for AI/ML empowered edge will consider capabilities for intertwining the slice design (network and service descriptors) with the EMDC semantic descriptors and especially service (use case) ontology. Slice design represents the process of adjusting vRAN configuration to the needs of a use case (traffic profile). The use cases' semantic description will be capturing traffic flows/demands/statistics with the indication of service requirements for successful service level agreements (SLAs) and where/when data can be accommodated by the network (in which cases, for which hardware, for which scenarios, etc). Statistical models of the traffic (and variants thereof) should be able to be built and provided to the orchestrator (backed-up by workload predictor / resource manager blocks). Modelling cross relations between workloads (e.g. an EMDC use case components vs vRAN SW components) will be performed in the integration stage, by considering optimal

alignment of multiple layers, such as semantic models of use cases as well as the network infrastructure (application/security/resource) oriented, policy-based network management mechanism and policy unification, and AI/ML models meta-descriptions with workflow descriptions.

The EMDC is expected to provide various levels of maturity for vRAN deployment to handle data at scale. The initial deployment considers the following steps: (i) vRAN is provided as container-based Kubernetes cluster including: core network, centralised unit (CU), distributed unit (DU), and physical layer; (ii) at this stage, the “baseline approach” for data handling will be prepared and the mechanisms for resource management will mainly be configured during the integration stage to suit the use case characteristics. The next stage deployment will follow a dedicated set of implementation steps to match with the proposed “extended approaches” for data handling. The following features for the implementation will be considered; (i) vRAN orchestration and resource management demand to have access to a semantic description of a use case; (ii) there should be KPIs/metrics together with rules that regulate their relation to underlying infrastructure of use case including sensors/actuators and network requirements; (iii) assurance of security requirements for a use case to highlight not only performance but the relation of performance and security goals (and trade-offs) of a use case depending on the completeness and shape of the ontologies of use cases. An important aspect and the foreseen outcome of the smooth integration between 5G vRAN and EMDC would be flexibility of disaggregated workload placement for the vRAN.

However, in order to monitor the decision-making process and handling the applications, a data collection framework needs to be designed. In the following section, a data collection framework implemented in a local test bed is presented. A practical deployment framework is considered which demonstrates 5G vRAN components deployed as the Kubernetes pods in the EMDC. In order to provide access to a comprehensive set of metrics of 5G vRAN, a framework based on Prometheus [289] exporters, Grafana [290] for visualisation and InfluxDB are utilised.

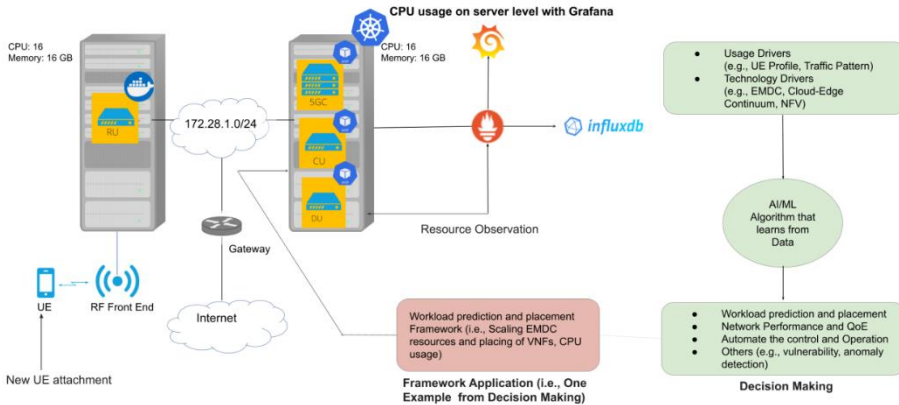


Figure 97 Data Collection Framework

With such instrumentation it is possible to accurately profile resource consumption of both (i) computing and (ii) radio metrics. Based on such profiling various AI/ML models (e.g., ARIMA, LSTM, and N-BEATS) can be trained in order to be able to predict resource demand. The details of this demonstration are described in [291]. In our experiment, the scenario model consisted of 6902 observations made each second, which is almost 2 hours of CPU usage from vRAN.

Figure 97 shows a data collection framework which demonstrates practical deployment of a 5G vRAN as the Kubernetes workload in the edge server represented by two legacy computers (desktop PCs). The mobile network is a full-featured 5G packaged into SW components according to the functional splits defined by 3GPP. Besides the general-purpose computers, the only specialised HW is the radio head device represented by the USRP node. The radio stack software (vRAN) is split into the RU, DU, CU and core components. The RU, DU and CU are hosting the following functions the PHY-Low (RU), PHY-High (DU), MAC (DU), RLC (DU), PDCP protocol layer (CU). The lower down the radio stack towards the PHY layer (RU), the tighter are requirements for interfaces' latency and throughput. Owing to above mentioned architecture an important benefit is the flexible control in migrating or scaling selected vRAN SW components along the edge-cloud continuum.

Such scaling and migration features are identified in the 3GPP specs defining slicing mechanism (3GPP SA5 with TS28.531, TS28.526; ETSI GS ZSM with ZSM003). All the yellow boxes in figure 2, except for RU, can be subject to placement in various partitions of the access and metro network, assuming the required throughput/latency requirements are met. In order to provide access to a comprehensive set of metrics of 5G vRAN a framework based on Prometheus exporters, Grafana for visualisation and InfluxDB are utilised. With such

instrumentation it is possible to properly profile resource consumption of both (i) computing and (ii) radio metrics. Based on such profiling various AI/ML models can be trained to predict resource demand (e.g. CPU consumption, memory, throughput, etc) based on the observed user traffic evolution in time. The collected metrics are 5G vRAN performance measurements on uplink and downlink and resources consumption observation.

7.5 DATA USAGE FOR WORKLOAD SCHEDULING IN EMDC

To support data handling in such edge networks where high level of flexibility of load placement and migration are essential, vRAN profiling sessions were executed. During such profiling sessions key focus was mainly on the CPU usage, depending on the number of users. In the Figure 98 the 5G DU CPU consumption for the period of 2 hours is presented. This scenario was used to train the long short term memory (LSTM) model to forecast CPU consumption. The top figure presents the DU unit CPU metric evolution in time. The time duration of a session is 2 hours. It can be seen that variation of CPU is relatively high at times and can reach beyond the 100 percent in case 2 UEs are actively using the YouTube sessions. The bottom plot on the other hand demonstrates the time evolution of 2-minute CPU readings when the CPU usage is evolving towards exceeding the maximum thresholds. Owing to KPI metrics time evolution it is possible to apply AI/ML models like LSTM, ARIMA (transfer learning), N-Beats [212] [286] [287] or similar to build prediction models. With such prediction models it is then possible to properly scale vRAN resources depending on the user's traffic envelope changes. In our case scaling is considering the horizontal scaling of CU between edge servers. This way PDCP related, encryption based PDU processing can become more adapted to traffic changes according to a typical tidal effect (e.g. traffic volume changes during day or week).

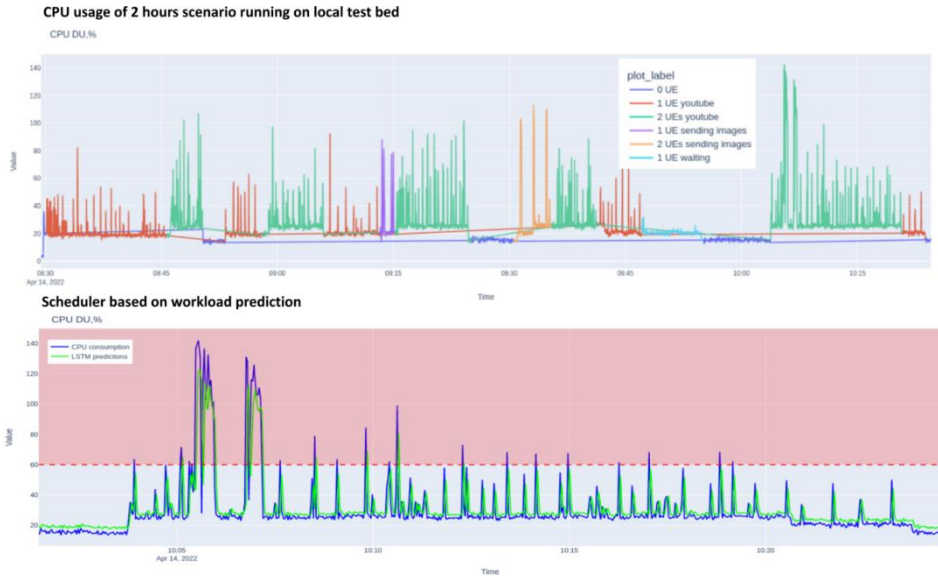


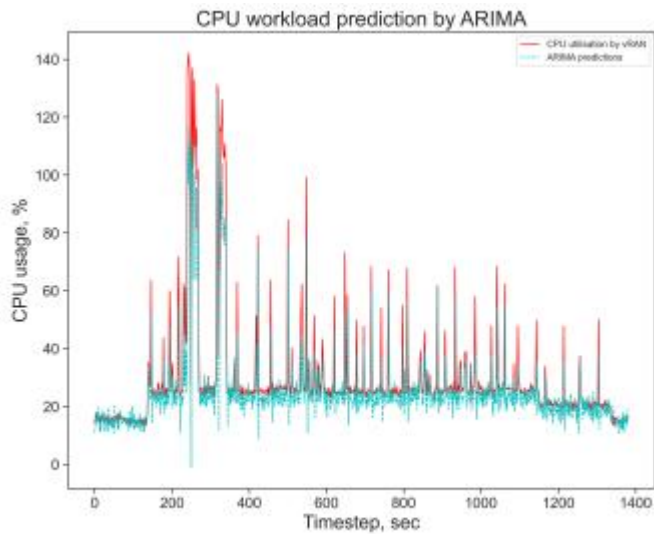
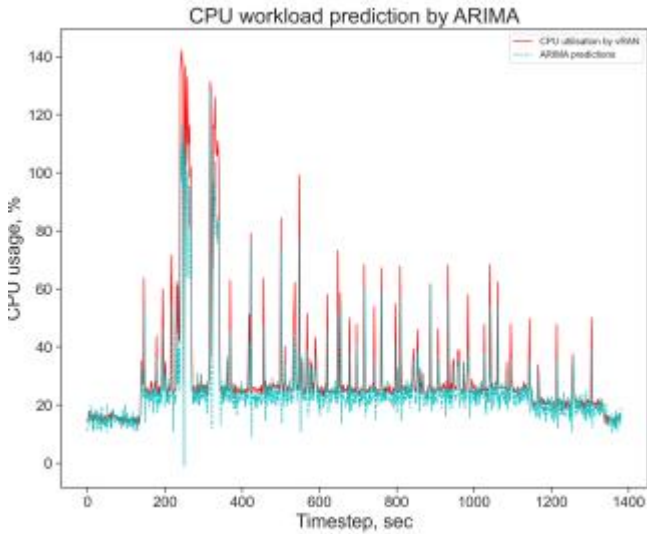
Figure 98 Data Usage for Workload Scheduling

This scenario observed several metrics such as throughput, primary resource block (PRBs), UEs connection, sending and receiving data etc. Different size data pictures were sent via email for transmitting purposes and also ran different quality Youtube videos ranging from 720p to 1440p for receiving purposes. As a result, the dataset from running the vRAN in this experimental setup was obtained. Obtained dataset was split to train and validation in 80:20 ratio, thus train data comprised 96 minutes of observation and for validation there was 24 minutes observation of CPU usage. For this experiment, a 5G network deployed on the EMDC was used which consists of servers on different network configurations.

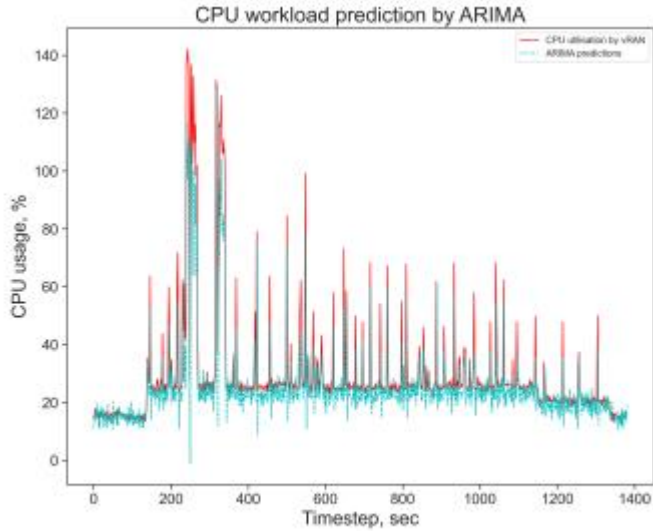
In the EMDC, the centralised unit (CU) and distributed unit (DU) can be deployed as VNF on one server with Kubernetes cluster and radio unit (RU) connected as USRP (i.e. RF front end) on another server, where a physical layer is deployed as Docker Swarm [292]. A commercial UE was considered to establish connectivity for data sending and receiving. Based on the collected data, several ML models for workload prediction were built and results described in the following sections. In this section, performance of the proposed ML-based workload prediction algorithms was evaluated by collecting the data from the experimental setup. All results were generated by using the PyTorch library in Python [292] where ML-based models are trained and tested based on the data collected from the real-time test scenarios running in the testbed.

In what follows author describes the modelling process of the ARIMA. Firstly, obtained data from the experiment were tested on auto correlation and stationarity

to train the ARIMA model.



(b) LSTM model



(c) N-BEATS model

Figure 99 CPU workload prediction for proposed ML models

Applying Dickey-Fuller test [293] to the data, author concludes that our series is stationary, thus order of differencing is set to $d = 0$. For the ARIMA model, the autoregressive parameter p and moving average q as 2 and 20 were selected. As a result an ARIMA model with parameters (2, 0, 10) was obtained and it was trained on the training dataset. On each step the model was trained with 20 historical observations and then trained models predicted future CPU usage. At the same time weights of the trained model were adjusted based on current observations. In other words, transfer learning of the ARIMA model is an iterative process in which the model is retrained after prediction. Figure 99a presents the prediction on validation dataset of CPU utilisation by vRAN. With transfer learning approach performance of the model is much better than predictions based on the model without retraining.

The advantages of LSTM over recurrent neural networks (RNN) is the ability to “forget” or to “take into account” long-term dependencies. For training the LSTM model dataset was processed with a sliding window technique where the window was set to 151 records (2.5 minutes). After the transformation, the train dataset had 6599 windows with 151 records each and the validation dataset transformed to 1381 windows with the same number of observations. Before training initial weights of the model were provided with uniform distribution $U[-0.08, 0.08]$. For this modelling, the hyper parameter values are as follows: epoch to 600, learning rate to 0.03, hidden states to 55, recurrent layer to 1, and dropout to 0.1. Figure 99b shows the workload prediction for LSTM models. Performance of the LSTM model is quite lower than the ARIMA model with transfer learning, but the LSTM

model has advantage in application and deployment on the EMDC and it does not require additional computation resources in comparison with transfer learning with ARIMA. To train this model a multivariate time series data were used, which in addition to CPU usage parameter comprises number of connected user equipment (UEs) presented as categorical variables. In the first stage of N-BEATS training “vanilla” model was trained to find optimal initial learning rate. From the test it was obtained that initial learning rate equals $21 \cdot 10^{-E04}$. For other hyper parameters, the values are set as follows: number of blocks to 2, number of fully connected layers to 4, size of fully connected layers to [300, 2048]. Workload prediction of CPU usage by N-BEATS model presented in Figure 99c and its result is the worst among three models. To measure the accuracy of the specified models, the mean absolute error (MAE) and mean absolute percentage error (MAPE) were followed. MAE indicates the difference between predicted value and true value of an observation, while MAPE defines the statistical measurement accuracy of a ML algorithm for a fixed dataset. Both terms are referred to as a loss function for defining error by the model evaluation. To reflect the effectiveness of the ML models, the experimental values of MAE (5.09, 6.40, and 14.21) and MAPE (0.14, 0.20, and 0.38) in case of ARIMA, LSTM, and N-BEATS were acquired respectively.

7.6 MODIFIED APPROACH TO CPU/ENERGY CONSUMPTION PREDICTION

The work below introduces architectural updates of the previous section in order to analyse what is the influence of LSTM architectural updates on the provided performance metrics of the prediction model. Figure 100 shows the block diagram for LSTM1 model. The CPU time series data is first pre-processed and parsed into segments using a moving window with a step of one time step.

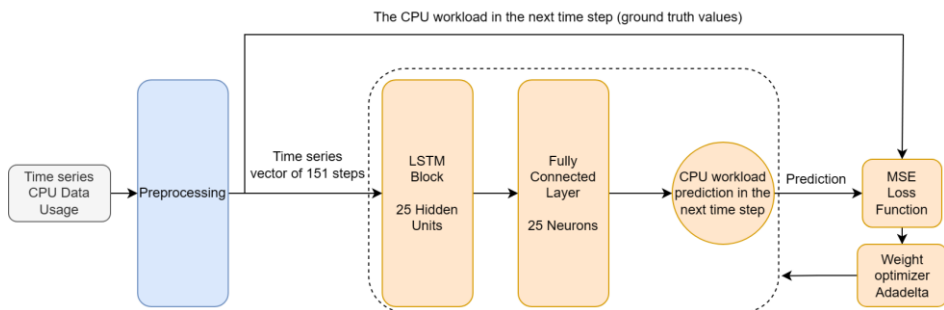


Figure 100 LSTM1 block diagram

Each generated segment has 151 time steps and each step represents one second. The next time step (152) is stored as the target variable. In order to perform training the dataset is split into 80% train and 20% test, then normalized to be

from -1 to 1. The normalized train set segments is fed to the model as in batch form. Then, the training of the model starts. The optimizer used is Adadelta with learning rate 10^{-3} . In each epoch, the model predicts the cpu load in the next time sample for the training data segments. The output is compared to the ground truth values of the CPU workload, then the weights are adjusted accordingly. The model is trained for 800 epochs. The train/validation curve is shown Figure 101.

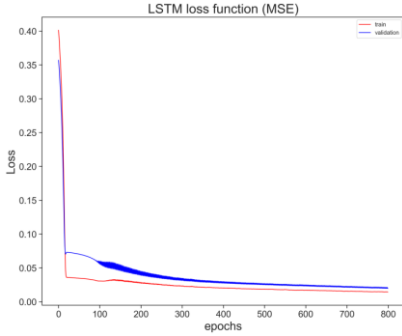


Figure 101 Train/validation curve for LSTM1

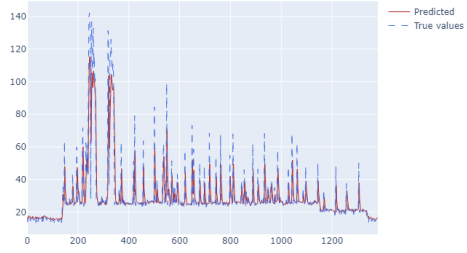


Figure 102 LSTM1 prediction versus true CPU workload.

It can be seen that both the train and validation mean squared error decrease as epochs increase with the train curve being slightly better the validation curve. In the test scenario, the trained model uses the segments from the test set (of length 151) to predict the next CPU time step at the 152th time step. Figure 102 shows the prediction from the LSTM1 model for the test set compared to the actual true values. It can be seen that LSTM1 reasonable performance of predicting the CPU workload and the RMSE for the test set is approximately 9. Figure 103 shows the architecture for the LSTM2 mode.

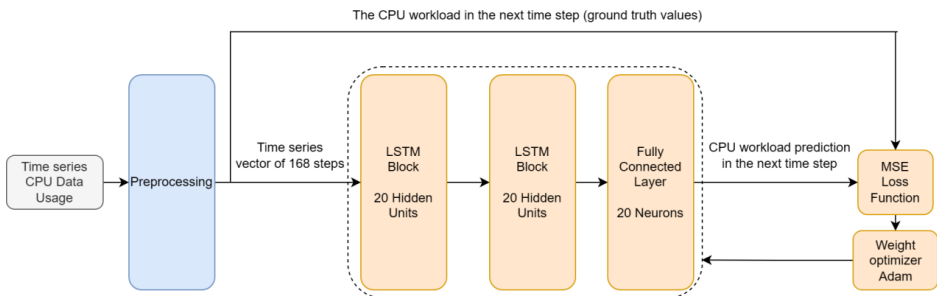


Figure 103 LSTM2 model architecture.

In the second case the CPU time series data is first per-processed and parsed into segments of length 168 with no overlap. The next time step (169) is stored as the

target variable. The dataset is split into 80% train and 20% test, then normalized to be from -1 to 1. The normalized train set segments is fed to the model as in batch form. Then, the training of the model starts. The optimizer used is Adam with learning rate 10^{-3} . In each epoch, the model predicts the cpu load in the next time sample for the training data segments. The output is compared to the ground truth values of the CPU workload, then the weights are adjusted accordingly. The model is trained for 800 epochs. The train/validation curve for LSTM2 is shown in *Figure 104*.

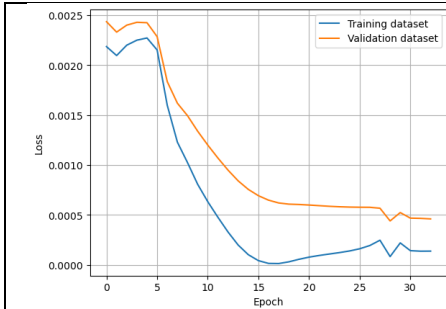


Figure 104 Train/validation curve for LSTM2

Datetime	Minute Second	
	Minute	Second
2022-04-14 08:29:28	29	28
2022-04-14 08:29:29	29	29
2022-04-14 08:29:30	29	30
2022-04-14 08:29:31	29	31
2022-04-14 08:29:32	29	32
...
2022-04-14 09:49:54	49	54
2022-04-14 09:49:55	49	55
2022-04-14 09:49:56	49	56
2022-04-14 09:49:57	49	57

Figure 105 Time index for each CPU workload reading

It can be seen that both train and validation MSE decrease as the epochs increase with the train MSE being slightly better than the validation MSE. The trained LSTM2 model is then used for CPU workload prediction on the data set. Segments of length 168 are fed into LSTM2 to predict the CPU workload in the 169th time step. The performance of the LSTM is compared to other regressors like decision tree and linear regression. These regressors are trained on a time index-based data set as in Figure 105. The input is the time of each CPU workload reading/recording in minutes and second, and the target is the CPU workload value. Finally, Figure 106 shows the prediction of LSTM2 and the other regressors based on the test set.

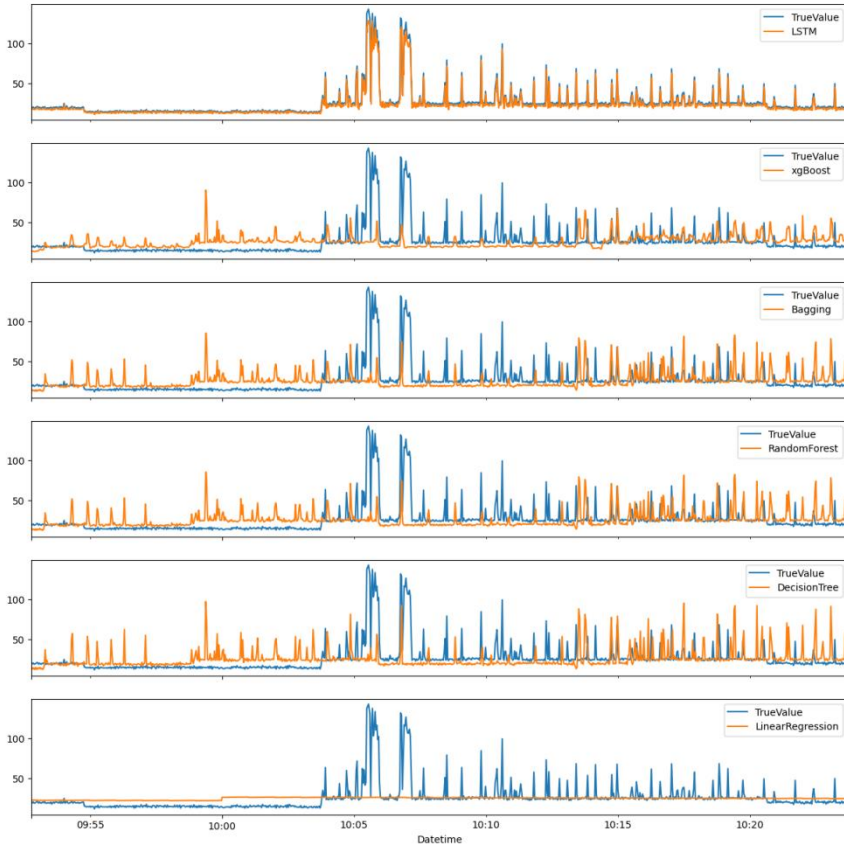


Figure 106 CPU workload prediction for different models

It can be seen that LSTM2 accuracy is highest at predicting the CPU workload, while the other regressors are not very effective. Table 37 shows the RMSE for these different models based on the test set.

Table 37 Comparison of the RSME of different predictors

Model	RMSE
LSTM1	9
LSTM2	4.6
XGBOOST	18
Bagging	19
Random Forest	19

Decision Tree	20
Linear Regression	16

7.7 APPLICABILITY OF THE RESULTS INTO THE 5G VRAN (AND BEYOND)

The architecture where above proposed solutions for prediction workloads could be further evaluated is the 5G open-RAN network, where prediction can be utilised to support admission control (Figure 107). The scope of such functional extension is related to the capability of offloading a certain part of the disaggregated vRAN to an external edge server, to relieve some computational resources of a current edge server as the traffic demand is growing and current resources are not enough. This feature assumes that offloading the CU-UP is the first step towards future more comprehensive “dynamic offloading” of vRAN elements (DU, RU, RIC, others). It is foreseen to happen by providing horizontal scaling of vRAN between the different edge servers (EMDCs) or between edge and cloud servers respectively. The role of scaling in more efficient RRM control is to assure: (a) offloading of selected 5G vRAN components to another server (edge or cloud), (b) combining it with prediction in order to be able to scale proactively - i.e. more smoothly and avoiding reactive approach. The key aspect to be achieved is the ability of dynamic creation of CU-UP that is instantiated on another physical node (edge server, cloud server). Once CU-UP is instantiated the local instance of the 5G radio stack (virtual containers) will instrument this remote CU-UP to be ready to handle user traffic on demand. But the “redirection” of selected UE traffic to this remote CU-UP will happen only at the stage of “new UE admission” and not during the session. The summary of benefits behind the application of auto-scaling solutions for disaggregated vRANs in 5G and 6G are:

- Scaling mechanism is an obligatory mechanism to maximize the virtualisation potential in the cloud native deployment and address more business scenarios for the vendors. The mechanism behind scaling is key in terms of:
 - availability - scaling will introduce new capacity for users without tearing down the whole vRAN (and hence causing down-time for existing users)
 - dynamic, on-the-fly RAN component redistribution/migration (RAM morphism) - mechanism to orchestrate distributed RAN components at run-time, developed in this feature, will be essential to the O-RAN 2nd wave evaluation.

Additional reason for such offloading can be that other server may provide availability of additional means to improve energy efficiency by applying acceleration hardware that can be utilised to offload some CPU for the “PDCP encryption” functionalities. Such accelerator can be for example the SmartNIC

(or other). The main offloading triggering function is the admission control (service), which will be augmented with capabilities to predict user traffic demand in future and indicate - when the currently available HW resources and/or radio resources will not be sufficient - i.e. when the resources required by vRAN will reach certain threshold. Here the target is to be able to redirect CU-UP processing of certain users' data flows (data bearers) to another machine (edge server) that has extra resources for this purpose. But this feature does not require "dynamic and fluent" adaptations of CU-UP with the ongoing connections but it is enough to have changes applied iteratively per "new connection request" level (i.e. admission decision) and not "during connection" of the user sessions (i.e. fluent vRAN scaling). The ultimate scaling of "any vRAN components", will be targeted by the future research.

In the ultimate shape the admission control and the orchestration (e.g. ETSI MANO, K8, VC) should be cooperating in order to identify optimal strategies for offloading any number of required components (e.g. RU, DU, CU, RIC, core, etc). Concerning the transport network between edge servers - the strong assumption here from EMDC compliant server side is, the perfect control over the links (QoS) between edge devices, which will be the SDN-based fiber link. Such link provides very fine-grained capabilities to reserve dynamically certain share of the transport network capacity in order to support the process of dynamic CU-UP scaling. This way it is completely feasible to send even some container images between edges on demand, or just download them (i.e. CU-UP binary and dependencies) from containers repository.

To achieve the above defined scope of functionalities, there is the need to have an efficient telemetry system that will be collecting required statistics (radio statistics, resource consumption) and sending it to the workload predictor. Predictions of resource utilisation are fed into the placement agent. With high resource demand observed, the placement agent should be able to scale RAN components between EMDCs (these EMDCs are connected with fiber links, and the links capacities are managed with SDN controller of the fiber network). In other words, the transport network between EMDCs are controlled with a SDN controller that can establish connection with another server, and particular QoS/capacity, and this way support the process of deploying there some processes (e.g. CU-UP1A in the Figure 107).

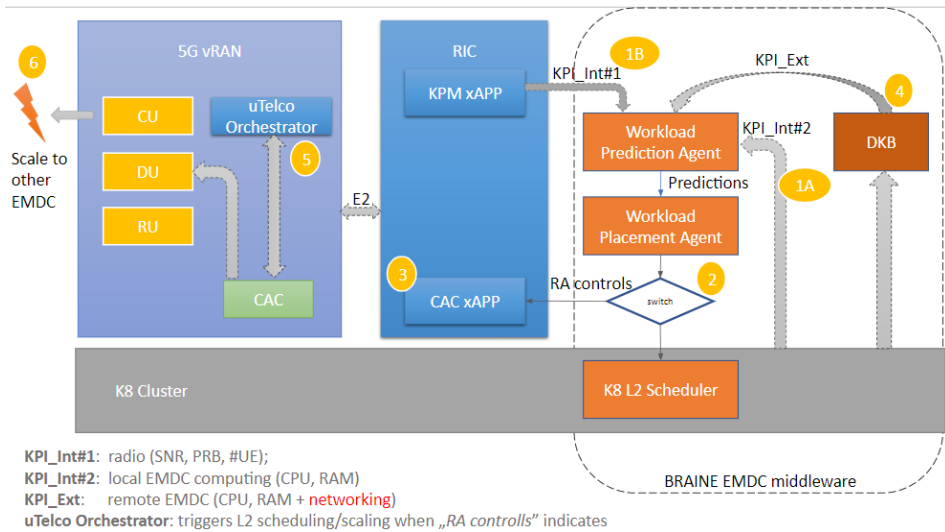


Figure 107 Proposed architecture enabling inclusion of the workload prediction and placement CAC agent for EMDC deployment

The flow of events in the figure above is as follows: (1A) first the collection of data is executed considering the local resources of the underlying EMDC HW (e.g. CPU, RAM), (1B) the custom based collection of radio interface statistics are collected by the RIC xApp (key performance monitoring) and provided over the ORAN A1 interface towards the earlier described workload prediction agent. Once the agent has access to the measurements (1A, 1B) it is capable of applying the models described in sections above (7.5, 7.6) in order to provision a relevant model that will help building the contextual understanding of the evolving future horizon with respect to expected resource consumption in next k-steps. In the point (2) of the figure it is highlighted that Workload Placement Agent would need to implement an appropriate algorithm that would support decision on selecting the most. The simple algorithm here is to utilize the up-to-date information provided by a distributed knowledge base implementation (4) and utilize it in order to be able to select the alternative server (e.g. EMDC2) which has enough computing power in order to deploy (place) the workload of the 5G ORAN based “CU-UP” component there. This action would require collaboration with the CAC algorithm, that is plugged into a RIC xApp (3) in the figure. Based on the cooperation between the CAC algorithm and the placement algorithm it will be possible to indicate to the micro-orchestrator of the underlying 5G ORAN based radio stack to perform the action of resource scaling on-demand (5). This will be possible to execute in advance based on the k-step prediction using the model as defined in the sections above (7.5, 7.6).

7.8 USE CASES FOR CLOUDIFIED, VIRTUALIZED AND DISAGGREGATED RAN

In relation to the disaggregation and scaling of a virtualized RAN (SD-RAN) discussed in this chapter, author presents here most likely use cases that highlight the need for cloud-edge-continuum solutions for such vRAN deployments:

- **UC1: secure 5G node with EMDC:** In this use case, SD-RAN deployed on an edge node is empowered with secure solutions to assure security level of E2E applications as well as securing edge-to-edge or edge-to-cloud communications. Solutions considered include quantum key distribution, distributed ledger technologies for local IoT as well as cryptography algorithms offloading to a smart NIC. In this case, disaggregation of SD-RAN would be needed in order to provide more flexibility in: (a) orchestrating various radio protocols within edge node (e.g. deploying medium access control (MAC) to the graphics processing unit (GPU) accelerator), (b) orchestrating some protocols of the same SD-RAN between edges, in order to exploit ultra fast fiber communication between edge nodes (e.g. coordinated at the level of SDN controller and RIC) or (c) workload optimization based on latency requirements where the latency could be relaxed by moving some workloads even to the cloud (e.g. packet data convergence protocol (PDCP) for acceleration).
- **UC2: utilizing SD-RAN with local breakout (LBO):** In this use case, the idea is to provide multiple instances of e.g. CU to match with the different user groups (e.g. premium, best effort). Representative example could be either video streaming from the internet or online gaming, in which case the “application server” can be (a) in the Internet or (b) deployed in the local edge server or even coexisting with SD-RAN (i.e. via local breakout). Here the disaggregation of SD-RAN could be focused on creating multiple instances of CU protocols. The “CU” of best-effort users could be directly connected to the core network part and then through the backhaul relevant data streams would normally be delivered to the Internet. Whereas the CU protocols for premium users could be deployed in a way that allows bypassing the core network and the Internet and instead go to the local content server. Naturally the latter approach would result in significantly lower delays, higher throughput and lower variation of both.
- **UC3: aggregating multiple instances of DU/CU in EMDC:** In this use case, one could be interested in maximizing gains of localizing the processing of multiple instances of SD-RAN in a single EMDC. In this case the key focus would be to assure sharing as many resources as needed (RAM, CPU, and storage) in order to speed-up the data exchange between CUs and/or DUs. This use case could be representative for implementing novel paradigms of network management towards 5G and beyond networks.

7.9 SWOT ANALYSIS OF SD-RAN DEPLOYMENT

The use cases presented in the above subsection significantly benefit from an additional support of cloudification and virtualization. More of such use cases can be further generated depending on the requirements of users as well as strategy of operators and availability of infrastructure. The key aspect here is to what extent the cloud-edge continuum solutions are required or even allowed considering the perspectives of security or latency. In some cases, pure-edge based deployment would be most preferable (e.g. UC1) whereas the equal share of cloud and edge looks optimal to properly support the UC2. It is important to consider various aspects when deciding about such deployment, and one of the most relevant decision-making nodes could be an orchestrator (e.g. MANO, Kubernetes) - a detailed description of orchestrators is in Section IV. The essential to notice here is the optimal split of responsibility between orchestration / NFVI / local-SDN-controller (e.g. RAN Intelligent Controller, fiber network SDN controller), especially that certain functionalities may be overlapping (or spilling out of one entity towards another). Intelligent functionalities of orchestrators can be customized and so improve value added services, e.g., (a) utility-based workload deployment in multi-cloud environment; (b) shared-knowledge based workload placement inside EMDC. Operators of future networks (e.g., B5G and 6G) would be benefited from flexibility in deploying various workloads in an adaptive and intelligent way considering various criteria for making the deployment optimal from the perspective of cost, security, and real-time performance. Considering the sole-perspective of SD RAN deployment, the degree of disaggregation for the radio protocols will further specify degrees of freedom when identifying optimal placement of RAN functionalities (e.g., L1, L2 etc). For the future deployment of RAN functionalities, overviewing the strength, weakness, opportunities, and threats (SWOT) is required to reflect the full benefit of the network performances. An Example of SWOT analysis for SD-RAN deployment in cloud, edge, and cloud-edge continuum is presented in Table 38. Some further consequences of following the particular of the options presented there are listed below: (i) Option1 uses cloud, and as consequence resources in the cloud are partitioned “as they are” so the level of granularity is communication service providers (CSPs) defined for the available node configurations. While in the Option2/Option3, the nodes capabilities partitioning is available under more flexible policies (ii) in order to deal with cloud resources of Option1, there are some security breaches: (i) images needs to be delivered to cloud over internet (i.e., slow and insecure) (ii) multi-cloud management platform has to ensure multiple ports opened to interact with various clouds (CSP) (iii) Option1 can utilize some abstraction of the “Orchestrators/virtual infrastructure managers (VIMs)” in the form of the scheduling abstraction layer (SAL) [294]. The SAL needs to be collecting available resources and their characteristics prior to workload deployment. The direct interaction with a cloud happens via Resource Manager (API), that registers all the available resources and presents

them to the SAL.

Table 38 SWOT analysis of SD-RAN deployment in Cloud-Edge SWOT

SWOT	SD-RAN in Cloud (Option1)	SD-RAN in Edge (Option2)	SD-RAN in Cloud-Edge Continuum (Option3)
	<ul style="list-style-type: none"> • Fine grained computing infrastructure • Multiple deployment options • Centralized security monitoring and attack prevention/detection capabilities 	<ul style="list-style-type: none"> • Data is more secure as it stays “local” to the users • Tight control of the network infrastructure • Small-footprint solutions by tailoring existing infrastructure 	<ul style="list-style-type: none"> • Flexible management of secure workloads • Real-time network functions can be deployed on edge while the non-RT in the cloud • Intelligence can be deployed “on both ends” or smoothly moved where needed
Strengths	<ul style="list-style-type: none"> • Smart offloading to multiple clouds (e.g. based on utility function) • Decreased requirement for local processing platform • Less geographically bounded computation 	<ul style="list-style-type: none"> • Lower barrier of entry for local edge providers • Limited data transition to/from the cloud • AI/ML based learning at a edge node or swarm intelligence when multiple edges are combined • Benefits of local 	<ul style="list-style-type: none"> • Learning based on data can happen incloud while use of the models in the edge • Deployment of workload can be based on prediction and utility function

SWOT	SD-RAN in Cloud (Option1)	SD-RAN in Edge (Option2)	SD-RAN in Cloud-Edge Continuum (Option3)
		processing alignment: MEC, caching, LBO	
Opportunities	<ul style="list-style-type: none"> • Network infrastructure to the cloud and between clouds with limited QoS guarantees • Due to data transmission to/from the cloud with higher risk of malicious activity • Low-latency applications (E2E) not feasible • exchanging large volumes of data over internet largely depends on access network quality 	<ul style="list-style-type: none"> • Investments are more distributed 	<ul style="list-style-type: none"> • More energy footprint are required
Weakness	<ul style="list-style-type: none"> • Data needs to be delivered over susceptible global internet 	<ul style="list-style-type: none"> • Cloud operators deploying edge solutions within communications service provider • Application servers running in e.g. Wavelength Zones without leaving the telecommunications network 	<ul style="list-style-type: none"> • Big data security • Application privacy

In Table 39, author has collected the following elements:

- Technique identification - to mention the particular mechanism
- 5G and B5G factor of influence - to investigate what aspect from B5G will be mainly influencing a particular technique for improvement and optimizations of results

- Foreseen solutions - to identify directions for network technique's modernization
- Challenges - to check what additional considerations are considered when planning an introduction of particular optimizations and improvements.

As can be seen, many techniques addressed in the Table 39 represent key decision-making components that are influenced by the B5G evolution. To assure smooth and data-driven operation of future networks the importance of modernizations at various stages can be seen. Among solutions considered (proposed by the authors based on the prior state of the art), we highlight the following:

- Bring learning and optimization by design
- Apply smart optimizations for accessing MEC
- Apply smart models for prediction at various layers
- Consider semantic fusion of infrastructure services and applications
- Apply AI/ML workload placement techniques
- Cross-domain optimization of resources
- Apply AI/ML to decrease the amount of required measurements
- Weak coupling theory.

Table 39 Commonalities and cross dependencies between RRM mechanisms and the workload prediction and placement techniques

Technique	5G and B5G factor of influence	Foreseen solutions	Challenges
Multiple-access schemes	<ul style="list-style-type: none"> • Density of networks • Density of UEs (e.g. mMTC) calls for more radio resources 	<ul style="list-style-type: none"> • Utilize remaining resources in non-orthogonal way (NOMA) 	<ul style="list-style-type: none"> • NOMA/OMA coexistence strategies • User clustering • Error propagation in SIC
Scheduling	<ul style="list-style-type: none"> • Sub-ms scheduling • Multiple factors need to be considered (e.g. energy efficiency, QoE, latency constraints, computation offloading) • Multi-RAT mechanisms 	<ul style="list-style-type: none"> • Multi-objective joint RRM and computation optimization • Apply cell-free in the network architecture • Bring learning and optimization by design (e.g. by means of ORAN RIC) 	<ul style="list-style-type: none"> • Computational complexity of 3D scheduling • Cell-free demands robust telemetry • Distributed processing brings new challenges • SD-RAN architectures requires network

Technique	5G and B5G factor of influence	Foreseen solutions	Challenges
	<ul style="list-style-type: none"> • Carrier aggregation • High demand for energy efficiency 		tailoring (non standardized yet) <ul style="list-style-type: none"> • SD-RAN architecture development need
MEC	<ul style="list-style-type: none"> • Greater push for services local breakout • Growing services mobility due to ultra-density of network antennas and UEs 	<ul style="list-style-type: none"> • Improve cross-domain orchestration with VNO • Apply smart optimizations for accessing MEC (e.g. with NOMA as alternative channel for offloading) 	<ul style="list-style-type: none"> • Efficient means and techniques to orchestrate workloads under multiple orchestrators (e.g. MEAO, VNO, transport SDN controller, fiber resources SDN controller) • Efficient interaction between RIC and MEC entities
Admission control	<ul style="list-style-type: none"> • Increase in handovers due to ultra-density and number of UEs moving • Multiple dimensions of resources to be considered (radio, computation, energy, spectrum, cross-domain and E2E slice resources) 	<ul style="list-style-type: none"> • Apply cell-free to remove handovers under umbrella of cooperative resource allocation • Apply smart models for prediction at various layers • Consider semantic fusion of infrastructure services and applications (use-cases) meta-data • Apply AI/ML workload placement techniques aligned with underlying 	<ul style="list-style-type: none"> • Understand the influence of multi-dimensional scheduling on CAC (new research needed) • Architectures are necessary at edge to further integrate and consolidate resources at various layers into consistent KPIs • Redesign of OSS/BSS/MANO systems to reflect new architectures to be more service oriented • slice provisioning and management

Technique	5G and B5G factor of influence	Foreseen solutions	Challenges
		NFVI orchestrators	should be addressed
Architecture of 5G RAN / Core	<ul style="list-style-type: none"> • Demand for ultra high capacity and controlled latency • Distributed nature of future cell-free networks • Service based implementation • High increase in measurement data volumes • Further levels of RAN disaggregation • Optical-wireless coexistence 	<ul style="list-style-type: none"> • Apply cell-free • Cross-domain optimization of resources: radio, wired, fiber and assuming various splits in the network potentially even at same time • Orchestrate and compose services at various levels (slice, SFC) • Apply ML/AI to decrease amount of required measurements • Consider new levels of flexibility in protocol instrumentation • Assure mechanisms to allow smooth alignment of wireless-optical 	
Network management for sustainability	<ul style="list-style-type: none"> • hyper-connected network management with extended reliability and latency • self-sustainability for maintaining KPIs in highly dynamic network architecture 	<ul style="list-style-type: none"> • Distributed orchestration with concepts of O-RAN, MEC, ZSM, ENI • Weak coupling theory 	<ul style="list-style-type: none"> • Goal directed RRM • Conflict resolution

7.10 SUMMARY

The radio access network (RAN) is transforming to an open, virtualized, programmable, and intelligent RAN. Open-RAN (O-RAN) architecture, defined by O-RAN Alliance, is the solution to such evolution. The innovative RAN Intelligent Controller (RIC) in O-RAN is no doubt where intelligence is hosted. The goal of this chapter was to investigate the importance of vRAN workload prediction in the EMDC architecture. For this investigation, the performance of three time series forecasting algorithms we evaluated together with a CPU usage data obtained by running VNFs in the EMDC. The evidence from this experiment indicates that the LSTM model shows advantages over other models in terms of application deployment on the EMDC platform. Further research might explore the optimal trade-off for the tasks between complexity and performance.

8 LEARNING BASED CAC AGENT DESIGN

8.1 INTRODUCTION

In this section author describes and evaluates a MDP-formulated problem to find optimal Call Admission Control policies for wireless networks with adaptive modulation and coding. The two classes of service (BE and UGS-priority) are considered and a variable capacity channel with constant error bit rate. Hierarchical Reinforcement Learning (HRL) techniques are applied to find optimal policies for multi-task CAC agent. In addition, this chapter validates several neural network training algorithms to deliver a training algorithm suitable for the CAC agent problem.

Contribution presented in this chapter is as follows: i) implementing CAC algorithm using Reinforcement Learning to deliver optimal actions that maximize the reward function in wireless 4G/5G networks with AMC; ii) use of the “options” approach to define the problem in a variable capacity channel; iii) providing an ANN training algorithm validation framework in order to validate different training neural network algorithms (backpropagation) using training set of learned Q-table from the reinforcement learning algorithm based on own design.

8.2 SYSTEM MODEL

Call admission control mechanisms allow the service provider to control the traffic flow to the network and deliver proper service quality to end customers. A CAC agent is therefore responsible for proper resource assignment in networks with QoS control and the incoming connections should be handled according to their different QoS requirements. The main parameter used for the resource assignment is the air interface capacity. Here, an example of a 4G PHY layer as specified in [295] is considered. But the same approach can be followed for any other PHY layers (5G, beyond). According to the standard and authors in [296], the total air interface capacity can be safely approximated as

$$C = \text{floor} \frac{BWn}{8000} 8000 \frac{N_{used}}{N_{FFT}} \frac{1}{1+G} c_r b_m \quad (8-1)$$

System parameters used to calculate relevant capacity are presented in the table below.

Table 40 System parameters

Parameter	Value
BW	3,5 MHz

n	8/7
N_{FFT}	256
G	1/8
c_r	1/2, 2/4
b_m	2,4
C	2,9; 4,36; 5,8 Mbps

In equation (8-1) the BW is the system bandwidth, n is the oversampling factor, N_{FFT} described the total sub-carriers available for the system, N_{used} describes the total sub-carriers available for the user, G is the cyclic prefix, c_r and b_m is the coding rate and number of bits per modulation respectively. To keep the bit error rate at a constant level of 10^{-3} (CWER) wireless systems use a technique called Adaptive Modulation and Coding (AMC).

The system can adapt to varying channel conditions (i.e. change of modulation or coding scheme) increasing or decreasing its throughput depending on estimated bit error rate. Parameters c_r and b_m in equation (8-1) have different values according to the system's current state (bit error rate) and can take any values in the range presented in Table 40. The calculated throughput thresholds in 4G network with AMC can be therefore defined as a vector $C = (c_1, c_2, \dots, c_M)$ where $m = 1, 2, \dots, M$ and M is the number of modulation coding schemes (MCSs). In order to capture the rate at which a network changes its modulation coding scheme several simulations were performed for 4G transmissions over a flat-fading Rayleigh channel using Matlab simulation. The Reed Solomon Convolutional coding scheme was used and the MCS thresholds were obtained from equation (8-1). By gathering the network statistics and calculating transition probabilities a 4G link can be modelled as a Finite State Markov Chain [297]. Such approximation has been used to mimic the flat-fading Rayleigh channel's changing rate in radio interface. For sake of simplicity only two traffic classes are considered in the model, where one of them carries VoIP packets [37], [272]. Therefore our proposed 4G network, modelled as MDP, supports two classes of traffic - Best Effort (BE) with constant bit rate and sending rate at 200 kb/s and Unsolicited Grant Service (UGS) with constant bit rate and sending rate at 64 kb/s. Incoming call requests are following the Poisson process with rate λ_c , and call holding times are derived from the exponential distribution (with mean μ_c equal to 1). Each accepted call consumes resources equal to b_i , where i is the incoming call class. To further simplify the problem, it is assumed that incoming calls and calls already accepted in the system share the same MCS. Thus, when the modulation coding scheme changes it applies to all present and future calls that will arrive during the time slot the MCS will last. In the next section, the description of the CAC agent problem is extended by describing it as a Markov Decision Process.

To determine the bit rate for the modulation readings (MCS) of the i -th terminal, a formula was used for the data collected from field measurements, based on which a simple resource scheduling mechanism was implemented:

$$B_{effective}(x, t) = \left[OFDM_{UL,total} - OFDM_{oh}(t) - \sum_{i \in S_{active_SS}} OFDM_i \right] * MCS_i * MCS_i(t) * OFDM_{sub} * I \quad (8-2)$$

where $B_{effective}(x, t)$ – is the monitored available throughput of a x connection at time t , $OFDM_{total}$ – total number of symbols available in the TDMA frame for the upstream direction, $OFDM_{oh}(t)$ – class-of-service specific signaling overhead, $OFDM_i$ – number of symbols used by i -users active in the cell, $OFDM_{sub}$ – number of subcarriers, $MCS_i(t)$ – number of symbols for a given modulation, I – the intensity of the influx of frames (in frames/sec). As it can be noticed, the impact of overhead can be significant for the throughput ($OFDM_{oh}(t)$) for a particular traffic class. Based on the analyses in the ns2 simulator, the following average overhead values were collected: rtPS - 3.2 symbols/frame, rtPS - 1.6 symbols/frame and BE class user - 0.3 symbols/frame.

8.3 MDP MODEL

The state space S

$$S = \left\{ s | s = (x_1, x_2); x_1, x_2 \geq 0, \sum_{i=1}^k x_i b_i \leq C \right\} \quad (8-3)$$

Action space A

$$A_S = \begin{cases} a = (1,1), & \text{if } s = (0,0) \\ a = (0,0), & \text{if } x_i b_i = C \\ a | a = (a_1, a_2); a_i \in \{0,1\}, & \text{otherwise} \end{cases} \quad (8-4)$$

Call arrival event

Call arrival events are indicated by the current arrival time in vector T. Call arrival event are represented by the vector:

$$\left\{ e = (e_1, e_2) | e_i \in \{0,1\}, \sum_{i=1}^k e_i = 1 \right\} \quad (8-5)$$

Reward function

$$r(s, a) = \sum_{i=1}^k w_i (x_i + a_i e_i), \quad \text{and } w = [1,10] \quad (8-6)$$

When a system is in state $s \in S$ and the decision maker (CAC agent) chooses an

action from the available action state space $a \in A_s$. The agent takes only actions at decision epochs which in our case correspond to call arrival times indicated by the vector T. In return for taking an action the agent receives a reward $r(s, a)$ and evolves to a new state s' . The agent stores a value for counting the number visit in an state-action pair - $visit(s, a)$.

The agent follows a greedy policy and choose actions according to the following formula:

$$a = \underset{a'}{\operatorname{arg\,max}}(visit(s, a')) \quad (8-7)$$

The Q-value update rule is as follows:

$$\delta = r(s, a) + \gamma * \underset{a'}{\operatorname{arg\,max}}(Q(s', a')) - Q(s, a) \quad (8-8)$$

$$\Delta Q(s, a) = \frac{1}{1 + visit(s, a)} * \delta \quad (8-9)$$

$$Q(s, a) = Q(s, a) + \Delta Q(s, a) \quad (8-10)$$

The Q-values are stored in a Look-Up-Table for further calculations.

In the second approach the arrival and departure times have not been calculated beforehand but calculated the transition probability matrix to drive the transition between states. All the parameters are specified as in the section 8.2. The transition matrix calculation is following the equations [186]. The uniformization was applied to transform a continuous-time Markov Chain with **non-identical decision times** into an equivalent continuous-time Markov process. The following formula is defined for uniformization:

$$\tau_c = \left(\sum_{i=1}^k \lambda_i a_i + C * \max\{\mu_1, \mu_2, \dots, \mu_k\} \right)^{-1} \quad (8-11)$$

The sojourn time is calculated as follows:

$$\tau(x, a) = \left(\sum_{i=1}^k \lambda_i a_i + \sum_{i=1}^k \mu_i x_i \right)^{-1} \quad (8-12)$$

The expected holding time is the expected time until next decision epoch after action “a” is taken in the present state “x”. From equations (8-11) and (8-12) and the arrival and departure rates the transition probability matrix is calculated as follows:

$$P(y|x, a) = \begin{cases} a_1 * \lambda_1 * \tau_c, & \text{if } y = x + (1,0) \\ a_1 * \lambda_2 * \tau_c, & \text{if } y = x + (0,1) \\ x_1 * \mu_1 * \tau_c, & \text{if } y = x - (1,0) \\ x_2 * \mu_2 * \tau_c, & \text{if } y = x - (0,1) \\ 1 - \frac{1}{\tau(x, a)}, & \text{if } y = x \end{cases} \quad (8-13)$$

8.4 THE CAC AGENT DEFINITION

In this section the approach and results of implementing the RL-based CAC agent are presented. First the approach to verify and validate the approach are presented in the diagram in the Figure 108. After implementation of the RL algorithm the following system parameters were used, in order to perform a systematic approach to verification and validation of the created CAC agent.

Table 41 Modelled system settings (N=2 traffic classes) [153]

Parameter	Value
C	10
b_1	1
b_1	2
$\lambda_1 = \lambda_2$	1 (call/sec)
$1/\mu_{b1} = 1/\mu_{b2}$	500 sec

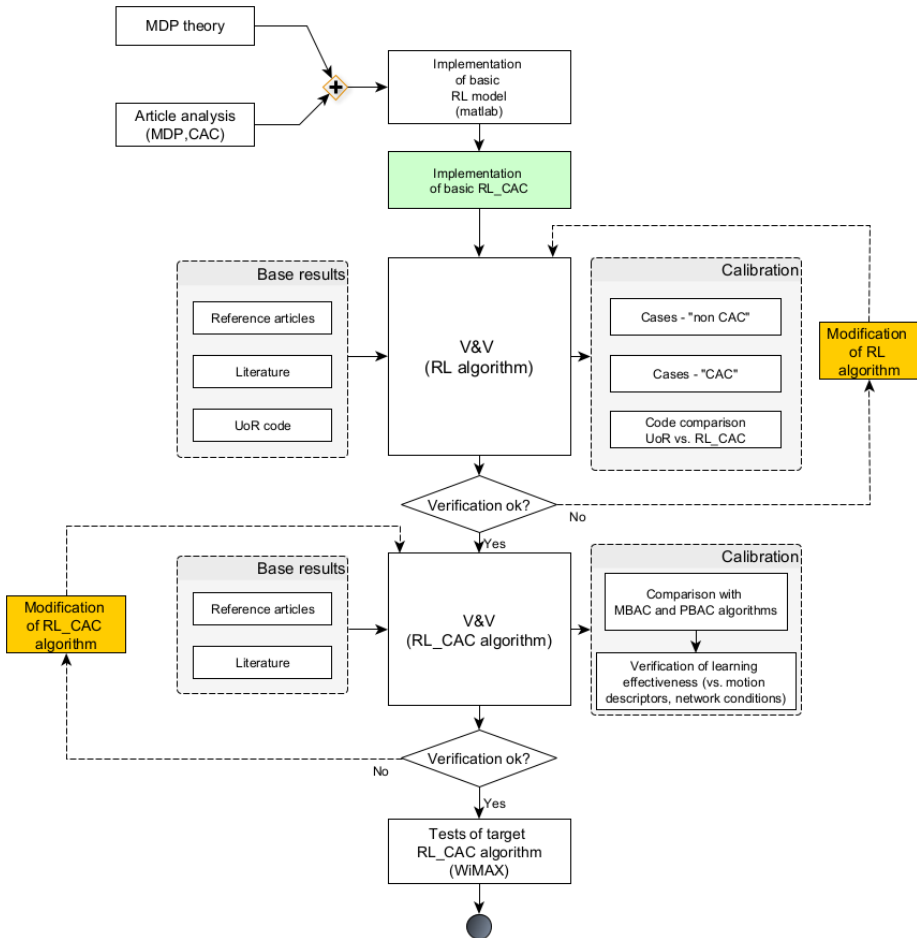


Figure 108 Conceptual diagram of activities needed to deliver relevant RL-based CAC agent

Measurements carried out for the above settings return optimal policies consistent with those presented by the authors in [153]. The results below refer to the “Step3” in the diagram above (see Figure 108). The results are compared to validate the proper calibration of the model. It can be seen that the results achieved with both methods (using Linear Programming and Machine Learning) are the same. As expected both methods (LP, Q-Learning) have successfully rendered the optimal policy to achieve maximum revenue in the long-run.

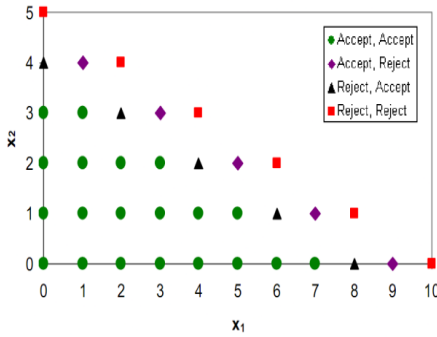


Figure 109 Results from simulation with MDP and Linear Programming

$$(\lambda_1 = 1 \frac{call}{s}, \lambda_2 = 1 \frac{call}{s}), \text{ after [153]}$$

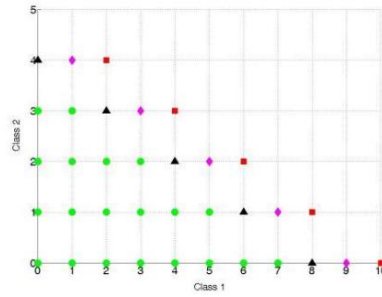


Figure 110 Results from simulation with MDP and Q-learning

$$(\lambda_1 = 1 \frac{call}{s}, \lambda_2 = 1 \frac{call}{s}); \text{ own results of the author}$$

8.5 CALL ADMISSION CONTROL – MDP FORMULATED PROBLEM

A CAC agent problem can be considered as a sequential decision-making model and thus modelled as an MDP [184]. A decision maker, agent, or controller observes the current state of the system and chooses actions according to the state information. The problem as an unconstrained MDP is defined and the MDP is defined as $\langle S, A, T, R, \gamma \rangle$, where S is the state space, A is the action space, T denotes the transition probability between states, R is a reward function and γ is a discount factor. The work in [195] and [192] is followed to formulate the MDP CAC problem. However, the notion of “options” is adopted to differentiate between policy sets utilized during a simulation run. An option consists of three components: a policy π , a termination condition β which is determined according to the channel changing distribution, and an initiation set $I \subseteq S$. Thus, an option $\langle \pi, \beta, I \rangle$ is available in state s and when it is chosen (option choice is given via w statistical distribution) the agent follows policy π . Here the policy corresponds to actions that can be followed (blocking or dropping) when in state s and a modulation coding scheme is changed - which similarly to the ARAC algorithm from 4 assumes such change requires triggering the CAC algorithm. The agent chooses an action according to the policy π , transits to a new state and checks the termination condition β . The agent continues until the termination condition is not meet. The role of the CAC agent is therefore to find a policy π^* for selecting actions with maximum expected reward. For the wireless system with the AMC

case the learning task is to find optimal blocking (π_b) and dropping (π_d) policies.

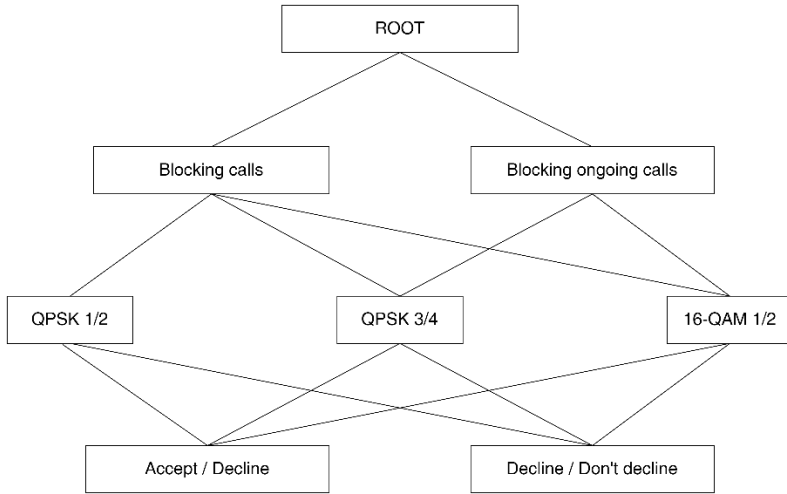


Figure 111 Hierarchical reinforcement learning

An option consists of three components: a policy $\pi: S \times A$, a termination condition β which is determined according to the channel changing distribution, and an initiation set $I \subseteq \Omega$. Thus, an option $\langle I, \pi, \beta \rangle$ is available in state s and when it is chosen (option choice is given via w statistical distribution) the agent follows policy.

8.6 REINFORCEMENT LEARNING

Reinforcement learning algorithms utilise the information from the environment to deliver optimal policies for a MDP formulated problem. A CAC agent can learn the relation between actions chosen at specific states, by following a custom policy, and the rewards he obtained following that policy. In addition, the agent utilises this knowledge to increase his future performance thus yielding an optimal policy for the problem. Q-learning algorithm can be used to improve an agent's performance and maximize his future rewards received from the system. It has been proven that the algorithm converges to the optimal policy under the assumption that each state-action pair is visited infinitely often [195]. For each state action pair a value is calculated and stored in a look-up table. The next section introduces reinforcement learning as a method for solving the MDP CAC problem. Q values are computed after a call arrival, or call drop, following an action. The Q learning approach uses the following error function to update the Q-values:

$$\Delta Q(s, a) = (r_t + \gamma \max_a [Q_t(s, a)] - Q_t(s, a)) \tag{8-14}$$

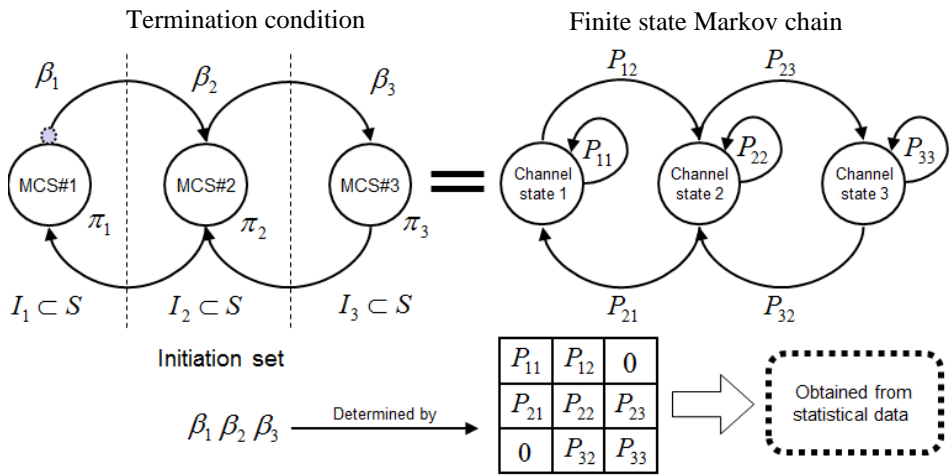


Figure 112 Diagram showing the approach to perform “options” based learning

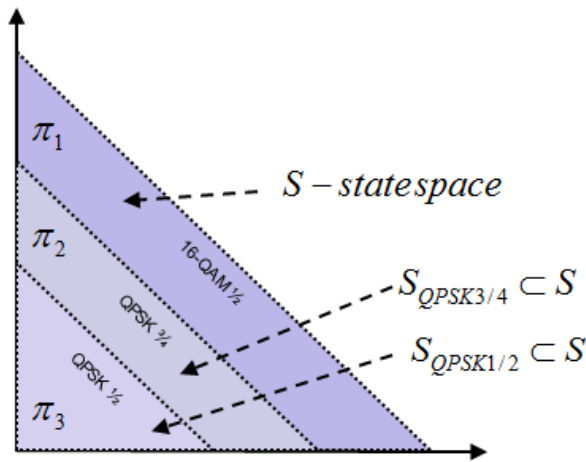


Figure 113 The visualization of the “options” concept for wireless system capacity modelling

In equation (8-14) γ is a discount factor introduced earlier in this chapter and $\alpha = \frac{1}{\text{visit}_t(s_t, a_t) + 1}$ is the learning rate. Here the variable $\text{visit}_t(s_t, a_t)$ represents the number of times the state-action pair has been visited by the CAC agent up to time t . In order to regularly visit state-action pairs, the learning is started with formulating a custom policy and the agent starts by following that policy [195]:

$$\pi = \arg \min_a \text{visit}_t(s_t, a) \quad (8-15)$$

This ensures that states will be visited equally often, and the convergence of the algorithm is accelerated. Furthermore, the problem of finding a CAC policy for 5G/4G/other system that supports an AMC, can be defined as a set of subtasks. In a network with different modulation coding schemes the subtasks correspond to finding blocking and dropping policies for each MCS. Using the options approach local optimal blocking and dropping policies for each modulation and coding scheme can be obtained (it was decided to narrow down the number of MCSs to three cases - QPSK 1/2, QPSK 3/4 and 16-QAM 1/2) which constitute to the overall optimal policy for a wireless system with the AMC. The results of the simulation are presented in Section 8.7. The next section introduces Artificial Neural Networks and their relevance when approximating the Q value function for the CAC problem.

8.6.1 Artificial Neural Networks

To additionally enhance convergence speed, and only insignificantly reduce the fidelity of the Q algorithm, Artificial Neural Networks can be used to approximate the policy function. An Artificial Neural Network with Q-learning algorithm was used in [195]. The authors show in the article that using ANN's to output values of Q can increase the CAC agent performance and outperforms the approach with look up tables in terms of convergence speed. However, no studies were carried out towards selecting the most suitable training algorithm for ANN.

8.6.2 Q-Learning approximation by ANN

One of the problems when using time difference algorithms is the insufficient exploration of states during training. The situation in which, given the agent's parameters, certain states will never be visited is normal and results from the fact that they are, under the given simulation assumptions, unreachable for the agent. In contrast, rarely visited states will not converge in a short time due to poor exploration of these states.

The problem of generalization in the Q-learning algorithm is a well-known problem [298]. Therefore, in order to speed up the process of finding optimal strategies, while not causing a decrease in the accuracy of the Q-learning algorithm, thus author introduces an artificial neural network to the system to approximate the strategy function. In [299], the author proves that such a combination significantly speeds up the process of convergence of the value function and is a more efficient approach than the use of ordinary tables storing Q values. According to the authors in [300], the greatest challenges currently faced by neural networks concern the processing of non-stationary signals, variable in time, accompanying non-linear systems with variable parameters (e.g. biomedical signals). Another of the so far unresolved problems is the increase in the ability to generalize, which, according to the authors, is far from ideal. A

related question is the choice of the structure (architecture) of the neural network. In the conventional approach there are no prescriptions for the comparison of the performance of different network structures. In the probabilistic perspective one can instead apply to this problem the principles of Bayesian model selection (MacKay 2003; Sivia 2006; von Toussaint et al. 2006). Regarding the ANN architecture it is crucial to tune it properly – authors in [301] show the different modifications made towards improving the target accuracy of prediction before achieving the satisfactory results (i.e. improvement from the baseline 46% up to 96%). The optimal solution was attained for certain „learning rate” and „more neurons in hidden layer”. Although too many neurons in hidden layer may lead to overfitting.

As shown in the literature review, reinforcement learning can be unstable or even lead to discrepancies in learning if a non-linear function such as a neural network is used as an approximator to represent the action-value component (also represented as Q) function. Although it is known from the literature that Q-learning converges to the optimal policy by interacting with the environment. As shown in [302], a single properly selected ANN network architecture can successfully learn control policies for different environments with very minimal input knowledge, using the same algorithm, network architecture and hyperparameters for each game, using the same data that would be available to the actual player. The key elements that determine the high effectiveness of learning are the innovations presented by the authors, including replay memory, separate Q-networks and ANN with convolutional architecture.

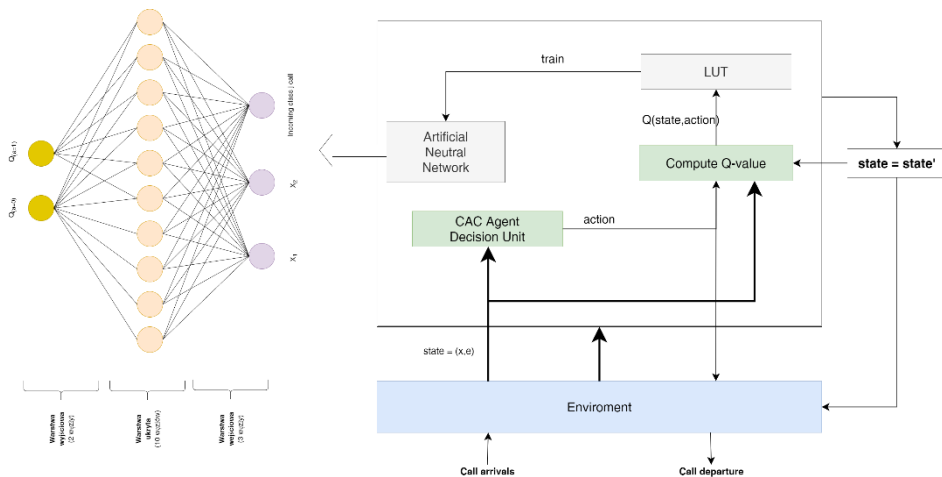


Figure 114 CAC agent with artificial neural network

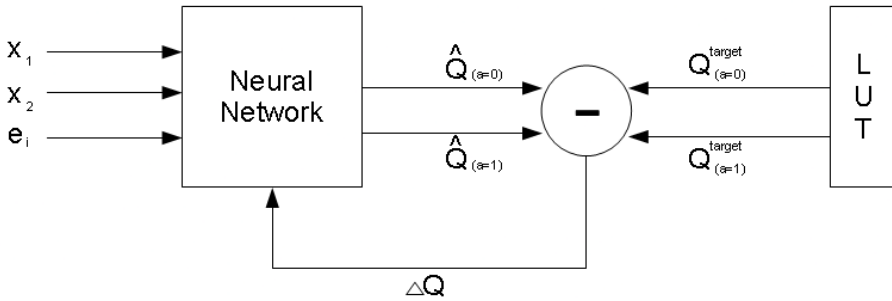


Figure 115 Back-propagation neural network [Source: own]

Hence in scope of this work author employs different ANN training algorithms to evaluate the performance of each training type in terms of prediction accuracy for the CAC problem. A set of training data consisting of 76% of original converged Q values from LUT was constructed to train the neural network. A feed-forward neural network with back-propagation and several training-algorithms were utilized (the algorithms are part of Matlab’s Toolbox and were used for the evaluation process). Each algorithm is used for training for a time duration of 180 seconds and each training is repeated 10 times. First the accuracy of each algorithm is evaluated by estimating the accuracy of the algorithm for each repetition. Second, the Q-values from 10 repetitions are added together to estimate the average Q-value for the training algorithm.

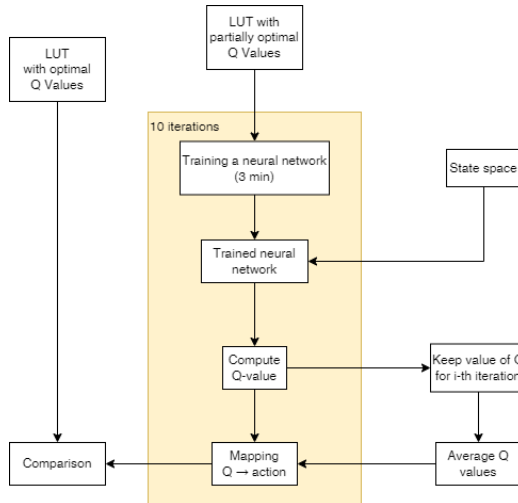


Figure 116 Methodology of testing the accuracy of Q value prediction for a neural network.

The results for prediction accuracy assessment of ANN with different training algorithms are summarized in Section 8.7. As a result of using the ANN the computation time needed for convergence decreases and less memory is needed to store Q-values.

8.7 VALIDATION OF THE MODEL

In this section results for the CAC agent in wireless networks, modelled as Markov Decision Process and solved using Reinforcement Learning with options approach, are presented. As it is considered the two traffic classes - Best Effort and Unsolicited Grand Service. The system receives a reward equal to “1”, when a BE call is accepted and a reward equal to “5”, when a UGS call is accepted. It is assumed that that BE calls are dropped when the system throughput decreases. For each call drop the system incurs a cost of “-5”.

The results of the RL-CAC algorithm are compared to a CAC agent that accepts every incoming call without QoS control (Complete Sharing CAC) and are depicted in Figure 118 and Figure 119. The results for blocking probabilities have been improved in the case of RL-CAC as compared to CS-CAC for the UGS traffic class. Moreover it can be seen in Figure 120 the reward received from the system is higher for RL-CAC due to optimal decisions. The next step of evaluation would be to define some constraints on blocking probabilities for UGS calls and evaluate the model with optimal dropping policies.

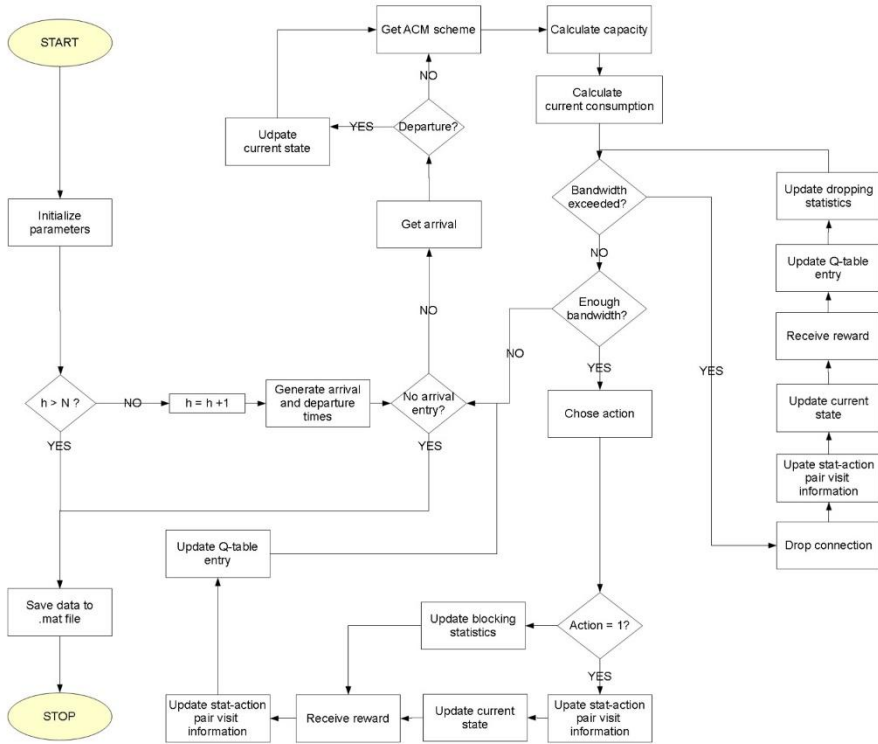


Figure 117 Block diagram for the RL-CAC learning algorithm.

The comparison of ANN Q-value prediction accuracy using different training algorithms is presented in Figure 122. The goal was to estimate Q values for states less frequently visited by CAC agent. The output from the ANN was then compared to the optimal policy obtained through simulation. The results show that the algorithm with the lowest variance of prediction accuracy was TRAINBR. However, when averaging over the Q values predicted by ANN it can be observed in conclusion that all algorithms have a prediction error between 2-10%.

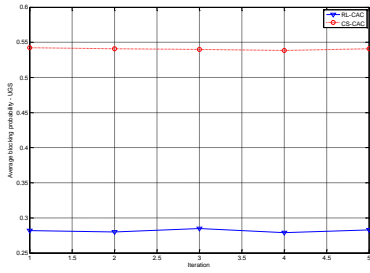


Figure 118 Average blocking probability for UGS class call

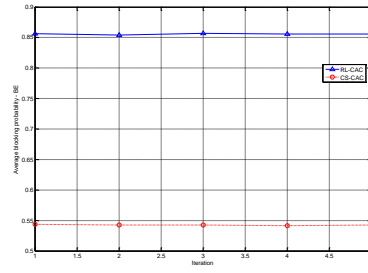


Figure 119 Blocking probability of BE calls

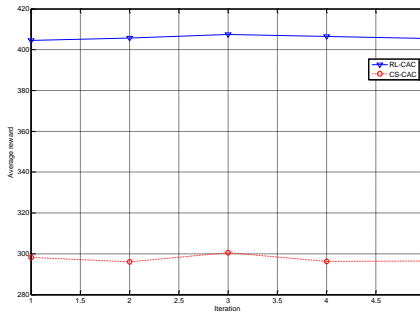


Figure 120 Average reward

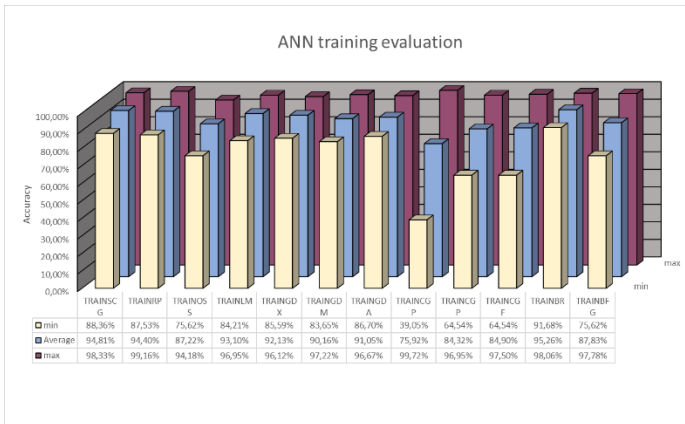


Figure 121 Results of prediction accuracy for various ANN

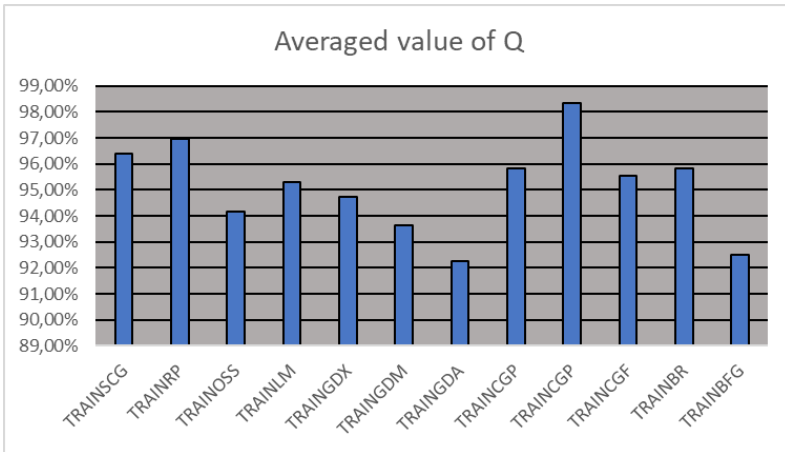


Figure 122 Average Q-value based on different ANN type (after training)

When the training is repeated several times the TRAINCGP, TRAINSCG and TRAINRP algorithms can be used as they achieve the lowest prediction error Figure 121. It has been decided to use the most stable training algorithm with lowest variance and highest average accuracy - TRAINBR (Bayesian regularisation). The algorithm updates the weight and bias values according to Levenberg-Marquardt optimization. It works by minimising a combination of squared errors and weights and then determining the correct combination to produce a network that generalises well.

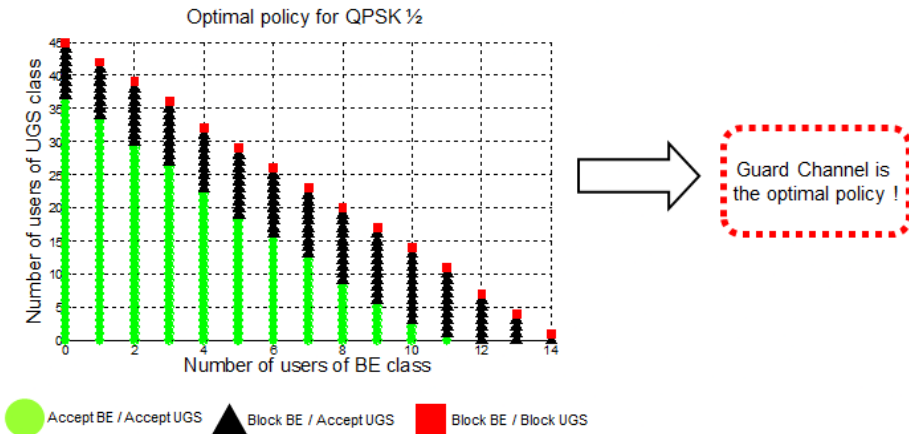
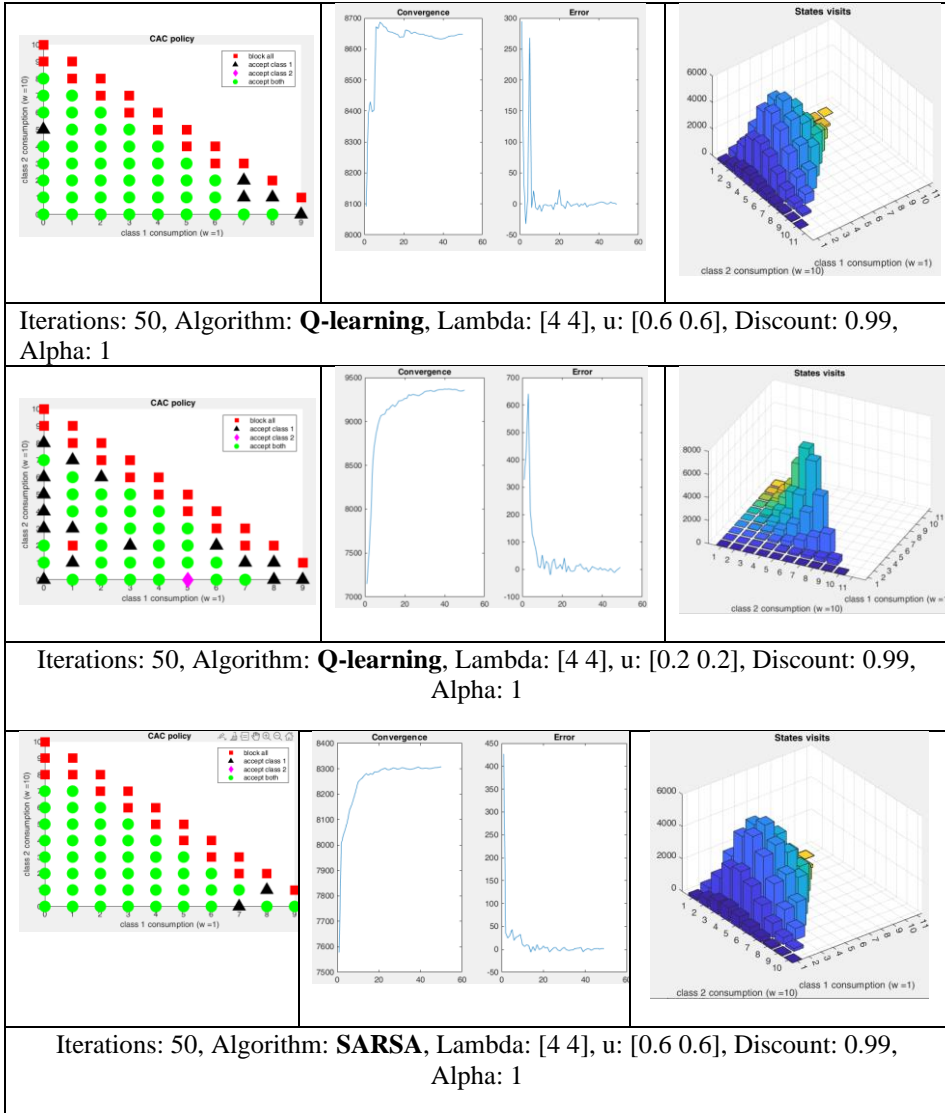


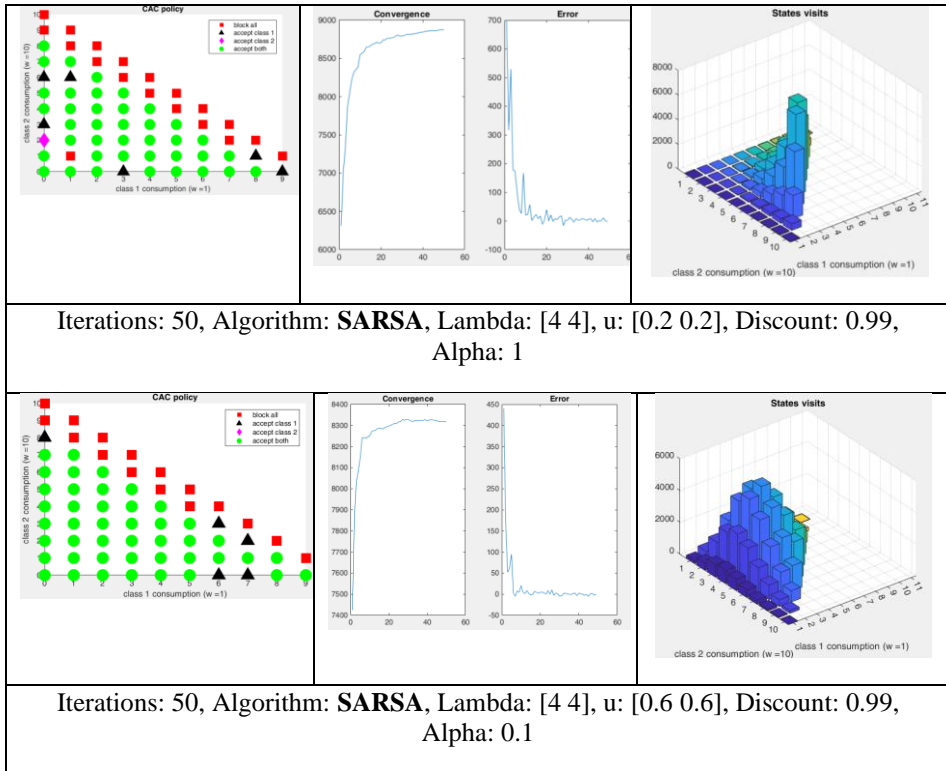
Figure 123 The example of policy after performing learning.

The policy resulting from the MDP-based optimization will depend on the user traffic parameters. The number of policies have been identified, and presented

based on the training campaign where various traffic settings are used (Table 42). The training is shown for the 50 episodes (i.e. decision epochs), various reinforcement learning algorithms are used (Q-learning, SARSA). The intensity of arrivals is constant across samples, but the holding time μ , is varying across plots.

Table 42 Sample results of training for short number of episodes and two learning algorithms (Q-learning, SARSA)





8.8 APPLICABILITY INTO 5G/BEYOND NETWORKS

Given the benefits of the MDP/ANN solution considering its high flexibility, low cost of utilisation ($O(1)$) after training, fast and accurate training – this solution looks relevant to be used as the controller algorithm in the commercial networks. The training could be happening based on data aggregated at the level of the non-RT RIC (or near-RT RIC) and the use of trained controller would be subject of ready model deployment to the near-RT RIC. At the level of near-RT RIC the xApp would be provided to deal with model exploitation when responding to the requests for data bearer or UE admission control. The concept of aligning the model with the admission control concept is presented in the Figure 124. As can be seen policy (as shown in the Figure 123) would be transferred to the CAC xApp for consuming the CAC model as a lookup table, that is used every time the data bearer request is delivered to the near-RT RIC via the E2 interface.

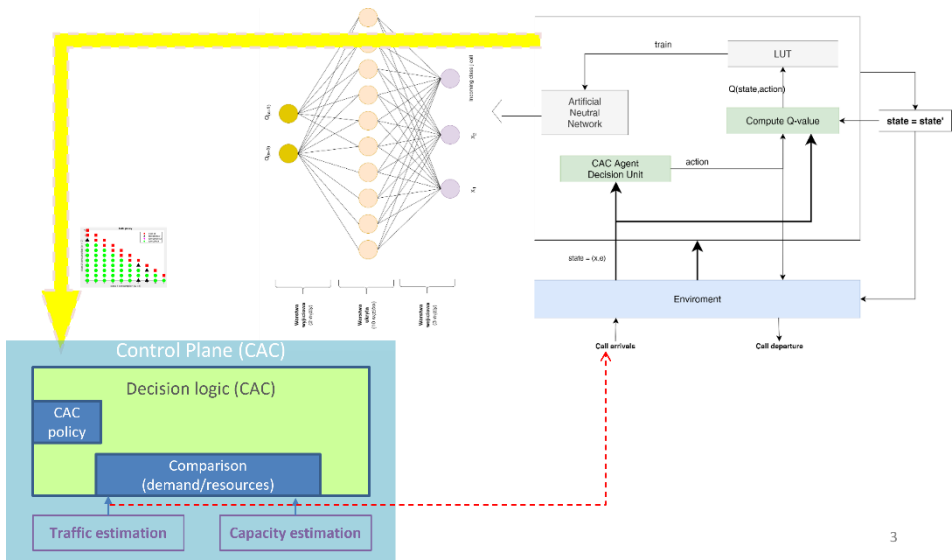


Figure 124 Concept of utilizing the model trained by the CAC agent with an ANN network

8.9 SUMMARY

In this chapter, the problem of CAC agent in wireless systems with Adaptive Modulation and Coding was investigated. Reinforcement learning techniques were utilised in order to deliver the optimal CAC policies. Also, to provide a platform for validating different policies, the options approach was used to choose policies defined over a common state set. This allows the CAC agent to choose optimal policies under variable channel conditions. The results of simulation were compared to a complete sharing CAC in terms of blocking probability and average reward. Furthermore, several back propagation training algorithms for artificial neural networks were simulated and compared results to find an algorithm suitable for the CAC problem. The study has shown that when using a feed-forward neural network with four nodes in the input layer, ten nodes in the hidden layer and two output nodes, the training algorithm with Bayesian regularisation is appropriate for the CAC problem (high accuracy and small variance). However, at the moment of writing this thesis no feasible dropping policies were calculated leaving it as a subject for further study.

9 ANALYSIS OF RESEARCH RESULTS

In this chapter author is concluding the chapters 4-8 with the analysis of results in alignment with the methodological goals as presented in chapter 3. Analysis here directly links to the postulated state (SP_i) targeted by the solutions developed in the frame of this thesis' key research questions the the hypothesis on the performance network admission and congestion control, which is attainable with the modernizations proposed by the author.

9.1 OVERVIEW OF ACHIEVEMENTS

Below in the Table 43 author has collected the achievements throughout the chapters with own contributions (4-8). The table identifies the modelling targets and the related achievements, provided as the result of the research work that has led towards obtaining responses to the research questions and confirm the thesis.

Table 43 Outline of analytical analyses performed in chapters 4-8

Chapter	Target of modeling	Achievement
4, 8	Analysis of admission control agent decisions making quality in 4G/5G/beyond 5G wireless networks in the form of Grade of Service metric assessment (P_B, P_D, BW_{util}).	The tool used is the 4G/5G network simulation (4) and the reinforcement learning as function approximation (8) to provide modernized admission control algorithms, also with symbol reservation schemes, for the future wireless networks.
5	Statistical analysis procedure to match the distribution of delays and packet losses, based on data from experiments (measurements of delays and losses in networks of various operators)	The result of the study is: i) a statistical model of delays and packet losses for a radio channel implemented in the form of a macro in the XLS program ii) an original solution for the emulation of the 4G/5G network in connection with the simulation of signalling overheads for the 4G/5G network
6	User-observed quality (QoE) testing based on a model using objective metrics resulting from network measurement (i.e. throughput, losses, delays, jitter) to support decisions on switching traffic streams between 4G/WiFi networks	Statistical analysis consists in analysing the instantaneous and statistical behaviour of the QoE performance as perceived by the user but based on network objective measurements in case of multi-RAT scenario (6, Fig.16, 17)
7	Time series analysis of the use of computing power (in the form of CPU metrics) consumed by	This chapter provides various prediction techniques to improve the admission capabilities with regards

	<p>software modules of the 5G ORAN network, i.e. CU, DU, RU. The purpose of the analysis is to predict future load values to support CU compute scaling decisions.</p>	<p>to scaling modern 5G/beyond 5G workloads among the edge servers, when local capacity is drained. For model quality metric the following metrics are used: RMSE, accuracy of prediction (for k-steps prediction).</p> <p>Nonlinear approximation methods: LSTM</p> <p>Correlation methods: ARIMA, N Beats</p> <p>Alternative regressors: XGBoost, Bagging, Random Forrest, Decision Tree, Linear Regression</p>
--	--	---

In the remaining sections of the chapter 9, author presents the analysis of results presented in the chapters 4-8. Ultimately the overall results are summarized in chapter 10 together with future work.

9.1.1 Quality analysis based on the GoS metric (Problem2)

This section is directly connected with solving the research problem (Problem2) stated in the Chapter1 of this thesis.

The research in chapter 4 analyses supplementary questions about the impact of the dynamics of the system (such as resource optimization techniques e.g. AMC) has on admission control and quality of service. It is by assuming that future networks will be more dense and chances for either handover or change of MCS will increase. To evaluate the performance of envisaged algorithms and assess their impact on the system, author has developed a cell-level simulation environment that relies on the proposed methodology. System under test (SUT) controls resources using either novel admission control (AC) mechanism ARAC or its predecessor EMAC, introduced in [117]. The algorithms are both traffic – aware and designed for controlling the VBR traffic with burst arrivals. One of them relies on calculating simple exponential weighted moving average (EWMA) of the overall resource consumption, whereas the other in the process of resource estimation differentiates between the new and the ongoing connections, thus providing more accurate resource estimations. Simulation results show that both of presented algorithms can provide appropriate QoS levels in the tested configuration. However, ARAC provides protection against connections arriving in large batches. Therefore, average delays of ARAC are generally lower than that of EMAC and reach the difference of approximately 23 – 25ms at maximum (depending on the VoIP codec used). These differences could prove crucial in a system with non – negligible core network delays. Results of CAC comparison prove that proposed ARAC algorithm decreases the delay

experienced by VoIP connections the higher the arrival rate at the cost of increased blocking probability.

To understand full potential of the introduced own modernizations, it was identified that the legacy EMAC offers a lower probability of rejecting new calls (P_B), it cannot be considered as an algorithm with higher performance than author's own modernization called ARAC. The fact that EMAC is not able to provide a QoS guarantee to growing number of arrivals, even despite of accepting these connections, makes it not satisfying solution if QoS assurances are required. When applying a CAC algorithm, an uncontrolled decrease in the QoS level caused by system overload should be minimized. On the other hand, the proposed ARAC algorithm ensured the appropriate level of QoS in all simulated cases (Figure 51) independently of the averaging interval length in range 10-500 super frames (K). The ARAC offers a probability of new call blocking comparable to another algorithms known from the literature nscARAC, and as has been shown to improve it by ca. 10% in certain scenarios. These are scenarios where the ratio of $t_{conn}/K < 1$ which in practice means that duration of connections is relatively short compared to averaging window. This is the situation where we expect growth of density of APs, thus also shortening its effective radius (10-100meters for nano/femto cells respectively) resulting duration of a connection would be shortened reasonably. And the decision of choosing the length of the measurement interval is characterized by much greater flexibility due to more stringent control of resources (connections recently admitted/started or finished). For ARAC, both the probability of rejecting a new call and the probability of terminating an existing call prematurely remain almost constant, regardless of the length of the measurement interval.

Moreover, the symbol reservation schemes proposed in chapter 4 have shown that the WCSRS scheme can be considered only in systems with coverage gaps, i.e. with poor SNR conditions. Symbol Reservation Schemes directly impact bandwidth utilization and blocking probability, therefore more attention should be brought to the topic of SRS combined with adaptive persistent scheduling and admission control. Author so far has not identified many past research that would be combining the symbol/PRB reservation in order to control the admission control probabilities (P_B, P_D). Combining the reservation scheme with particular admission control algorithm provides additional means for enforcing QoS for a connection in a persistent way – especially in the case of mobility, where the SINR changes at the cell edge.

Previous measurements in the field is enhanced by improving the fidelity level of the proposed 4G/5G simulator. To compare SUT's performance using either nbLDPC or legacy CTC (Convolutional Turbo Coding) codes in a mobile channel, a method called Link-To-System interface (L2S) has been implemented. A method based on mutual information (MI) called RBIR (Mutual Information Per Received Bit | Received coded Bit Information Rate) was selected. The simulation environment builds on top of the NS2 and MATLAB software

packages. For admission control simulations with nbLDPC and CTC codes the conclusion can be drawn that achievable gain of nbLDPC can only be observed if users experience relatively good channel conditions. For higher modulations less MCS transitions for nbLDPC codes are observed, which results in lowering dropping probability and slightly increasing average system throughput. Nevertheless, if users' experience is moderate due to bad channel conditions, gain achieved thanks to nbLDPC codes becomes insignificant.

According to another group of simulations, the impact of using different error correction codes is most significant in case of varying SNR conditions. Using more robust FEC schemes (e.g. 5G nbLDPC) results in higher average cell throughputs while keeping similar dropping probability levels - when coping with bad SNR conditions. When coping with good SNR conditions – it results in lower dropping probabilities while keeping similar average cell throughputs.

It is worth noting that the future networks 5G/6G will need to be more dense to enable the expected traffic growth (x20 by the 2030/35) but also the cooperative RRM approaches will become more widespread. It is simply due to the fact that the spectral efficiency of spatial diversity techniques like mMIMO or beamforming has limits of ca. 8-10bit/s/Hz and will not be enough to cover the demand. This way in order to meet traffic demands in the hot-spots of the densely populated cities where there can be 8k-18k inhabitants per square kilometre, amount of bandwidth that would need to be provided (re-farmed) is at the level of 1200Mhz or more [13]. And such spectrum share, would first need to be commissioned by the regulators. An alternative is to provide the hybrid hetnet layer of dense small-cells (also pico-femto) which would replace the need for such high share of the spectrum band. Here a practical example is an opportunity to replace additional 1250Mhz for a densely populated city at a kilometre square area covered with 7.2 macro cells with 177 small cells (13x13 cells grid) to assure same area capacity. Such density to be practical requires cooperative RRM approaches, where the radio units (RU) are treated as a homogeneous pool of resources that is managed by e.g. cell-free scheduler. Such scheduler is specific as it allocates resources based on a 3D approach (UE-RU-PRB) across the whole network (or network cluster). This way as shown by authors in [11], [10] the average SINR can be increased especially at the so called “cell-edge” locations, i.e. in places where an UE is far from any nearby cell centre. In this approach handovers are no longer necessary, and it is the scheduler that is capable of allocating (and reallocating) users to APs with high granularity e.g. TTI-level dynamics.

It is important to notice that the CAC parameters considered in the chapter, could be considered as additional actions to design other learning algorithms in the future. The following actions can be considered to assure better control over resources (e.g. as parameters of a MDP state): a) adjustment of the length of the measurement period K , b) activate selected variant of the “symbol reservation” (CCSS, RFSRS, WCSRS), c) modify threshold of guard channel for mDHCAC. Moreover in case particular deployment of 5G/6G would need to deal with

computing resources limitations, algorithms available could be tuned based on their required computation resources.

9.1.2 Quality of congestion control based on packet loss and delay models (Task2)

The target of the work for chapter 5 was to define a complete, balanced, end-to-end evaluation framework for innovative and modernized congestion control algorithms (in 4G/5G/beyond) to enable appropriate optimizations adapting transmitted stream of data to the underlying network capacity (e.g. remote monitoring of autonomous cars) for improving capabilities of reliable mobile testing in laboratories, based on field test data. Special target of this work is to support the teleoperation of remote vehicles. Here the very recent EU project that is currently in the early phase of showcasing such use-cases with real trucks and other vehicles in the city of Antwerp is the project 5G-Blueprint [234]. The project is using the commercial 4G/5G access networks to teleoperate above mentioned vehicles, to deliver good understanding of the available capabilities in that domain.

After performing comprehensive suite of drive tests in various locations across Poland, author has carried out statistical tests regarding the modelling of the HARQ mechanism and the impact of throughput on delays and losses. The main aim was to be able to inject the results of statistical analysis (as a tuned model for packet delays and losses) into the emulator developed by the author.

Model validation is carried out separately for each scenario (loss simulation, delay simulation), its implementation is aimed at comparing the distribution of field measurement and simulation results. The resulting simulation tool can be utilized in order to inject delay and packet loss into the TBONEX emulator. After performing the new set of field tests it is possible to apply the methodology presented in the current section and tune parameter values, so that they can be then used inside the developed simulator (MS Excel file with VB Macros). The simulator tuned based on the prepared set of parameters has been utilized in the emulator to enable simulations based on these particular channel models.

The overall aim of the applicability of the non-reference QoE metrics here, was to use it to indicate perceived level of video quality, and based on its measured levels adjust the settings of the video source (e.g. transcoder, camera) and/or the settings of the radio resources of the wireless link. In case of highly variable channel with low SINR (or high variance) the need to adjust QoS such links will be increasing. Through the use of the proposed emulator it is possible to evaluate such algorithms for different scenarios and settings. And thus the adaptability of the video traffic can also be better profiled by collecting appropriate datasets from such (many) experiments.

The first set of tests (Series 1) in the chapter 5 presents the measurements

collected using IP camera. It was noticed a negative impact of the use of transcoder on the quality of QoE results. To confirm this assumption we started series of tests where instead of Streamer we used live feed from IP camera. These measurements were carried out using the emulator and aimed at determining whether the observed disturbances in video quality is the result of the use of a streamer or the emulator:

- collected results allows concluding that the video from the camera behaves much more stable than its counterpart video file streamed using Streamer also the case when emulation was activated
- tests aimed at comparing the use of various transport protocols: TCP and UDP show that that despite the expected effects (UDP experiences more lost blocks while TCP freezes more often) the difference in the overall quality is not significant
- examining the influence of utilizing delay distribution (Gaussian) instead of instantaneous values of delay from traces into the emulator. The original assumption at design stage of the TBONEX was to use such approach (delay distributions are derived from number of sequential samples in the trace file) to reduce the number of traffic control commands inside the router enforcing QoS parameters at the IP level. Comparing the results, however, it turned out that this is true only for artificial traffic (non-video) because introducing delay distribution, even small in the mean value, caused significant problems with video quality (e.g. introducing freezing).

Second group of tests (Series 2) – provided in the chapter 5 presents measurements collected using the Streamer and a locally stored video file. To spot the issue that author has identified i.e. “*where the video disruption is generated in the E2E path*” at first stage considering the Server, wireless link, and the Client. The suite of tests was planned with several different settings: different components enabled/turned off, various bit rates etc. It was possible to confirm that the *transcoder introduces noticeable degradation* of video quality due to its high consumption of resources. In addition, the same tests were performed on machines with different processing power (i5, i7 CPUs). It turned out that the transcoder is quite resource consuming, and the quality of transcoding quite heavily depends on the computing power used.

Series 3 shows the measurements made with the streamer, that streams locally stored video file, but this time transmitted through the emulated channel, and with an emphasis on the use of the TCP protocol. The TCP was used to evaluate how well its built-in AIMD traffic flow control can synergize with the objectives of security scenario (where video from a camera is used to help accomplish a task to the operator). The purpose of this test was to see whether a change of the transport protocol from UDP (or reliable UDP) to TCP will positively affect the previously observed problems (degradations in quality under increased jitter). In addition the focus was also on checking the performance of the end-to-end solution configured close to its target shape and with more advanced test

scenarios. The collected results suggest that the degradation of the quality when comparing simple emulation script (stable radio conditions, LOS) with more extreme scenario (low mobility with partial NLOS) is not very large, but certainly it is noticeable. In another attempt to similar parallel measurements were performed with the same parameters using the Greenpacket router that was sending the same video as the real 4G network. This test confirmed that the results obtained using real networks present high level of similarity to those using emulation, **suggesting that emulator design was successful**. The summary of the above measurements (Series1-3) is presented below:

- Introduction of the Streamer increases the chance for packet re-ordering
- The increase of packet reordering at the receiver side, automatically coincides with greater chance of freezing
- Simple scenarios (i.e. where radio conditions were almost perfect) resulted in worst QoE - the reason was the misbehaving operation of the “delay smoothing” technique in TBONEX. After identifying the issues, it was removed from the future tests.
- Increase in video rate showed that QoE metrics were improved largely (Blockloss, Blockiness)
- The level of one-way-delay for tests with TCP was substantially higher than for tests with UDP.
- Although there were attempts made to modify the settings of the VLC reordering buffers there was practically no influence perceived during our tests.
- Trying to evaluate the difference in QoE depending on the video player used (VLC or the LiveView player) no significant difference between the resulting QoE were identified. The FFPlay was used only with UDP as author was not able to successfully configure it with the TCP protocol.

Although the proposed emulator model (TBONEX) does not strictly support algorithms related to dual connectivity which can be important enabler of the future networks (e.g. using the LWA type of solutions, or protocols similar to MP-TCP), the tool certainly can be utilized to verify such algorithms effectiveness in varied network conditions e.g. varied number of users, varied channel signal strength or varied user behaviour with different priorities attached to them. This can be achieved by either simulating such algorithms by implementing dedicated scripts or with the use of newer hardware mentioned in section above.

9.1.3 Service quality (QoE, QoS) analysis with multi-RAT decision agent (Task2)

The current section relates to the research problem of quantitative assessment of the quasi-optimization of admission and congestion control parameters and algorithms benefit from combining it with intelligent, QoE-based traffic control in multi-RAT networks using novel RAN controller architectures in the control

plane. Based on the experiment defined, designed and performed and the novel two applications for the RAN controller, author has collected the following feedback from the measurements.

A total of four configurations or use cases have been tested, two with single RAT (LTE radio access), but with different scheduling criteria (RR and PF), and two with the two selected RATs (LTE and Wi-Fi), but with different criteria for switching decision (SINR or video QoE). The findings of this research demonstrates that the multi-RAT activation through RANC provides much better QoE for the user compared to a single RAT system. The above-mentioned results indicate that the marriage of SDN-based networks with existing infrastructure of Internet service provider (ISP) is a valuable contribution to traffic steering. An ability to dynamically and adaptively switch traffic between different RATs based on appropriate policies enables new capabilities for handling priority or best effort traffic in collocated networks. The success of such approach will depend on the operators' and infrastructure providers' interest in collaborating with e.g., SD RAN vendors (like 4G, 5G RAN) where the existing capacity of "the pipe" will gain more utilisation by availability of OFDMA based "scheduled radio unit" that augments Wi-Fi access points at an indoor location. As a follow up of this research author foresees the interest in continuing the multi RAT traffic control with the WiFi6 (OFDMA scheduler), 60 GHz links (WiGig), and more than 100 GHz links (Terahertz) for indoor locations.

9.1.4 Quality of predicting CPU consumption to leverage the admission control actions (Task3)

The chapter 7 is touching the emerging topic of the role of an edge and AI in the future development of networks to support various applications, verticals and use-cases. Here the tight and close cooperation of the workload prediction/placement algorithms with a) admission control and congestion control on one side and b) the applications supported with AI/ML middleware is crucial to assure real-time guarantees but also help in the emergence of new applications and use-cases (eHealth, smart city surveillance, optimized waste control in PCB designs etc.).

Several machine learning models including traditional regression models and neural network model (LSTM) for time series forecasting were implemented. LSTM model is known to be good at time series forecasting but traditional ML regression models are also considered here because of the nature of the problem to be addressed. It is believed that workload of edge servers in wireless networks usually follows a certain pattern. Therefore, such prediction problem, by extracting features from the time, can be solved by regression models. These models are trained with one-year, hourly-granularity data and are then evaluated with root mean squared error metrics. The evaluation demonstrated that the 2-layer LSTM model outperformed other regression models, with RMSE loss of 653. In addition, K-step forecast was performed. Assuming the LSTM model has

1-week input (i.e. 168-long sequence), one can predict the CPU/energy consumption of the following 2432 hours, with subsequent true values observed. This is open-loop forecasting because in reality, a system can predict the value of the next timestamp given an input sequence. Then it observes the true value of the next timestamp, adds to the input and keeps predicting for the future timestamps. Other ML models, on the other hand, perform prediction by extracting features of a given timestamp. Such predictions are compared with true values. The LSTM model proves the most accurate predictions given the true values. In the end, a conceptual guideline to implement such models in ORAN is provided.

9.1.5 Quality of the intelligent CAC agent for wireless networks (Task4)

The research work on synthesizing controller for admission of new requests in future wireless networks is proving that the appropriate definition of the problem of optimization of both blocking and dropping probabilities in the hierarchical system approach can lead to improved probability of blocking the UGS connections and at the cost of increased blocking probability of the best-effort connections. The rewards for accepting connections when using the RL-CAC approach will be significantly higher with the proposed approach than with the legacy CS-CAC approach. Moreover, author has implemented own approach to boosting the learning process with the use of ANN networks to learn the trajectory of rewards – by monitoring changes of the Q values in the Q-Learning approach. The learning approach attempts modelling the optimization problem as hierarchical learning with constraints on both blocking and dropping probabilities. It has been shown that on one hand the learned policies can match the guard-band CAC schemes but also as proven in literature the further extensions of the reward function, definition of state can bring more benefits.

9.2 REMARKS

As it can be seen in the above sections the defined research problems (Problem1-2) as well as the supportive tasks (Task1-5) were successfully addressed by the author. It can be concluded in general that the planned achievements were reached. Author has delivered multiple models and algorithms that enables fulfilling the research problems and tasks and as such contribute to the identified challenges for the admission and congestion in the future networks (Table 43). The ARAC algorithm provides a customizable solution that can serve the needs of operators especially in the current fast evolution of the networks (5G/6G). Moreover when combined with the symbol reservation schemes it can support resource persistent scheduling as well as serve the congestion control purposes. With the ARAC accounting for dynamics in the resource consumption characterizing the future networks can be managed.

To further customize the admission control the algorithms based on learning, and especially utilizing various forms of MDP agents are very suitable. Especially interesting is the flexibility of defining algorithm objectives and actions that it can consider. The scaling of resources of the 5G/6G ORAN based disaggregated networks, is at the moment the very important target that is tracked by the author.

Moreover the multi-RAT solutions as indicated in the chapter 6, provide important support, especially when combined with QoE or SINR feedback loop towards the RAN controller, that can analyse the status of such quality feedback and based on this can adjust the selection of underlying RAT technology. Such approach can be suitable to any RAT technologies which can be subject to intelligent switching (e.g. dual connectivity, carrier-aggregation) that may provide more options for traffic offloading to operators.

The complete framework for low-cost laboratory configuration (chapter 5), that enables evaluation of congestion control (and admission control) algorithms has been proposed, utilizing the statistical model developed by the author. It has been shown in the comprehensive set of tests that such framework can provide a useful tool especially when designing control algorithms for video delivery in uplink. It is of special attractiveness nowadays when more and more proof-of-concept solution for teleoperation of vehicles is showing up on the horizon. Even if not for the actual driving at high speeds due to network coverage issues, definitely valid for remote assistance when a vehicle has failed to drive autonomously.

Eventually the very promising solution for workload prediction has been developed, evaluated and implemented in a form of a python scripts (based on multiple predictor models). This solution is of particular interest in order to provide efficient support to the future network scaling which will grow in importance due to expected increase in the role of edge technologies (e.g. EMDC) as presented in previous chapters.

10 CONCLUSIONS

On the basis of the analysis, own research, formulated problems and in order to prove the thesis, conclusions of a cognitive, practical and further proceedings can be formulated.

10.1 SCIENTIFIC CONCLUSIONS

By successfully resolving the research problems identified in the chapter 1 (**Problem1, Problem2**) author has achieved the main goals of the thesis. In turn it has enabled addressing the main hypothesis of the work presented in this dissertation. In order to refer to the results of work discussed above (chapter 9) below the few distinct dimensions are discussed:

- Author has proposed novel contributions that address the gaps identified for research in the domain of “admission control for future wireless networks” in the section 2.12. Summary of models developed and achievements accomplished across the chapters with own contributions is presented in Table 43.
- The main methods used in this dissertation to deliver results of the model validation was by using: simulators (5G Vienna, NS2), emulators (own model developed based on field data, in chapter 5) but also by using real-life 5G ORAN network testbed (where author has followed the self-identified methodology of scenario-based data collection to deliver data for predictive model building). The simulation experiments has followed the well known methodologies of system and link level simulations (SLS, LL). Moreover author has cross-validated the wireless coverage models with own measurements performed using self-designed tools (e.g. RaspberryPi based QoS probe). The author intention was to compare realistic measurements with simulated environment. It can be seen that the differences between modelled coverage and the measured one are essential (see e.g. Figure 59).
- The scope of performed experiments has been selected in such a way that the coverage of addressed experiments is most comprehensive. The Table 12 shows that experiments were carried in various directions (UL/DL), with different feedback loops (QoE, QoS, CPU), different radio technologies (4G, 5G), multiple traffic sources (VoIP, video, priority/non-priority traffic). Admission control was studied as a separate phenomena (chapters 4, 8) but also in combination with congestion control (chapters 0-7). This way the outcomes present the relations between elements of the analyses system (of future wireless networks) form a multiple directions. Especially considering the combination of the resource control in radio (symbols, PRBs) at the level of GoS metrics where blocking and dropping probabilities have been thoroughly considered for comparing the results of multiple configurations of both environment and algorithms. Moreover the computing resources consumption were studied for the future deployment of 5G/B5G networks at edge in order to propose prediction based solution for scaling disaggregated functions of the radio stack (CU-UP) between different physical data centres (EMDC). Author has been identified most performing prediction architectures of LSTM that exceed accuracy of prediction of alternative models (n-Beats, ARIMA, legacy regressors). It has been shown that the achieved RMSE is reasonably lower as compared to these approaches (Table 37).
- The results on wireless link modelling, presented in chapter 5 of the use of QoE performance metrics to close the loop for congestion control, shows that using such well defined and validated models with the

properly designed emulation framework can substantially support the process of designing novel controllers for the uplink video delivery, especially in the case of mobile terminals (e.g. cars, drones). The statistical models defined in Figure 69, Figure 70, has been successfully validated for statistical correctness and introduced into the XLS based tool (simulator) as well as the own-developed emulator facility. The tool can be tuned for any new data collected from the field tests (e.g. automated with drone mission) to provide extended customization capability.

- In case of own-research the solutions proposed for admission control, author has referenced the results existing already in the literature as baseline, in order to assess the current status and based on such knowledge divide alternative solutions. The results presented in chapter 4 for the ARAC modifications highlighted in Figure 26 and Table 18 (algorithm compensates for batch arrivals, modulation changes, connection finalizations) show that the algorithm can work well in case of the future networks densification where intensity of handovers and modulations changes is expected to grow substantially. The latency improvement for VoIP traffic attributed to the priority traffic class (but with legacy scheduler) can be seen in figures Figure 39 - Figure 43, despite the fact of compensating for dynamics of MCS changes, sessions finished and batch arrivals. Also when comparing the proposed ARAC modernization with the precursor from literature nscARAC it can be seen that although more compensation has been included by author (i.e. algorithm is more robust to changes), the results of the two algorithms are comparable and even ARAC provides 5-8% lower P_B for arrival intensity 50-250 conn/min. ARAC can account for the additional resources released by recently terminated calls which is visible for highest level of connection arrivals (departures) - Figure 47. In addition ARAC when compared to the EMAC guarantees QoS for all connections admitted and there are no connections with “bad QoS”, i.e. which network failed to assure quality to. In addition to the ARAC author has also contributed to the modernization of the DHCAC algorithm by providing the updated definition of the decision rule (in equations: (4-4) and (4-5)). The expected results of such modification where studied theoretically and depicted in the figure (Figure 21), showing more relevant estimation of „non-GBR” resources in case of more bandwidth
- The generic algorithm for GBR traffic admission was introduced in Figure 22 in order to serve as meta algorithm that deals with degradation of “non-GBR” connections that can now be better addressed once the congestion control modelling framework was defined in chapter 5. These results are crucial to support practical prototypes for teleoperation of cars/vehicles e.g. [234].

- Moreover with the design of the RL-based admission control that considers the blocking and dropping probability limitations in a hierarchical approach, it has been shown that such approach leads to improved GoS (lower P_B and P_D). This is an attractive result as further improvements can be introduced when augmenting state definition with the new actions e.g. “scale CU-UP” once the framework introduced in the chapter 7 in the future research. Especially interesting from the point of view of addressing computing resources overload is the k-step prediction enabled by LSTM solution which in practice can lead to additional time to scale resources without penalty of increased P_B and P_D when the seasonal of temporal computing demand fluctuations. Moreover additional user traffic can be added to the network that without such solutions could have been overloaded earlier.

By targeting the **Problem1**, the second major goal was accomplished owing to assessing the status of existing prior-art CAC algorithms, identified directions for its development, proposed modernized solutions (ARAC algorithm, QoE framework, LSTM workload prediction suited to 5G/B5G ORAN virtualized networks, RL-CAC agent for learning CAC policies online with both Q-Learning and SARSA boosted with ANN for shorter learning times, as well as the intelligent and QOE based switching of video traffic between cellular-licensed and unlicensed (WiFi based) access networks (chapter 6). The proposed solutions provide a portfolio of solutions that is not only useful in the transition form the current 4G to 5G networks but can be robust enough to assure continued evolution towards 6G networks in near future. Solving the **Problem2** and thus accomplishing the goal behind it, was possible by proposing the modernized CAC algorithms (mDHCAC, ARAC) which are not only improving quality in the existing 4G/5G networks, but are suitable to deal with their evolution in the future – especially considering networks trends of i) densification, ii) centralized control and cooperative RRM, iii) increased use and role of AI/ML solutions especially in the control plane.

The accomplishment of the **Task1** addressed in the chapter 1, has been achieved by proving that it is possible, valuable and future-proof to deliver models (statistical model for losses and packets delays in wireless mobile channel) that can be deployed in a simple lab-based environment, in order to deal with high-fidelity wireless link modelling and design QoE/QoS based video controllers (and supported with local radio feedback). This result is of special importance to the use-cases with tele-operated vehicles. By completing the **Task2** – author has defined the role of intelligent traffic offloading presented in chapter 6 can be very useful tool for operators, especially in the case where currently on the competitive market, there are initiatives where subscribers are incentivized to become also operators of local networks (WiFi, 5G). Proposed algorithm (Figure 82) can support the regular networks for daily operation but can as well be suitable solution for improved delivery of capacity for crisis management, where ad-hoc capacity delivery (e.g. by means of carrier aggregation, dual-connectivity

or multi-RAT) can be promising solution, leading to improved control of the QoE when supported by intelligent RAT switching by QoE/SINR based algorithms at RAN controller (e.g. RIC). The proposed solution is actually suitable also for the dual-connectivity as well as the carrier aggregation approaches, as they share commonalities with the multi-RAT context. Where the only differences is the type of spectrum as well as interfaces and means to activate additional resources. As regards the unification of the approach the deployment of QoE-RANC and SINR-RANC algorithms as xApps can largely support this goal. The research towards efficient prediction algorithm aimed at addressing the **Task3** goals for the scaling of disaggregated RAN networks now and in the future, has proven that properly tuned LSTM mechanism applied to workload prediction is capable to provide additional (lead) time for performing OAM activities, at operator side, in order to prepare provisioning more instances of selected protocols (CU-UP in this thesis) on demand and well aligned with the existing orchestration frameworks (e.g. Kubernetes). Combining the proposed time-series prediction with workload placement (i.e. as introduced in chapter 8) would additionally contribute to accomplishing the **Task4**, by simply extending the state definition by adding actions related to CU-UP scaling. This way further customization and synergy of learning at various control plane modules (workload prediction, workload placement) but aligned under common RL-CAC algorithm, can further bring benefits to customers and operators. Such approach naturally coexists with policy (and intent) based network management as available via already existing interfaces of e.g. A1 (ORAN). Reaching objectives indicated by the **Task5** by providing the methodology enabling the above and future modernizations of control plane RRM algorithms (irrespective if they are related to optimization, modernization or innovation) was done by introducing the ML learning-based workload prediction and placement framework. This framework is suited well with the current and future architectures of edge computing platforms. Such splatforms are characterized by common denominators of: the data-driven nature, learning-based operation, cross-layered approach (to deconflict optimization directions), etc. The Figure 95 shows the proposed evolution in the way such mechanisms should be designed and cross-dependencies between them should be considered. Nevertheless, the utmost importance for attaining is the high energy efficiency and energy consumption minimization - which should be always be priority for the future proof network designs.

10.2 PRACTICAL CONSIDERATIONS

The solutions provided by author bring a **consolidated design and provisioning framework** that can be interesting suite of models and tools for future small-scale operators (e.g. neutral host) where the learning and deployment of models, should be efficiently deployable on the edge servers with highly varying traffic loads (AI/ML based application workloads) where the 5G/B5G workloads to some extent compete for computing resources. Dynamic scalability of resources (CU-

UP to start with) gives operators additional leverage for more sustainable network deployments. Low-cost and validated in the field, laboratory solutions supporting efficient design of control algorithms for uplink video delivery will be interesting enhancement of the future use-cases where e.g. drones can be instructed with QoS/QoE probes in order to perform regular scanning of coverage area in order to help tuning the algorithms for dynamic (and potentially also learning-based) video controllers, that aim at improved quality of remote control and e.g. crisis situation assessment.

10.2.1 Recommendations for video controller synthesis

The practical dimension related to the research performed in the chapter 5 can be summarized by the collection of identified recommendations for the controller designers:

- Transcoder along with tools necessary for feeding video to it in case of streaming local file (transcoder must be provided with a video stream from external source e.g. video player or camera) is a resource draining task. The tests executed by the author, showed that, computers with significantly more processing power are introducing visibly less QoE degradation
- The use of IP camera instead of the complete “Streaming Server” component in tests showed, that transcoder combined with the FFMpeg is most resource hungry part of currently tested architecture. Test performed with an IP camera, shows that with similar settings of video stream at transmitter, degradation of video QoE was significantly lower than with use of transcoder, especially considering the Freezing metric. In this case, different stream profiles specified in camera settings was used to mimic changes done by transcoder. Tests also proved that (non-transcoded) stream from the IP camera is slightly more resilient against emulation, but such situation is probably an effect of camera lower requirement for resources in comparison to transcoder.
- Usage of UDP or TCP protocols may vary based on user specific applications and needs. As the approach followed in the thesis is mostly based upon video with TCP protocols (e.g. RTMP), most of tests were performed with usage of TCP. However, the series of tests were performed for purpose of both protocols comparison. As expected from a nature of mentioned protocols, the use of QoE probe indicated that UDP introduces significantly higher Blockloss value while TCP introduces significantly higher Freezing value (it is assumed that it is caused by retransmissions). Although overall quality seemed better with use of TCP, the tests showed that UDP will serve better in scenarios where video fluency is prioritized, while TCP will shine in case where smooth playback is not as important.

- The use of MIMO antennas greatly improves QoE when in NLOS or with the mobility enabled. The most recommended solution here is to utilize the miniPCI version of Teltonika modem or an alternative external 4G/5G router also from Teltonika. These modems have been integrated with the prototype Controller.

10.2.2 Recommendations for video controller architectures

Summarizing the findings of the QoE metric-based congestion controller evaluation, it is recommended to consider the following adjustments when designing video controllers for use with wireless mobile networks:

- 1) **Adjust video bitrate followed with the resolution updates** - it can be problematic to decrease video's bit rate significantly while not changing its resolution.
- 2) **Transcoder (TR) buffer adjustment** - It is important to know that the transcoder by default is set to inject a small latency, which means that it contains a buffer of customized frame size that causes about delay according to how much frames such buffer contains. Therefore, the operator is able to customize the buffer size. If the buffer's size increases the quality of a video is better but it decreases the reaction rate at the transmitter, when receiving information about signal quality of a network.
- 3) **Video codec vulnerability to network packet losses** (e.g. H264) – considering optimizations of the feedback loop it needs to be considered that even single video frame lost may cause further loss effect. Therefore it might be an option (for future) to consider using MJPEG standard as alternative solution for adaptation of video encoding, which is not causing many bad effects for entire traffic after losing few frames.
- 4) **Video stream modifications at camera level** - dealing with dynamically changing network conditions it is also possible to not perform any major changes within transcoder in terms of bit rate of a video stream, but instead change the camera's profile, which can be pre-set before scenario. But this is an alternative solution not evaluated in this thesis due to its scalability issues. Each camera vendor offers its proprietary API, with various configuration parameters, and technical specificities to be considered for successful realization of such adjustment.
- 5) **To deal with QoE evaluation and emulation on the single machine**, the provided computing resources need to consider the minimum requirements, otherwise the distortions at QoE level may not only be due to wireless emulation but due to not enough computing power. And such situations might not be easy to identify.

10.2.3 Implications of CPU prediction for the RRM in future networks

Let us assume availability of a well-established O-RAN network or testbed. In a general AI/ML workflow defined by O-RAN alliance, the machine learning (ML) models need to be trained and validated before deployment. Therefore, the LSTM models introduced in chapter 7, should be trained in a machine learning and training host (MTH). Such host can be located in vicinity of non-real-time RIC (NRT RIC) or at Service Management and Orchestration (SMO) but outside the NRT RIC or even near-real-time RIC (nRT RIC). Since it takes time to train a decent deep neural network model, it may be more realistic to host MTH in SMO (outside NRT RIC). Data (such as workload statistics and measurements) from edge servers is collected and used to train the initial model. The fully trained and well validated model is then sent to the NRT RIC which acts as a “machine learning inference host” (MIH). This is where prediction is performed. The NRT RIC can send the inference to the nRT RIC via A1 interface so that the latter can perform such action. This could be done with the xApp in a nRT RIC. For example, with a high workload forecasted for a particular edge server, the traffic steering xApp may operate to offload traffic to servers with lower workload (forecasted) in advance to avoid possible congestion. The performance of the model is monitored and retrained when degradation is detected.

The architecture of utilizing LSTM based workload prediction models was proposed, considering industrial trends of using cloud orchestrators like the Google Kubernetes along with AI/ML middleware architectures for edge data centres [22]. Such solution was introduced into the commercial 5G ORAN system as it has been identified by vendor as essential enabler of highly adaptive network.

10.3 RECOMMENDATIONS FOR FUTURE WORK

Author would like to indicate that although the assumed research problems (and related tasks) were accomplished – as presented in the previous sections and chapters (4-8) – the research work performed has brought multiple inspirations for the further research. Below the most important ones are mentioned:

- The research problem addressed in the chapter 7 has been accomplished to the level of studying the optimal model for prediction of the CPU consumption and as such properly addresses the research questions and task identified. Still the next step could be to i) build open data-set with more comprehensive set of measurement scenarios to deliver reliable data for the community, to enable model design and evaluation as according to the best knowledge of the author availability of such models is very limited, ii) moreover the design of the workload placement agent will need to follow and augment the workload prediction model in order to assure complete solution. Author has already contributed above

proposal of predication based CU-UP scaling, towards proof-of-concept implementation in the real 5G ORAN network.

- The CAC algorithms described in chapters 4 and 8 should further be studied from the particular perspective of the cell-free networks as well as the NOMA networks. These two according to the prior art analysis (chapter 2) and author best understanding are important directions to study due to expected growth in deployments of small-cell networks in the coming years. The process already started and operators will soon see the outbreak of the small cells in their vicinity – these can be crowdsources (community driven) approaches or deployments driven by e.g. the OTT players in the competitive market (e.g. Google). Here of special interest is the general model for CAC that can deal with multiple technologies at same time (e.g. multi-RAT, NOMA, cell-free) as the networks will be combining multiple techniques to maximize the capacity. And the increase in available spectrum, within the mid-band but also available at the mmWave bands will only be able to offload some part of the traffic (say 20-45%).
- The above future extensions would particularly need to consider
 - Private, private public network configurations
 - Cooperation between cellular and cell-free networks
 - Various mix of services (i.e. verticals or slices)
 - Capacity sharing in the context of the neutral host architectures.

As future work author plans to deal with video controllers that learn from experience (i.e. history based) or can adjust to the location (also based on some previously observed network parameters in a location).

11 REFERENCES

- [1] 6G Infrastructure Association Vision and Societal Challenges Working Group Societal Needs and Value Creation Sub-Group, “What societal values will 6G address? Societal Key Values and Key Value Indicators analysed through 6G use cases,” 2022, doi: 10.5281/zenodo.6557534.
- [2] “USTAWA z dnia 16 lipca 2004 r. Prawo telekomunikacyjne (Dz.U. 2022 poz. 1648).” 2022.
- [3] É. Estaunié, *Traité pratique de télécommunication électrique (télégraphique-téléphonie)*. 1904.
- [4] J. Zander and P. Mähönen, “Riding the data tsunami in the cloud: myths and challenges in future wireless access,” *IEEE Communications Magazine*, vol. 51, no. 3, pp. 145–151, Mar. 2013, doi: 10.1109/MCOM.2013.6476879.
- [5] J. Flizikowski, *Rozprawa o innowacji*. Uniwersytet Technologiczno-Przyrodniczy w Bydgoszczy, 2021.
- [6] European Commission, “EUROPE 2020 A strategy for smart, sustainable and inclusive growth,” 2010.
- [7] “Overview of IMT standards for Global Wireless Communication,” 2023.
- [8] B. M. Khorsandi, “Latest Trends of Use cases for 6G in Hexa-X,” *Nokia Strategy & Technology Hexa-X WPI Lead*. 2023.
- [9] Tarik Taleb *et al.*, “6G System architecture: A service of services vision,” *ITU Journal on Future and Evolving Technologies*, vol. 3, no. 3, pp. 710–743, Dec. 2022, doi: 10.52953/DGKO1067.
- [10] V. Ranjbar, A. Girycki, M. A. Rahman, S. Pollin, M. Moonen, and E. Vinogradov, “Cell-Free mMIMO Support in the O-RAN Architecture: A PHY Layer Perspective for 5G and Beyond Networks,” *IEEE Communications Standards Magazine*, vol. 6, no. 1, pp. 28–34, Mar. 2022, doi: 10.1109/MCOMSTD.0001.2100067.
- [11] F. Kooshki, M. A. Rahman, M. M. Mowla, A. G. Armada, and A. Flizikowski, “Efficient Radio Resource Management for Future 6G Mobile Networks: A Cell-less Approach,” *IEEE Networking Letters*, pp. 1–1, 2023, doi: 10.1109/LNET.2023.3263926.
- [12] F. Kooshki, A. G. Armada, M. M. Mowla, and A. Flizikowski, “Radio Resource Management Scheme for URLLC and eMBB Coexistence in a Cell-Less Radio Access Network,” *IEEE Access*, vol. 11, pp. 25090–25101, 2023, doi: 10.1109/ACCESS.2023.3256528.
- [13] Coleago Consulting, “Estimating the mid-band spectrum needs in the 2025-2030 time frame Global Outlook A report by Coleago Consulting Ltd,” 2021. [Online]. Available: www.coleago.com
- [14] M. Camelo *et al.*, “DAEMON: A Network Intelligence Plane for 6G Networks,” in *2022 IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 1341–1346. doi: 10.1109/GCWkshps56602.2022.10008662.
- [15] J. Vardakas, “MARSAL’s network architecture specifications - Deliverable D2.2,” Jan. 2022. Accessed: Jun. 19, 2023. [Online]. Available: https://www.marsalproject.eu/wp-content/uploads/2022/09/D2_2_V1.0.pdf
- [16] J. Zhang, Y. Wei, E. Björnson, Y. Han, and X. Li, “Spectral and energy efficiency of cell-free massive MIMO systems with hardware impairments,” in *2017 9th*

- International Conference on Wireless Communications and Signal Processing (WCSP)*, 2017, pp. 1–6. doi: 10.1109/WCSP.2017.8171057.
- [17] A. P. Guevara, C.-M. Chen, A. Chiumento, and S. Pollin, “Partial Interference Suppression in Massive MIMO Systems: Taxonomy and Experimental Analysis,” *IEEE Access*, vol. 9, pp. 128925–128937, 2021, doi: 10.1109/ACCESS.2021.3113167.
- [18] F. Kaltenberger, C. Roux, M. Buczkowski, and M. Wewior, “The OpenAirInterface application programming interface for schedulers using Carrier Aggregation,” in *2016 International Symposium on Wireless Communication Systems (ISWCS)*, 2016, pp. 497–500. doi: 10.1109/ISWCS.2016.7600955.
- [19] M. O. Ojijo and O. E. Falowo, “A Survey on Slice Admission Control Strategies and Optimization Schemes in 5G Network,” *IEEE Access*, vol. 8, pp. 14977–14990, 2020, doi: 10.1109/ACCESS.2020.2967626.
- [20] A. Lacava *et al.*, “Programmable and Customized Intelligence for Traffic Steering in 5G Networks Using Open RAN Architectures.” 2022.
- [21] I. Sarrigiannis, K. Ramantas, E. Kartsakli, P.-V. Mekikis, A. Antonopoulos, and C. Verikoukis, *Online VNF Lifecycle Management in a MEC-enabled 5G IoT Architecture*, vol. PP. 2019. doi: 10.1109/JIOT.2019.2944695.
- [22] “The BRAINE,” May 01, 2021. <https://cordis.europa.eu/project/id/876967> (accessed Jun. 19, 2023).
- [23] S. Pramanik, A. Ksentini, and C. F. Chiasserini, “Characterizing the Computational and Memory Requirements of Virtual RANs,” in *2022 17th Wireless On-Demand Network Systems and Services Conference (WONS)*, 2022, pp. 1–8. doi: 10.23919/WONS54113.2022.9764455.
- [24] “6G SandBox,” 2023.
- [25] A. Akbar, S. Jangsher, and F. A. Bhatti, “NOMA and 5G emerging technologies: A survey on issues and solution techniques,” *Computer Networks*, vol. 190, p. 107950, 2021, doi: <https://doi.org/10.1016/j.comnet.2021.107950>.
- [26] ComSoc WTC SIG RSMA, “2nd Season of RSMA Webinar Series,” 2022. <https://sites.google.com/view/ieee-comsoc-wtc-sig-rsma/events/webinar-series> (accessed Jun. 19, 2023).
- [27] Y. Liu *et al.*, “Evolution of NOMA Toward Next Generation Multiple Access (NGMA) for 6G,” *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 4, pp. 1037–1071, 2022, doi: 10.1109/JSAC.2022.3145234.
- [28] KEYSIGHT TECHNOLOGIES BELGIUM, “6G-SANDBOX (Supporting Architectural and technological Network evolutions through an intelligent, secureD and twinning enaBled Open eXperimentation facility),” *EU Commission - contract no. 101096328*, Jan. 01, 2023. <https://6g-sandbox.eu/> (accessed Jun. 20, 2023).
- [29] M. K. Bahare *et al.*, “The 6G Architecture Landscape European Perspective,” 2023. doi: 10.5281/zenodo.7313232.
- [30] D. Lowenstein and C. Mueth, “Implementing a Digital Twin, Design and Test, Test and Measurement Strategy,” in *2022 IEEE AUTOTESTCON*, 2022, pp. 1–6. doi: 10.1109/AUTOTESTCON47462.2022.9984739.
- [31] A. N. Manjeshwar, P. Jha, A. Karandikar, and P. Chaporkar, “Enhanced UE Slice Mobility for 5G Multi-RAT Networks,” in *2019 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2019, pp. 1–6. doi: 10.1109/NFV-SDN47374.2019.9039982.

- [32] H. Choi, D. Han, and W. Na, "Research Challenge on MPTCP in 5G/6G Networks," in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 2022, pp. 856–858. doi: 10.1109/ICTC55196.2022.9952783.
- [33] M. Agiwal, H. Kwon, S. Park, and H. Jin, "A Survey on 4G-5G Dual Connectivity: Road to 5G Implementation," *IEEE Access*, vol. 9, pp. 16193–16210, 2021, doi: 10.1109/ACCESS.2021.3052462.
- [34] A. Guidotti *et al.*, *The path to 5G-Advanced and 6G Non-Terrestrial Network systems*. 2022. doi: 10.1109/ASMS/SPSC55670.2022.9914764.
- [35] A. Flizikowski, R. Kozik, M. Majewski, and M. Przybyszewski, "Evaluation of guard channel admission control schemes for IEEE 802.16 with integrated nb-LDPC codes," in *2009 International Conference on Ultra Modern Telecommunications & Workshops*, IEEE, Oct. 2009, pp. 1–8. doi: 10.1109/ICUMT.2009.5345468.
- [36] A. Flizikowski, W. Hołubowicz, M. Przybyszewski, and S. Grzegorzewski, "Admission Control and System Capacity Assessment of WiMAX with ACM and nb-LDPC Codes Simulation Study with ViMACCS ns2 Patch," in *2010 24th IEEE International Conference on Advanced Information Networking and Applications*, IEEE, 2010, pp. 1077–1084. doi: 10.1109/AINA.2010.154.
- [37] A. Flizikowski, M. Przybyszewski, and W. Hołubowicz, "Evaluation of Measurement Based Admission Control Algorithms for IEEE 802.16 Networks in Simulations with L2S Physical Layer Abstraction and nbLDPC Codes," 2010, pp. 447–459. doi: 10.1007/978-3-642-16295-4_50.
- [38] A. Flizikowski, R. Kozik, M. Majewski, and M. Przybyszewski, "Performance comparison of guard channel admission control schemes for IEEE 802.16 system with various turbo code FEC schemes," in *2009 IEEE 28th International Performance Computing and Communications Conference*, IEEE, Dec. 2009, pp. 360–365. doi: 10.1109/PCCC.2009.5403853.
- [39] A. Flizikowski, R. Kozik, H. Gierszal, M. Przybyszewski, and W. Hołubowicz, "WiMAX cell level simulation platform based on ns-2 and DSP integration," *International Journal of Electronics and Telecommunications*, vol. 56, no. 2, pp. 169–176, Jun. 2010, doi: 10.2478/v10177-010-0022-3.
- [40] A. Flizikowski, M. Majewski, and M. Przybyszewski, "Evaluation of Optimal Resource Management Policies for WiMAX Networks with AMC: A Reinforcement Learning Approach," 2011, pp. 459–467. doi: 10.1007/978-3-642-23154-4_50.
- [41] A. Flizikowski, M. Majewski, M. Przybyszewski, and W. Houbowicz, "Evaluation of QoS and QoE in Mobile WIMAX – Systematic Approach," in *Quality of Service and Resource Allocation in WiMAX*, InTech, 2012. doi: 10.5772/27891.
- [42] A. Flizikowski, M. Przybyszewski, T. Olejniczak, and M. Płócienniczak, "An approach to video-streaming tests in mobile WIMAX using low-cost time-reference," *Studia Informatica*, vol. 36, no. 2, pp. 43–58, 2015.
- [43] A. Flizikowski, T. Marciniak, T. A. Wysocki, and O. Oyerinde, "Selected Aspects of Non orthogonal Multiple Access for Future Wireless Communications," *Mathematics in Computer Science*, vol. 17, no. 2, p. 10, 2023, doi: 10.1007/s11786-023-00561-y.
- [44] M. A. Rahman, A. Flizikowski, S. Pietrzyk, and M. M. Mowla, "Design and

- Experimental Validation of Radio Access Network Controller Prototype for Multi-RAT Technologies with Scheduler Strategies,” in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, IEEE, Sep. 2021, pp. 1506–1511. doi: 10.1109/PIMRC50174.2021.9569396.
- [45] A. Flizikowski, E. Alkhovik, M. Munjure Mowla, and M. Arifur Rahman, “Data Handling Mechanisms and Collection Framework for 5G vRAN in Edge Networks,” in *2022 IEEE Conference on Standards for Communications and Networking (CSCN)*, IEEE, Nov. 2022, pp. 36–41. doi: 10.1109/CSCN57023.2022.10051118.
- [46] A. Flizikowski, E. Alkhovik, M. M. Mowla, and M. A. Rahman, “Importance of Workload Prediction of Virtualized RAN in the Importance of Workload Prediction of Virtualized RAN in the Edge Micro Data Center Edge Micro Data Center,” *TechRxiv*, 2022, doi: 10.36227/techrxiv.21644708.v1.
- [47] A. A. I. M. M. M. F. A. A. M. A. B. Z. D. Mämmelä, “Sustainability in 6G Networks: Vision and Directions.,” *TechRxiv*, 2022.
- [48] T. M. Chen and S. S. Liu, *ATM Switching Systems*. USA: Artech House, Inc., 1995.
- [49] M. Dąbrowski, W. Burakowski, and A. Bęben, “On effectiveness of conditional admission control for IP QoS network services with REM scheme,” *Journal of Telecommunications and Information Technology*, vol. nr 2, pp. 33–39, 2002.
- [50] A. Kumar, D. Manjunath, and J. Kuri, *Communication Networking: An Analytical Approach*, 1st ed. Elsevier, 2004.
- [51] A. Flizikowski and M. Plócienniczak, “Framework for Evaluating QoE for Remote Control of Autonomous Cars in Mobile Wireless Networks,” 2017, pp. 170–182. doi: 10.1007/978-3-319-68720-9_20.
- [52] International Telecommunication Union, “I.371 : Traffic control and congestion control in B-ISDN,” 2004.
- [53] R. N. M. Hassan, *Call Admission Control for Wireless Broadband Networks Paperback*. LAP Lambert Academic Publishing, 2015.
- [54] Juan Ramiro and Khalid Hamied, *Self-Organizing Networks*. Wiley, 2011. doi: 10.1002/9781119954224.
- [55] International Telecommunication Union, “Y.1291: An architectural framework for support of Quality of Service in packet networks,” 2004.
- [56] European Telecommunications Standards Institute, “ETSI TS 123 203 V14.3.0 (2017-05) Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; Policy and charging control architecture (3GPP TS 23.203 version 14.3.0 Release 14),” 2017.
- [57] International Telecommunication Union, “Y.2171: Admission control priority levels in Next Generation Networks,” 2006.
- [58] M. Polese, L. Bonati, S. D’Oro, S. Basagni, and T. Melodia, “Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges,” *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1376–1411, 2023, doi: 10.1109/COMST.2023.3239220.
- [59] A. Chowdary, G. Chopra, A. Kumar, and L. R. Cenkeramaddi, “Enhanced User Grouping and Pairing Scheme for CoMP-NOMA-based Cellular Networks,” in *2022 14th International Conference on COMMunication Systems & NETWORKS*

- (COMSNETS), IEEE, Jan. 2022, pp. 319–323. doi: 10.1109/COMSNETS53615.2022.9668568.
- [60] K. Bahram *et al.*, “Survey on 5G Second Phase RAN Architectures and Functional Survey on 5G Second Phase RAN Architectures and Functional splits splits,” 2022, doi: 10.36227/techrxiv.21280473.v1.
- [61] M. Alavirad *et al.*, “O-RAN architecture, interfaces, and standardization: Study and application to user intelligent admission control,” *Frontiers in Communications and Networks*, vol. 4, Mar. 2023, doi: 10.3389/frcmn.2023.1127039.
- [62] R. Bryś, J. Pszczółkowski, and M. Ruszkowski, “Mechanizmy QoS płaszczyzny sterowania w systemach specjalnych - wyniki badań symulacyjnych,” *Przegląd Telekomunikacyjny + Wiadomości Telekomunikacyjne*, vol. nr 11, pp. 1574–1584, 2011.
- [63] M. A. Callejo-Rodriguez *et al.*, “EuQoS: End-To-End QoS over Heterogeneous Networks,” in *2008 First ITU-T Kaleidoscope Academic Conference - Innovations in NGN: Future Network and Services*, IEEE, May 2008, pp. 177–184. doi: 10.1109/KINGN.2008.4542264.
- [64] H. Tarasiuk, R. Janowski, and W. Burakowski, “Admissible traffic load of real time class of service for inter-domain peers,” in *Joint International Conference on Autonomic and Autonomous Systems and International Conference on Networking and Services - (icas-isns'05)*, IEEE, 2005, pp. 63–63. doi: 10.1109/ICAS-ICNS.2005.16.
- [65] A. Sang and S. Li, “A predictability analysis of network traffic,” in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, 2000, pp. 342–351 vol.1. doi: 10.1109/INFCOM.2000.832204.
- [66] B. Rong, B. Tremblay, M. Bennani, and M. Kadoch, “Integrating traffic aggregation mechanism into SIP based IP telephony over MPLS network,” in *GLOBECOM '05. IEEE Global Telecommunications Conference, 2005.*, 2005, pp. 5 pp.-. doi: 10.1109/GLOCOM.2005.1577749.
- [67] Y. Ou and L. Jianhua, “Call Admission Control and Scheduling Schemes with QoS Support for Real-time Video Applications in IEEE 802.16 Networks,” *J Multimed*, vol. 1, May 2006, doi: 10.4304/jmm.1.2.21-29.
- [68] “5G NR QoS (Quality of Service) | 5G QoS as per 3GPP NR Standard.” <https://www.rfwireless-world.com/5G/5G-NR-QoS.html> (accessed Jun. 21, 2023).
- [69] K. Kousias *et al.*, “Implications of Handover Events in Commercial 5G Non-Standalone Deployments in Rome,” in *Proceedings of the ACM SIGCOMM Workshop on 5G and Beyond Network Measurements, Modeling, and Use Cases*, in 5G-MeMU '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 22–27. doi: 10.1145/3538394.3546041.
- [70] S. Alotaibi, “Key Challenges of Mobility Management and Handover Process In 5G HetNets,” *IJCSNS International Journal of Computer Science and Network Security*, vol. 22, no. 4, doi: 10.22937/IJCSNS.2022.22.4.18.
- [71] European Telecommunications Standards Institute, “ETSI TS 138 300 V15.4.0 (2019-04) 5G;NR; Overall description; Stage-2 (3GPP TS 38.300 version 15.4.0 Release 15),” 2019. Accessed: Jun. 21, 2023. [Online]. Available:

- https://www.etsi.org/deliver/etsi_ts/138300_138399/138300/15.04.00_60/ts_138300v150400p.pdf
- [72] R. J. and H. P. J. Gibbens, "Effective bandwidths for the Multi-type UAS channel," *Queueing Systems*, pp. 17–28, 1991.
- [73] "Andrew S. Tanenbaum - Computer Networks".
- [74] European Commission, "Self-Optimisation and Self-Configuration in Wireless Networks," *SOCRATES*, Jan. 01, 2008. <https://fp7-socrates.org/> (accessed Jun. 20, 2023).
- [75] S. Ghosh and A. Konar, *Call Admission Control in Mobile Cellular Networks*, vol. 437. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-30997-7.
- [76] Milbrandt J. and Menth M., "Experience-Based Admission Control with Type-Specific Overbooking," *Lecture Notes in Computer Science*, vol. 4268, 2006, doi: https://doi.org/10.1007/11908852_7.
- [77] P. Duc, D. Chuong, and V. M. N. Vo, *A Model of Traffic Prediction Based Admission Control in OBS Nodes*. 2019. doi: 10.1109/RIVF.2019.8713683.
- [78] I. Ivars, "Probe-based admission control in IP networks," Jan. 2003.
- [79] T. Jo, *Machine Learning Foundations*. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3-030-65900-4.
- [80] Kun Wu, "OPNET Implementation of Endpoint Admission Control Algorithms," 2003.
- [81] I. Vilà, O. Sallent, and J. Pérez-Romero, "On the Design of a Network Digital Twin for the Radio Access Network in 5G and Beyond," *Sensors*, vol. 23, no. 3, p. 1197, Jan. 2023, doi: 10.3390/s23031197.
- [82] A. C. E. T. T. 3GPP Organizational Partners (ARIB, "3GPP TS 23.203 V9.9.0 (2011-06) 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Policy and charging control architecture (Release 9)," 2011. Accessed: Jun. 22, 2023. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/23_series/23.203/
- [83] A. C. E. T. T. T. 3GPP Organizational Partners (ARIB, "System architecture for the 5G System (5GS) 3GPP TS 23.501 V0.0 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2;0 (2017-01)," 2017, Accessed: Jun. 22, 2023. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/23_series/23.501/
- [84] International Telecommunication Union, "Y.3113: Requirements and framework for latency guarantee in large-scale networks including the IMT-2020 network," 2021. Accessed: Jun. 22, 2023. [Online]. Available: <https://www.itu.int/rec/T-REC-Y.3113-202102-I/en>
- [85] European Telecommunications Standards Institute (ETSI), "ETSI TS 123 501 V16.6.0 (2020-10) 5G; System architecture for the 5G System (5GS) (3GPP TS 23.501 version 16.6.0 Release 16)," 2020. Accessed: Jun. 22, 2023. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/123500_123599/123501/16.06.00_60/ts_123501v160600p.pdf
- [86] N. Nathani and G. C. Manna, "A Quality of Service Based Model for Supporting Mobile Secondary Users in Cognitive Radio Technology," in *Cognitive Radio in 4G/5G Wireless Communication Systems*, IntechOpen, 2018. doi:

- 10.5772/intechopen.80072.
- [87] R. K. and D. P. K. and S. S. D. Laishram Romesh and Mangang, "An Adaptive Call Admission Control in WiMAX Networks with Fair Trade off Analysis," in *Advances in Communication, Network, and Computing*, J. Das Vinu V. and Stephen, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 426–430.
- [88] K. E. Suleiman, A.-E. M. Taha, and H. S. Hassanein, "Understanding the interactions of handover-related self-organization schemes," in *Proceedings of the 17th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems*, New York, NY, USA: ACM, Sep. 2014, pp. 285–294. doi: 10.1145/2641798.2641830.
- [89] J. Prados-Garzon, O. Adamuz-Hinojosa, P. Ameigeiras, J. J. Ramos-Munoz, P. Andres-Maldonado, and J. M. Lopez-Soler, "Handover implementation in a 5G SDN-based mobile network architecture," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2016, pp. 1–6. doi: 10.1109/PIMRC.2016.7794936.
- [90] E. Hossain, L. Le, and D. Niyato, *Radio Resource Management in Multi-Tier Cellular Wireless Networks*. 2013. doi: 10.1002/9781118749821.
- [91] 3GPP, "Telecommunication Management; Self-Organizing Networks (SON); Concepts and Requirements," *3rd Generation Partnership Project (3GPP), TS 32.500*, 2008.
- [92] H. Y. Lateef, A. Imran, M. A. Imran, L. Giupponi, and M. Dohler, "LTE-advanced self-organizing network conflicts and coordination algorithms," *IEEE Wirel Commun.*, vol. 22, no. 3, pp. 108–117, 2015, doi: 10.1109/MWC.2015.7143333.
- [93] A. Bayazeed, K. Khorzom, and M. Aljnidi, "A survey of self-coordination in self-organizing network," *Computer Networks*, vol. 196, p. 108222, 2021, doi: <https://doi.org/10.1016/j.comnet.2021.108222>.
- [94] D. K. Arrowsmith and R. J. Mondragón, "Modelling Network Data Traffic," 2006.
- [95] M. Newman, "Power Laws, Pareto Distributions and Zipf's Law," *Contemporary Physics - CONTEMP PHYS*, vol. 46, Dec. 2004, doi: 10.1080/00107510500052444.
- [96] P. P. Mariño, *Optimization of Computer Networks - Modeling and Algorithms*. Chichester, UK: John Wiley & Sons, Ltd, 2016. doi: 10.1002/9781119114840.
- [97] MOHANAD AMER and SHISHIRA PUTTASWAMY, "Traffic Model for Cellular Network Analysis - Master Thesis," 2019.
- [98] E. Gelenbe, *System Performance Evaluation*. CRC Press, 2000. doi: 10.1201/9781482274530.
- [99] M. Zukerman, "Introduction to Queueing Theory and Stochastic Teletraffic Models," Jul. 2013.
- [100] P. Schulz, "Queueing-Theoretic End-to-End Latency Modeling of Future Wireless Networks." 2019.
- [101] M. J. Dąbrowski, "Zaawansowane metody sterowania przyjmowaniem nowych wywołań w sieciach QoS IP," 2004.
- [102] A. Morton, Y. E. Mghazli, M. Dolly, P. Tarapore, G. Ash, and C. Dvorak, "Y.1541-QOSM: Model for Networks Using Y.1541 Quality-of-Service Classes RFC 5976," 2010. Accessed: Jun. 22, 2023. [Online]. Available: <https://datatracker.ietf.org/doc/rfc5976/>

- [103] Kelly, *Do sprawdzenia*. 1991.
- [104] Stasiak, *Do sprawdzenia A*. 2009.
- [105] Stasiak, *Do sprawdzenia B*. 2009.
- [106] ProbabilisticCAC, “ProbabilisticCAC.”
- [107] N. Ansari, H. Liu, Y. Q. Shi, and H. Zhao, “On modeling MPEG video traffics,” *IEEE Transactions on Broadcasting*, vol. 48, no. 4, pp. 337–347, 2002, doi: 10.1109/TBC.2002.806794.
- [108] L. J. D. la Cruz, E. Pallares, J. J. Alins, and J. Mata, “Self-similar traffic generation using a fractional ARIMA model. Application to the VBR MPEG video traffic,” in *ITS’98 Proceedings. SBT/IEEE International Telecommunications Symposium (Cat. No.98EX202)*, 1998, pp. 102–107 vol.1. doi: 10.1109/ITS.1998.713099.
- [109] Y. Liang and M. Han, “Dynamic Bandwidth Allocation Based on Online Traffic Prediction for Real-Time MPEG-4 Video Streams,” *EURASIP J Adv Signal Process*, vol. 2007, no. 1, p. 087136, 2006, doi: 10.1155/2007/87136.
- [110] Oliveira2016, *Oliveira2016*. 2016.
- [111] J. Gamboa, “Deep Learning for Time-Series Analysis,” Jan. 2017.
- [112] R. Bonetto and M. Rossi, “Smart Grid for the Smart City,” in *Designing, Developing, and Facilitating Smart Cities: Urban Design to IoT Solutions*, V. Angelakis, E. Tragos, H. C. Pöhls, A. Kapovits, and A. Bassi, Eds., Cham: Springer International Publishing, 2017, pp. 241–263. doi: 10.1007/978-3-319-44924-1_12.
- [113] W. Hruday and L. Trajković, “Mobile WiMAX MAC and PHY layer optimization for IPTV,” *Math. Comput. Model.*, vol. 53, pp. 2119–2135, 2011.
- [114] K. Suh *et al.*, “Push-to-Peer Video-on-Demand System: Design and Evaluation,” *Selected Areas in Communications, IEEE Journal on*, vol. 25, pp. 1706–1716, Jan. 2008, doi: 10.1109/JSAC.2007.071209.
- [115] K. Shuaib and F. Sallabi, “Smoothing of video transmission rates for an LTE network,” in *2010 IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications*, IEEE, Oct. 2010, pp. 713–719. doi: 10.1109/WIMOB.2010.5644857.
- [116] A. A. Djamal, E. Meddour, T. Ahmed, and T. Rasheed, “Cross layer design for optimized video streaming over heterogeneous networks,” in *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*, New York, NY, USA: ACM, Jun. 2010, pp. 933–938. doi: 10.1145/1815396.1815609.
- [117] J. Lakkakorpi and A. Sayenko, “Measurement-Based Connection Admission Control Methods for Real-Time Services in IEEE 802.16e,” in *2009 Second International Conference on Communication Theory, Reliability, and Quality of Service*, IEEE, Jul. 2009, pp. 37–41. doi: 10.1109/CTRQ.2009.9.
- [118] J. Aein, “A Multi-User-Class, Blocked-Calls-Cleared, Demand Access Model,” *IEEE Transactions on Communications*, vol. 26, no. 3, pp. 378–385, 1978, doi: 10.1109/TCOM.1978.1094081.
- [119] D. Chiarotto, L. Badia, and M. Zorzi, “Soft capacity of OFDMA networks is suitable for soft QoS multimedia traffic,” in *2013 IEEE International Conference on Communications (ICC)*, 2013, pp. 2496–2501. doi: 10.1109/ICC.2013.6654908.
- [120] A. G. Burr, A. Papadogiannis, and T. Jiang, “MIMO Truncated Shannon Bound

- for system level capacity evaluation of wireless networks,” *2012 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 268–272, 2012.
- [121] Tse David and Pramod Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.
- [122] T. Jämsä, J. Medbo, P. Kyösti, K. Haneda, and L. Raschkowski, “The 5G wireless propagation channel models,” in *5G Mobile and Wireless Communications Technology*, A. Osseiran, J. F. Monserrat, and P. Marsch, Eds., Cambridge: Cambridge University Press, 2016, pp. 357–380. doi: DOI: 10.1017/CBO9781316417744.014.
- [123] Bogucka H., *Projektowanie i obliczenia w radiokomunikacji. Wybrane zagadnienia*. 2005.
- [124] E. De Santis, A. Giuseppi, A. Pietrabissa, M. Capponi, and F. Delli Priscoli, “Satellite Integration into 5G: Deep Reinforcement Learning for Network Selection,” *Machine Intelligence Research*, vol. 19, no. 2, pp. 127–137, Apr. 2022, doi: 10.1007/s11633-022-1326-3.
- [125] European Commission, “Beyond Next Generation Mobile Networks BuNGee Project.” 2012. Accessed: Jun. 22, 2023. [Online]. Available: <https://cordis.europa.eu/docs/projects/cnect/7/248267/080/deliverables/001-BuNGeeD23UOYv1.pdf>
- [126] A. M. Ahmadzadeh, “Capacity and cell-range estimation for multitraffic users in mobile WiMAX,” Master thesis, , University College of Borås, Sweden, 2008.
- [127] T. Chen, R. Boreli, and T. Iyer, “A cross layer scheme for maximising the combination of wireless VoIP capacity and quality,” in *Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly*, New York, NY, USA: ACM, Jun. 2009, pp. 148–153. doi: 10.1145/1582379.1582413.
- [128] WiMAX_Analytical#A, *WiMAX_Analytical#A*.
- [129] A. C. E. T. T. T. 3GPP Organizational Partners (ARIB, “3GPP TS 38.306 NR; User Equipment (UE) radio access capabilities 38.306,” 2017. Accessed: Jun. 22, 2023. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3193>
- [130] A. A. El-Saleh, M. A. Al Jahdhami, A. Alhammadi, Z. A. Shamsan, I. Shaye, and W. H. Hassan, “Measurements and Analyses of 4G/5G Mobile Broadband Networks: An Overview and a Case Study,” *Wirel Commun Mob Comput*, vol. 2023, p. 6205689, 2023, doi: 10.1155/2023/6205689.
- [131] “Single User Throughput in 5G NR.” Accessed: Jun. 22, 2023. [Online]. Available: https://www.ofcom.org.uk/__data/assets/file/0033/195549/sut-model-700mhz-3.6-3.8ghz-spectrum.xlsm
- [132] S. Klisara, G. Nermin, and E. Avdagić-Golub, “Rough estimation of cell numbers in 5G networks using simple mathematical calculations,” *Science, Engineering and Technology*, vol. 1, no. 2, pp. 1–7, Oct. 2021, doi: 10.54327/set2021/v1.i2.15.
- [133] B. Sklar and F. (Fredric J.) Harris, *Digital communications : fundamentals and applications*.
- [134] M. P. Mota, D. C. Araujo, F. H. Costa Neto, A. L. F. de Almeida, and F. R. Cavalcanti, “Adaptive Modulation and Coding Based on Reinforcement Learning for 5G Networks,” in *2019 IEEE Globecom Workshops (GC Wkshps)*, 2019, pp.

- 1–6. doi: 10.1109/GCWkshps45667.2019.9024384.
- [135] A. C. E. T. T. 3GPP Organizational Partners (ARIB, “3GPP TR 36.942 Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios,” 2022. Accessed: Jun. 22, 2023. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2592>
- [136] E. Makridis, “Reinforcement Learning for Link Adaptation in 5G-NR Networks.”
- [137] A. Alexiou, C. Bouras, and E. Rekkas, “An improved MBMS power counting mechanism towards long term evolution,” *Telecommun Syst*, vol. 43, no. 1, p. 109, 2009, doi: 10.1007/s11235-009-9197-2.
- [138] N. Chukhno *et al.*, “The Use of Machine Learning Techniques for Optimal Multicasting in 5G NR Systems,” *IEEE Transactions on Broadcasting*, vol. 69, no. 1, pp. 201–214, Mar. 2023, doi: 10.1109/TBC.2022.3206595.
- [139] S. Satya Sri Ganesh Seeram, “Degree Project in Electrical Engineering, specializing in Communication Systems Second Cycle 30.0 credits Link Adaptation in 5G Networks Reinforcement Learning Framework based Approach.”
- [140] H. Gierszal, W. Hołubowicz, Ł. Kiedrowski, and A. Flizikowski, “Performance of non-binary LDPC codes for next generation mobile systems,” *International Journal of Electronics and Telecommunications*, vol. 56, no. 2, pp. 111–116, Jun. 2010, doi: 10.2478/v10177-010-0014-3.
- [141] Iisaku Shun-Ichi and Urano Yoshiyori, “Performance analysis of integrated communication system with heterogeneous traffic,” in *Proc. 11th Int. Teletraffic Congress*, 1985, pp. 72–77.
- [142] Randhawa 2015, “Uzupełnić”
- [143] C. P. Vassal, M. Tanelli, and M. Lovera, “Dynamic Trade-off Analysis of QoS and Energy Saving in Admission Control for Web Service Systems,” in *Proceedings of the 4th International ICST Conference on Performance Evaluation Methodologies and Tools*, ICST, 2009. doi: 10.4108/ICST.VALUETOOLS2009.7941.
- [144] E. Yavuz and V. C. M. Leung, “Efficient Methods for Performance Evaluations of Call Admission Control Schemes in Multi-Service Cellular Networks,” *IEEE Trans Wirel Commun*, vol. 7, no. 9, pp. 3468–3476, Sep. 2008, doi: 10.1109/TWC.2008.070280.
- [145] Ren-Hung Hwang, J. F. Kurose, and D. Towsley, “MDP routing in ATM networks using virtual path concept,” in *Proceedings of INFOCOM '94 Conference on Computer Communications*, IEEE Comput. Soc. Press, pp. 1509–1517. doi: 10.1109/INFCOM.1994.337531.
- [146] D. Nguyen, T. Nguyen, and X. Yang, “Multimedia wireless transmission with network coding,” in *Packet Video 2007*, IEEE, Nov. 2007, pp. 326–335. doi: 10.1109/PACKET.2007.4397057.
- [147] Asars A., “Determining the optimal interval of the parameter identification of self-similar traffic,” *Automatic Control and Computer Sciences*, doi: <https://doi.org/10.3103/S0146411609040075>.
- [148] Sang Soo Jeong, Jeong Ae Han, and Wha Sook Jeon, “Adaptive connection admission control scheme for high data rate mobile networks,” in *VTC-2005-Fall. 2005 IEEE 62nd Vehicular Technology Conference, 2005.*, IEEE, pp. 2607–2611. doi: 10.1109/VETECF.2005.1559021.

- [149] L. Breslau, S. Jamin, and S. Shenker, "Comments on the performance of measurement-based admission control algorithms," in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, 2000, pp. 1233–1242 vol.3. doi: 10.1109/INFCOM.2000.832506.
- [150] N. G. Bean, "Estimation and Control in ATM Networks," 1994.
- [151] Kacianka Severin and Hermann Hellwagner, "Adaptive video streaming for UAV networks," in *MoVid '15: Proceedings of the 7th ACM International Workshop on Mobile Video*, 2015, pp. 25–30. doi: <https://doi.org/10.1145/2727040.2727043>.
- [152] R. Zhu and J. Yang, "Buffer-aware adaptive resource allocation scheme in LTE transmission systems," *EURASIP J Wirel Commun Netw*, vol. 2015, no. 1, p. 176, 2015, doi: 10.1186/s13638-015-0398-y.
- [153] H. Chen, L. Huang, S. Kumar, and C.-C. J. Kuo, *Radio Resource Management for Multimedia QoS Support in Wireless Networks*. Boston, MA: Springer US, 2004. doi: 10.1007/978-1-4615-0469-6.
- [154] Guo Xin and Ma Wenchao, "Dynamic Bandwidth Reservation Admission Control Scheme for the IEEE 802.16e Broadband Wireless Access Systemsupelnić," in *Wireless Communications and Networking Conference (WCNC), 2007 IEEE*, 2007. doi: 10.1109/WCNC.2007.628.
- [155] Daehyoung Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans Veh Technol*, vol. 35, no. 3, pp. 77–92, Aug. 1986, doi: 10.1109/T-VT.1986.24076.
- [156] S. Kalikivayi, I. S. Misra, and K. Saha, "Bandwidth and Delay Guaranteed Call Admission Control Scheme for QoS Provisioning in IEEE 802.16e Mobile WiMAX," in *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*, IEEE, 2008, pp. 1–6. doi: 10.1109/GLOCOM.2008.ECP.246.
- [157] B. Zhu, K. Xue, H. Lu, and P. Hong, "Fair Connection Admission Control Scheme for IEEE 802.16e Systems," in *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*, IEEE, Oct. 2008, pp. 1–4. doi: 10.1109/WiCom.2008.736.
- [158] S. Wu, K. Y. M. Wong, and B. Li, "A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 2, pp. 257–271, 2002, doi: 10.1109/90.993306.
- [159] I. Koo, S. Bahng, and K. Kim, *Resource reservation in call admission control schemes for CDMA systems with non-uniform traffic distribution among cells*, vol. 1. 2003. doi: 10.1109/VETECS.2003.1207578.
- [160] N. Bambos, S. C. Chen, and G. J. Pottie, "Channel access algorithms with active link protection for wireless communication networks with power control," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 583–597, 2000, doi: 10.1109/90.879345.
- [161] Amir Mirzaeinia, "Latency and Throughput Optimization in Modern Networks: A Comprehensive Survey," *Preprint*, 2020.
- [162] R. Laishram, *Adaptive Call Admission Control for QoS Provisioning in WiMAX*. LAP Lambert Academic Publishing, 2010.

- [163] S. Xie and M. Wu, "Optimized call admission control in wireless networks," in *ICAIT '08: Proceedings of the 2008 International Conference on Advanced Infocomm Technology*, 2008, pp. 1–7.
- [164] S. Luo and Z. Li, "A Dynamic Hierarchical CAC Mechanism for IEEE 802.16d System," in *2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, IEEE, Dec. 2008, pp. 626–630. doi: 10.1109/PACIIA.2008.107.
- [165] P. Jiang and A. Şekercioğlu, "A dynamic call admission control scheme for optimization with QoS provisioning in multiclass cellular networks," in *Proceedings of the 4th ACM symposium on QoS and security for wireless and mobile networks*, New York, NY, USA: ACM, Oct. 2008, pp. 19–26. doi: 10.1145/1454586.1454590.
- [166] I. Lokshina, "Study about effects of self-similar IP network traffic on queuing and network performance," *International Journal of Mobile Network Design and Innovation*, vol. 4, no. 2, p. 76, 2012, doi: 10.1504/IJMNDI.2012.048487.
- [167] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-Layer combining of adaptive Modulation and coding with truncated ARQ over wireless links," *IEEE Trans Wirel Commun*, vol. 3, no. 5, pp. 1746–1755, 2004, doi: 10.1109/TWC.2004.833474.
- [168] B. Fong, N. Ansari, A. C. M. Fong, and G. Y. Hong, "On the scalability of fixed broadband wireless access network deployment," *IEEE Communications Magazine*, vol. 42, no. 9, pp. S12–S18, 2004, doi: 10.1109/MCOM.2004.1336719.
- [169] S.-K. Noh, Y.-H. Hwang, B.-H. Ye, and S.-H. Kim, "New CAC Algorithm using Adaptive Modulation Control," in *Network Control and Engineering for QoS, Security and Mobility, IV*, D. Gaïti, Ed., Boston, MA: Springer US, 2007, pp. 173–186.
- [170] S. Noh, S. Hong, Y. Hwang, and B. Yae, "New call admission control mechanisms considering adaptive modulation control," in *2006 8th International Conference Advanced Communication Technology*, 2006, pp. 6 pp. – 418. doi: 10.1109/ICACT.2006.205998.
- [171] O. Alanen, "Quality of Service for Triple Play Services in Heterogeneous Networks," 2007.
- [172] J. Augé, S. Oueslati, and J. Roberts, "Measurement-based admission control for flow-aware implicit service differentiation," in *2011 23rd International Teletraffic Congress (ITC)*, 2011, pp. 206–213.
- [173] Holma Harri and Toskala Antti, *LTE Small Cell Optimization: 3GPP Evolution to Release 13*. John Wiley & Sons Ltd, 2016. doi: 10.1002/9781118912560.
- [174] S. Hansun, "A new approach of moving average method in time series analysis," in *2013 Conference on New Media Studies (CoNMedia)*, 2013, pp. 1–4. doi: 10.1109/CoNMedia.2013.6708545.
- [175] S. Jha, *Engineering Internet QoS*, 1st ed. 2000.
- [176] S. Özokes, "EVALUATION OF ACTIVE QUEUE MANAGEMENT ALGORITHMS," 2005.
- [177] L. Liu, M. Lian, C. Lu, S. Zhang, R. Liu, and N. Xiong, "TCSA: A Traffic Congestion Situation Assessment Scheme Based on Multi-Index Fuzzy Comprehensive Evaluation in 5G-IoV," *Electronics (Basel)*, vol. 11, no. 7, p. 1032, Mar. 2022, doi: 10.3390/electronics11071032.

- [178] L. Leskelä, “Stabilization of an Overloaded Queuing Network Using Measurement-Based Admission Control,” *J Appl Probab*, vol. 43, no. 1, pp. 231–244, 2006, doi: DOI: 10.1239/jap/1143936256.
- [179] R. S. Sutton, D. Precup, and S. Singh, “Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning,” *Artif. Intell.*, vol. 112, no. 1–2, pp. 181–211, Aug. 1999, doi: 10.1016/S0004-3702(99)00052-1.
- [180] H. P. Phyu, D. Naboulsi, and R. Stanica, “Machine Learning in Network Slicing—A Survey,” *IEEE Access*, vol. 11, pp. 39123–39153, 2023, doi: 10.1109/ACCESS.2023.3267985.
- [181] C. F. Hayes *et al.*, “A practical guide to multi-objective reinforcement learning and planning,” *Auton Agent Multi Agent Syst*, vol. 36, no. 1, p. 26, Apr. 2022, doi: 10.1007/s10458-022-09552-y.
- [182] H. J. Damsgaard, A. Ometov, and J. Nurmi, “Approximation Opportunities in Edge Computing Hardware: A Systematic Literature Review,” *ACM Comput Surv*, vol. 55, no. 12, pp. 1–49, Dec. 2023, doi: 10.1145/3572772.
- [183] European Commission, “Approximate Computing for Power and Energy Optimisation.” 2020. Accessed: Jun. 22, 2023. [Online]. Available: <https://cordis.europa.eu/project/id/956090/pl>
- [184] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley and Sons Inc. Publication, 2005. doi: 10.1002/9780470316887.
- [185] E. Nordstrom and J. Carlstrom, “Call admission control and routing for integrated CBR/VBR and ABR services: a Markov decision approach,” in *IEEE ATM Workshop '99 Proceedings (Cat. No. 99TH8462)*, IEICE of Japan, pp. 71–76. doi: 10.1109/ATM.1999.786781.
- [186] N. Nasser and H. Hassanein, “An optimal and fair call admission control policy for seamless handoff in multimedia wireless networks with QoS guarantees,” in *IEEE Global Telecommunications Conference, 2004. GLOBECOM '04.*, IEEE, pp. 3926–3930. doi: 10.1109/GLOCOM.2004.1379104.
- [187] X. Yang and G. Feng, “Cost Minimization for Admission Control in Bandwidth Asymmetry Wireless Networks,” in *2007 IEEE International Conference on Communications*, IEEE, Jun. 2007, pp. 5484–5489. doi: 10.1109/ICC.2007.909.
- [188] W. Ni, W. Li, and M. Alam, “Determination of optimal call admission control policy in wireless networks,” *IEEE Trans Wirel Commun*, vol. 8, no. 2, pp. 1038–1044, Feb. 2009, doi: 10.1109/TWC.2009.080349.
- [189] Huan Chen, Chih-Chuan Cheng, and Hsi-Hsun Yeh, “Guard-Channel-Based Incremental and Dynamic Optimization on Call Admission Control for Next-Generation QoS-Aware Heterogeneous Systems,” *IEEE Trans Veh Technol*, vol. 57, no. 5, pp. 3064–3082, Sep. 2008, doi: 10.1109/TVT.2008.915521.
- [190] Y. Fang, I. Chlamtac, and Yi-Bing Lin, “Channel occupancy times and handoff rate for mobile computing and PCS networks,” *IEEE Transactions on Computers*, vol. 47, no. 6, pp. 679–692, Jun. 1998, doi: 10.1109/12.689647.
- [191] W. Chen, J. Yu, and F. Pan, “An Optimal CAC Scheme Based on GSMDP Model in Heterogeneous Wireless Networks,” in *2010 2nd International Workshop on Intelligent Systems and Applications*, IEEE, May 2010, pp. 1–6. doi: 10.1109/IWISA.2010.5473251.
- [192] A. Pietrabissa and F. Delli Priscoli, “MDP call control in variable capacity

- communication networks,” in *18th Mediterranean Conference on Control and Automation, MED'10*, IEEE, Jun. 2010, pp. 483–488. doi: 10.1109/MED.2010.5547715.
- [193] C. Chang, Chung-Ju Chang, and Kuen-Rong Lo, “Analysis of a hierarchical cellular system with reneging and dropping for waiting new and handoff calls,” *IEEE Trans Veh Technol*, vol. 48, no. 4, pp. 1080–1091, Jul. 1999, doi: 10.1109/25.775357.
- [194] G. H. S. Carvalho, R. W. L. Coutinho, and J. C. W. A. Costa, “Design of optimal Call Admission Control for WiMax/WiFi integration,” in *2009 SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC)*, IEEE, Nov. 2009, pp. 564–568. doi: 10.1109/IMOC.2009.5427519.
- [195] S.-M. Senouci, A.-L. Beylot, and G. Pujolle, “Call admission control in cellular networks: A reinforcement learning solution,” *International Journal of Network Management*, vol. 14, no. 2, pp. 89–103, Mar. 2004, doi: 10.1002/nem.510.
- [196] T. Mitchell, *Machine Learning*. McGraw-Hill, 1997.
- [197] Kwiatkowska 2007, “Uzupelnić”.
- [198] M. N. Qureshi, M. K. Shahid, M. I. Tiwana, M. Haddad, I. Ahmed, and T. Faisal, “Neural Networks for Energy-Efficient Self Optimization of eNodeB Antenna Tilt in 5G Mobile Network Environments,” *IEEE Access*, vol. 10, pp. 61678–61694, 2022, doi: 10.1109/ACCESS.2022.3181595.
- [199] Keysight, “Network Digital Twin Ecosystem,” Apr. 2022. <https://www.keysight.com/us/en/assets/3122-1404/technical-overviews/Network-Digital-Twin-Ecosystem.pdf> (accessed Jun. 24, 2023).
- [200] IEC Market Strategy Board, “Edge intelligence,” 2017.
- [201] “European edge computing consortium.” <https://ecconsortium.eu/> (accessed Jun. 24, 2023).
- [202] D. Artuñedo Guillen *et al.*, “Edge computing for 5G networks - white paper.” Zenodo, Mar. 2020. doi: 10.5281/zenodo.3698117.
- [203] A. Díaz Zayas *et al.*, “A Modular Experimentation Methodology for 5G Deployments: The 5GENESIS Approach,” *Sensors*, vol. 20, no. 22, 2020, doi: 10.3390/s20226652.
- [204] D. M. Gutierrez-Estevez, N. Dipietro, A. Dedomenico, M. Gramaglia, U. Elzur, and Y. Wang, “5G-MoNArch Use Case for ETSI ENI: Elastic Resource Management and Orchestration,” in *2018 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2018, pp. 1–5. doi: 10.1109/CSCN.2018.8581789.
- [205] R. Ferrus, O. Sallent, J. Perez-Romero, and R. Agusti, “On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration,” *IEEE Communications Magazine*, vol. 56, no. 5, pp. 184–192, May 2018, doi: 10.1109/MCOM.2017.1700268.
- [206] A. Slalmi, H. Chaibi, R. Saadane, A. Chehri, and G. Jeon, “5G NB-IoT: Efficient network call admission control in cellular networks,” *Concurr Comput*, vol. 33, no. 22, p. e6047, Nov. 2021, doi: <https://doi.org/10.1002/cpe.6047>.
- [207] X. Costa-Pérez, J. Swetina, T. Guo, rajesh mahindra, and S. Rangarajan, “Radio Access Network Virtualization for Future Mobile Carrier Networks,” *IEEE Communications Magazine*, vol. 51, p. 27, Jul. 2013, doi: 10.1109/MCOM.2013.6553675.
- [208] A. Ahmed and E. Ahmed, *A Survey on Mobile Edge Computing*. 2016. doi:

- 10.1109/ISCO.2016.7727082.
- [209] S. Hu, W. Shi, and G. Li, “CEC: A Containerized Edge Computing Framework for Dynamic Resource Provisioning,” *IEEE Trans Mob Comput*, vol. 22, no. 7, pp. 3840–3854, 2023, doi: 10.1109/TMC.2022.3147800.
- [210] T. Subramanya, D. Harutyunyan, and R. Riggio, “Machine learning-driven service function chain placement and scaling in MEC-enabled 5G networks,” *Computer Networks*, vol. 166, p. 106980, Jan. 2020, doi: 10.1016/j.comnet.2019.106980.
- [211] X. Foukas and B. Radunovic, “Concordia: Teaching the 5G VRAN to Share Compute,” in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, in SIGCOMM '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 580–596. doi: 10.1145/3452296.3472894.
- [212] D. G. da Silva, M. T. B. Geller, M. S. dos S. Moura, and A. A. de M. Meneses, “Performance evaluation of LSTM neural networks for consumption prediction,” *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 2, p. 100030, 2022, doi: 10.1016/j.prime.2022.100030.
- [213] J. Zhong, S. Duan, and Q. Li, “Auto-Scaling Cloud Resources using LSTM and Reinforcement Learning to Guarantee Service-Level Agreements and Reduce Resource Costs,” *J Phys Conf Ser*, vol. 1237, no. 2, p. 022033, Jun. 2019, doi: 10.1088/1742-6596/1237/2/022033.
- [214] Subhash Chopra, “Intelligent 5G L2 MAC Scheduler,” 2021.
- [215] Y. Garí, D. A. Monge, E. Pacini, C. Mateos, and C. García Garino, “Reinforcement learning-based application Autoscaling in the Cloud: A survey,” *Eng Appl Artif Intell*, vol. 102, p. 104288, Jun. 2021, doi: 10.1016/j.engappai.2021.104288.
- [216] Y. Lin, Y. Gao, and W. Dong, “Bandwidth Prediction for 5G Cellular Networks,” in *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*, 2022, pp. 1–10. doi: 10.1109/IWQoS54832.2022.9812912.
- [217] K. Wang, S. Shah, H. Chen, A. Perrault, F. Doshi-Velez, and M. Tambe, “Learning MDPs from Features: Predict-Then-Optimize for Sequential Decision Problems by Reinforcement Learning.”
- [218] S. Balhara *et al.*, “A survey on deep reinforcement learning architectures, applications and emerging trends,” *IET Communications*, Jul. 2022, doi: 10.1049/cmu2.12447.
- [219] S. Taherizadeh and V. Stankovski, “Dynamic Multi-level Auto-scaling Rules for Containerized Applications,” *Comput J*, vol. 62, no. 2, pp. 174–197, Feb. 2019, doi: 10.1093/comjnl/bxy043.
- [220] I. Vinogradova, “Multi-Attribute Decision-Making Methods as a Part of Mathematical Optimization,” *Mathematics*, vol. 7, no. 10, 2019, doi: 10.3390/math7100915.
- [221] S. Song, J. Jung, M. Choi, C. Lee, J. Sun, and J.-M. Chung, “Multipath Based Adaptive Concurrent Transfer for Real-Time Video Streaming Over 5G Multi-RAT Systems,” *IEEE Access*, vol. 7, pp. 146470–146479, 2019, doi: 10.1109/ACCESS.2019.2945357.
- [222] V. F. Monteiro, M. Ericson, and F. R. P. Cavalcanti, “Fast-RAT Scheduling in a 5G Multi-RAT Scenario,” *IEEE Communications Magazine*, vol. 55, no. 6, pp. 79–85, 2017, doi: 10.1109/MCOM.2017.1601094.
- [223] N. A. Elmosilhy, M. M. Elmesalawy, and A. M. A. Elhaleem, “User Association

- With Mode Selection in LWA-Based Multi-RAT HetNet,” *IEEE Access*, vol. 7, pp. 158623–158633, 2019, doi: 10.1109/ACCESS.2019.2949035.
- [224] M. S. Afaqui, C. Cano, V. Kotzsch, C. Felber, and W. Nitzold, “Implementation of the 3GPP LTE-WLAN Inter-Working Protocols in Ns-3,” in *Proceedings of the 2019 Workshop on Ns-3*, in WNS3 '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 25–32. doi: 10.1145/3321349.3321355.
- [225] M. S. Afaqui, C. Cano, V. Kotzsch, C. Felber, and W. Nitzold, “Real-time operation of LTE/Wi-Fi interworking via NS-3 and SDR interfacing,” in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 981–982. doi: 10.1109/INFCOMW.2019.8845047.
- [226] F. Elsharif and E. K. P. Chong, “Risk-Averse Traffic Allocation for Multi-RAT Connectivity in HetNets,” in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, 2020, pp. 803–811. doi: 10.1109/CCWC47524.2020.9031114.
- [227] Y. He, Z. Zhang, and Y. Zhang, “A Big Data Deep Reinforcement Learning Approach to Next Generation Green Wireless Networks,” in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6. doi: 10.1109/GLOCOM.2017.8254717.
- [228] M. Anany, M. M. Elmesalawy, and A. M. A. El-Haleem, “Matching Game-Based Cell Association in Multi-RAT HetNet Considering Device Requirements,” *IEEE Internet Things J*, vol. 6, no. 6, pp. 9774–9782, 2019, doi: 10.1109/JIOT.2019.2931448.
- [229] M. Katz and F. Fitzek, *WiMAX Evolution: Emerging Technologies and Applications*. 2008. doi: 10.1002/9780470740118.
- [230] A. Raja and C. Flanagan, “Real-Time, Non-intrusive Speech Quality Estimation: A Signal-Based Model,” in *Genetic Programming*, M. O’Neill, L. Vanneschi, S. Gustafson, A. I. Esparcia Alcázar, I. De Falco, A. Della Cioppa, and E. Tarantino, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 37–48.
- [231] J. Yun, M. J. Piran, and D. Y. Suh, “QoE-Driven Resource Allocation for Live Video Streaming Over D2D-Underlaid 5G Cellular Networks,” *IEEE Access*, vol. 6, pp. 72563–72580, 2018, doi: 10.1109/ACCESS.2018.2882441.
- [232] F. Metzger *et al.*, “Context Monitoring for Improved System Performance and QoE,” in *Autonomous Control for a Reliable Internet of Services: Methods, Models, Approaches, Techniques, Algorithms, and Tools*, I. Ganchev, R. D. van der Mei, and H. van den Berg, Eds., Cham: Springer International Publishing, 2018, pp. 23–48. doi: 10.1007/978-3-319-90415-3_2.
- [233] J. Nawała, L. Janowski, and M. Leszczuk, “Modeling of Quality of Experience in No-Reference Model,” 2017.
- [234] 5G-Blueprint Consortium, “5G-Blueprint.” Sep. 01, 2020. Accessed: Jun. 21, 2023. [Online]. Available: <https://www.5gblueprint.eu/>
- [235] Mushroom Networks, “Mushroom networks home page.” <https://www.mushroomnetworks.com> (accessed Jun. 22, 2023).
- [236] Servision, “Servision home page.” <http://www.servision.net/solutions/mobile> (accessed Jun. 22, 2023).
- [237] Candelatech, “Candelatech.” http://www.candelatech.com/wiser50_product.php (accessed Jun. 22, 2023).
- [238] “Ixia home page.”

- [239] B. Rickenbach, P. Griffin, J. Rush, J. Flanagan, B. Adamson, and J. Macker, "Adaptive data delivery over disadvantaged, dynamic networks," in *2011 - MILCOM 2011 Military Communications Conference*, 2011, pp. 1628–1632. doi: 10.1109/MILCOM.2011.6127542.
- [240] H. Luo, S. Ci, D. Wu, and H. Tang, "Adaptive Wireless Multimedia Communications with Context-Awareness Using Ontology-Based Models," in *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, 2010, pp. 1–5. doi: 10.1109/GLOCOM.2010.5683106.
- [241] I. Kofler, "In-Network Adaptation of Scalable Video Content," *SIGMultimedia Rec.*, vol. 2, no. 4, pp. 7–8, Dec. 2010, doi: 10.1145/2039331.2039335.
- [242] A. Flizikowski, M. Majewski, and M. Przybyszewski, "QoE assessment of VoIP over IEEE 802.16 networks with DaVinci codes using E-model," in *2010 Future Network & Mobile Summit*, 2010, pp. 1–8.
- [243] E. Laias, *Performance analysis and enhancement of QoS for fixed WiMAX networks*. LAP LAMBERT Academic Publishing, 2009.
- [244] A. Diefenbach, T. Ginzler, S. McLaughlin, J. Sliwa, T. A. Lampe, and C. Prasse, "TACTICS TSI architecture: A European reference architecture for tactical SOA," in *2016 International Conference on Military Communications and Information Systems (ICMCIS)*, IEEE, May 2016, pp. 1–8. doi: 10.1109/ICMCIS.2016.7496584.
- [245] G. Vasileios *et al.*, "Interoperability of security and quality of Service Policies Over Tactical SOA," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2016, pp. 1–7. doi: 10.1109/SSCI.2016.7850077.
- [246] R. Rigolin F. Lopes, A. Viidanoja, M. Lhotellier, A. Diefenbach, N. Jansen, and T. Ginzler, *A Queuing Mechanism for Delivering QoS-constrained Web Services in Tactical Networks*. 2018. doi: 10.1109/ICMCIS.2018.8398695.
- [247] P. Szymaniak *et al.*, "Nowe rozwiązania w zakresie systemów do strumieniowania obrazu ruchomego w sieciach bezprzewodowych," *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne*, ISSN 1230-3496, e-ISSN 2449-7487, vol. 88, no. 6, pp. 576–579, 2016.
- [248] T-Nova Project Consortium, "Service mapping," 2016.
- [249] G. Bianchi *et al.*, "Superfluidity: a flexible functional architecture for 5G networks," *Transactions on Emerging Telecommunications Technologies*, vol. 27, no. 9, pp. 1178–1186, Sep. 2016, doi: 10.1002/ett.3082.
- [250] X. Dutreilh, A. Moreau, J. Malenfant, N. Rivierre, and I. Truck, "From Data Center Resource Allocation to Control Theory and Back," in *2010 IEEE 3rd International Conference on Cloud Computing*, IEEE, Jul. 2010, pp. 410–417. doi: 10.1109/CLOUD.2010.55.
- [251] F. Fund, C. Wang, T. Korakis, M. Zink, and S. Panwar, "GENI WiMAX Performance: Evaluation and Comparison of Two Campus Testbeds," in *2013 Second GENI Research and Educational Experiment Workshop*, 2013, pp. 73–80. doi: 10.1109/GREE.2013.23.
- [252] J. Pinola, I. Harjula, A. Flizikowski, M. Safianowska, A. Ahmad, and S. S. Mhatre, "EuWireless RAN Architecture and Slicing Framework for Virtual Testbeds," in *Testbeds and Research Infrastructures for the Development of Networks and Communications*, H. Gao, K. Li, X. Yang, and Y. Yin, Eds., Cham: Springer International Publishing, 2020, pp. 131–149.
- [253] ISI Grece, "6G-BRICKS," *Project website*, Jan. 01, 2023. <https://6g-bricks.eu/>

- (accessed Jun. 19, 2023).
- [254] S. Barmounakis and N. Alonistioti, "A Survey of Existing Testbeds for Programmable Networks," in *Wiley 5G Ref*, 2019, pp. 1–23. doi: <https://doi.org/10.1002/9781119471509.w5GRef115>.
- [255] T. S. Rappaport, *Wireless communications: Principles and practice*. Prentice Hall, 1996.
- [256] B. Li, C. Lin, and S. T. Chanson, "Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks," *Wireless Networks*, vol. 4, no. 4, pp. 279–290, 1998, doi: 10.1023/A:1019116424411.
- [257] J. Wicher, "Problemy identyfikacji systemów technicznych ze szczególnym uwzględnieniem układów mechanicznych," IPPT PAN, 1976.
- [258] Z. Bubnicki, *Teoria i algorytmy sterowania*. Warszawa: Wydawnictwo Naukowe PWN, 2005.
- [259] Stefan Ziemba, *Problemy teorii systemów*. Kraków: Ossolineum, 1980.
- [260] E. W. Knightly and N. B. Shroff, "Admission control for statistical QoS: theory and practice," *IEEE Netw*, vol. 13, no. 2, pp. 20–29, 1999, doi: 10.1109/65.768485.
- [261] Christos Politis, "EUHT evaluation report - ITU evaluation proecess for IMT-2020," 2020.
- [262] M. Fuentes *et al.*, "5G New Radio Evaluation Against IMT-2020 Key Performance Indicators," *IEEE Access*, vol. 8, pp. 110880–110896, 2020, doi: 10.1109/ACCESS.2020.3001641.
- [263] K. Koutlia, B. Bojovic, Z. Ali, and S. Lagén, "Calibration of the 5G-LENA system level simulator in 3GPP reference scenarios," *Simul Model Pract Theory*, vol. 119, p. 102580, Sep. 2022, doi: 10.1016/j.simpat.2022.102580.
- [264] P. K. Gkonis, P. T. Trakadas, and D. I. Kaklamani, "A Comprehensive Study on Simulation Techniques for 5G Networks: State of the Art Results, Analysis, and Future Challenges," *Electronics (Basel)*, vol. 9, no. 3, 2020, doi: 10.3390/electronics9030468.
- [265] H. Kaschel, S. Cordero, and E. Costoya, *Analysis and Evaluation of Radio Mobile program on Line of Sight paths, with SRTM and ASTER DTEDs and its v11.6.6 / v9.1.6 versions*. 2019. doi: 10.1109/CHILECON47746.2019.8988013.
- [266] A. Privalov and A. Tsarev, *Hybrid Model Of Human Mobility For DTN Network Simulation*. 2016. doi: 10.7148/2016-0419.
- [267] S. Kasampalis, P. Lazaridis, Z. Zaharis, A. Bizopoulos, S. Zettas, and J. Cosmas, *Comparison of Longley-Rice, ITM and ITWOM propagation models for DTV and FM broadcasting*. 2013.
- [268] P. Schulz, *Queueing-Theoretic End-to-End Latency Modeling of Future Wireless Networks*. Technische Universität Dresden, 2020.
- [269] T. Kazaz *et al.*, "Orchestration and Reconfiguration Control Architecture ORCA-a 5G Experimental Environment," 2017.
- [270] A. Flizikowski, R. Kozik, H. Gierszal, M. Przybyszewski, and W. Houbowicz, "WiMAX system level simulation platform based on ns-2 and DSP integration," *BROADBANDCOM 2009-Selected Papers on Broadband Communication, Information Technology and Biomedical Applications*, 2009.
- [271] X. Guo, W. Ma, Z. Guo, and Z. Hou, "Dynamic Bandwidth Reservation Admission Control Scheme for the IEEE 802.16e Broadband Wireless Access Systems," in *2007 IEEE Wireless Communications and Networking Conference*,

- IEEE, 2007, pp. 3418–3423. doi: 10.1109/WCNC.2007.628.
- [272] A. Flizikowski, R. Kozik, M. Majewski, and M. Przybyszewski, “Evaluation of guard channel admission control schemes for IEEE 802.16 with integrated nb-LDPC codes,” in *2009 International Conference on Ultra Modern Telecommunications & Workshops*, 2009, pp. 1–8. doi: 10.1109/ICUMT.2009.5345468.
- [273] A. Flizikowski, R. Kozik, M. Majewski, and M. Przybyszewski, “Performance comparison of guard channel admission control schemes for IEEE 802.16 system with various turbo code FEC schemes,” in *2009 IEEE 28th International Performance Computing and Communications Conference*, 2009, pp. 360–365. doi: 10.1109/PCCC.2009.5403853.
- [274] “TS23.501 System architecture for the 5G System (5GS),” *3GPP*. 2017.
- [275] “TR 26.939 Guidelines on the Framework for Live Uplink Streaming (FLUS),” *3GPP*. 2018.
- [276] “P.912 : Subjective video quality assessment methods for recognition tasks,” *ITU*. Mar. 15, 2016.
- [277] “[cur].”
- [278] Flizikowski Adam, “INFSCO-ICT-216203 DaVinci D5.4.1 v1.0,” 2010.
- [279] “TS 38.214 NR; Physical layer procedures for data.” 2017.
- [280] G. Tesauro, N. K. Jong, R. Das, and M. N. Bennani, “A Hybrid Reinforcement Learning Approach to Autonomic Resource Allocation,” in *2006 IEEE International Conference on Autonomic Computing*, IEEE, pp. 65–73. doi: 10.1109/ICAC.2006.1662383.
- [281] “[aina].”
- [282] Christian Horn, “Using Multipath TCP to better survive outages and increase bandwidth,” *Red Hat Blog*, Jun. 09, 2022.
- [283] K. Yamagishi and T. Hayashi, “Parametric Packet-Layer Model for Monitoring Video Quality of IPTV Services,” in *2008 IEEE International Conference on Communications*, IEEE, 2008, pp. 110–114. doi: 10.1109/ICC.2008.29.
- [284] A. Ahmad, A. Floris, and L. Atzori, “Timber: An SDN based emulation platform for QoE Management Experimental Research,” in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, May 2018, pp. 1–6. doi: 10.1109/QoMEX.2018.8463387.
- [285] O. A. Montesinos López, A. Montesinos López, and J. Crossa, “Overfitting, Model Tuning, and Evaluation of Prediction Performance,” in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, O. A. Montesinos López, A. Montesinos López, and J. Crossa, Eds., Cham: Springer International Publishing, 2022, pp. 109–139. doi: 10.1007/978-3-030-89010-0_4.
- [286] D. Lee, D. Lee, M. Choi, and J. Lee, “Prediction of Network Throughput using ARIMA,” in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, 2020, pp. 1–5. doi: 10.1109/ICAIIIC48513.2020.9065083.
- [287] L. Deng, K. Ruan, X. Chen, X. Huang, Y. Zhu, and W. Yu, “An IP Network Traffic Prediction Method based on ARIMA and N-BEATS,” in *2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2022, pp. 336–341. doi: 10.1109/ICPICS55264.2022.9873564.
- [288] A. Flizikowski, E. Alkhovik, M. Munjure Mowla, and M. Arifur Rahman, “Data Handling Mechanisms and Collection Framework for 5G vRAN in Edge

- Networks,” in *2022 IEEE Conference on Standards for Communications and Networking (CSCN)*, IEEE, Nov. 2022, pp. 36–41. doi: 10.1109/CSCN57023.2022.10051118.
- [289] S. Barrachina-Munoz, M. Payaro, and J. Mangues-Bafalluy, “Cloud-native 5G experimental platform with over-the-air transmissions and end-to-end monitoring,” in *2022 13th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, IEEE, Jul. 2022, pp. 692–697. doi: 10.1109/CSNDSP54353.2022.9908028.
- [290] M. Mekki, S. Arora, and A. Ksentini, “A Scalable Monitoring Framework for Network Slicing in 5G and Beyond Mobile Networks,” *IEEE Transactions on Network and Service Management*, vol. 19, no. 1, pp. 413–423, Mar. 2022, doi: 10.1109/TNSM.2021.3119433.
- [291] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” Dec. 2019.
- [292] L.-M. Tufeanu, A. Martian, M.-C. Vochin, C.-L. Paraschiv, and F. Y. Li, “Building an Open Source Containerized 5G SA Network through Docker and Kubernetes,” in *2022 25th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2022, pp. 381–386. doi: 10.1109/WPMC55625.2022.10014753.
- [293] L. N. Smith, “Cyclical Learning Rates for Training Neural Networks,” Jun. 2015.
- [294] M. Rozanska and G. Horn, “Autonomous Multi-Cloud Application Deployment and Optimized Management Using Open Source Frameworks,” in *2020 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, Los Alamitos, CA, USA: IEEE Computer Society, Aug. 2020, pp. 252–253. doi: 10.1109/ACSOS-C51401.2020.00072.
- [295] “IEEE 802.16. Part 16: Air Interface for Broadband Wireless Access Systems,” 2009.
- [296] K. Yu, X. Wang, S. Sun, L. Zhang, and X. Wu, “A Statistical Connection Admission Control Mechanism for Multiservice IEEE 802.16 Network,” in *VTC Spring 2009 - IEEE 69th Vehicular Technology Conference*, IEEE, Apr. 2009, pp. 1–5. doi: 10.1109/VETECS.2009.5073347.
- [297] X.-T. Vu, D.-T. Nguyen, and T. A. Vu, “An finite-state Markov channel model for ACM scheme in WiMAX,” in *TENCON 2009 - 2009 IEEE Region 10 Conference*, IEEE, Nov. 2009, pp. 1–6. doi: 10.1109/TENCON.2009.5396142.
- [298] B. Jang, M. Kim, G. Harerimana, and J. W. Kim, “Q-Learning Algorithms: A Comprehensive Classification and Applications,” *IEEE Access*, vol. 7, pp. 133653–133667, 2019, doi: 10.1109/ACCESS.2019.2941229.
- [299] J. Su, J. Liu, D. B. Thomas, and P. Y. K. Cheung, “Neural Network Based Reinforcement Learning Acceleration on FPGA Platforms,” *SIGARCH Comput. Archit. News*, vol. 44, no. 4, pp. 68–73, Jan. 2017, doi: 10.1145/3039902.3039915.
- [300] J. Barzykowski, A. Domańska, and M. Kujawińska, *Współczesna metrologia. Zagadnienia wybrane*. Wydawnictwo Naukowo-Techniczne, 2007.
- [301] M. Kulin, C. Fortuna, E. De Poorter, D. Deschrijver, and I. Moerman, “Data-Driven Design of Intelligent Wireless Networks: An Overview and Tutorial,” *Sensors*, vol. 16, no. 6, p. 790, Jun. 2016, doi: 10.3390/s16060790.
- [302] V. Mnih *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, doi: 10.1038/nature14236.

12 ANNEX A

12.1 MEASUREMENT TOOLS VALIDATION

The diagram Figure 125 shows the connectivity layout for performing validation of an open-source measurements tools (MGEN, Wireshark) with the commercial IXIA XM2 traffic generator and analyser.

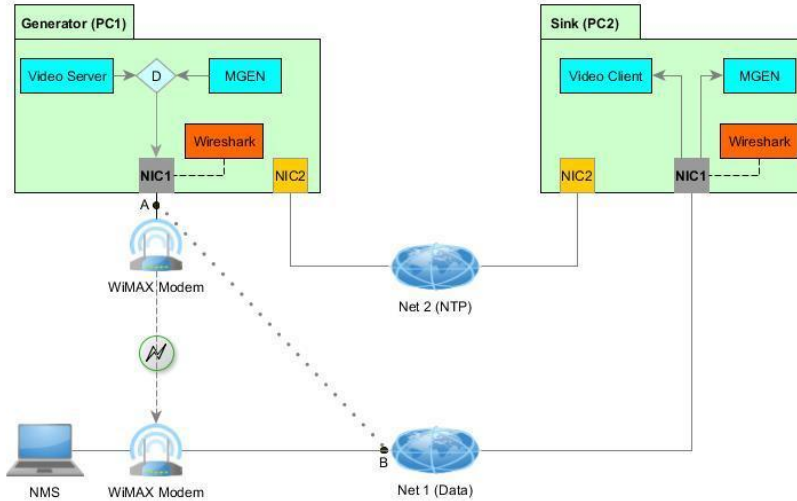


Figure 125 Initial system set-up for validation measurements

Both source and destination hosts (PC1, PC2) were running Wireshark and collecting all the packets transmitted through network interface. Captured packets were processed offline by a dedicated script used to calculate IP packets end-to-end performance statistics. The script uses files created on source and destination hosts, de-encapsulates both data sets there, and from each packet identifies information about its length, timestamp, and some additional statistics for higher-layer protocols such as RTP e.g. sequence number of RTP packet. This information is then used for calculating delay, throughput and jitter.

This section focuses on the preparatory stage before measurement campaign of video streaming services in 4G/5G networks. Author has successfully prepared a complete environment that relies on low-cost IP performance measurement software. It has been shown that the software provides accurate measurements when validated using professional packet generator/analyser IXIA XM2. It has been also shown that it is quite straightforward and cost effective to realise GPS synchronisation using the self-developed platform based on Raspberry Pi (cost of a unit ca 100 EUR) with GPS module attached. Moreover it has been shown that making the environment more portable with virtual machines is not feasible due

to inherent problems with time synchronisation between physical and virtual machines. During this stage of tests a great number of laboratory tests were performed. These results show that current level of delays as well as the throughput is sufficient to enable video transmission in stationary conditions (and good radio coverage). The same behaviour has been confirmed for mobile tests with mobility in the range of 10-30km/h in the range of 1km from base station.

12.2 RADIO TRACES

In order to collect traces the dedicated hardware was designed and developed. It was based on Raspberry-PI computer equipped with GPS module and 4G modem. The GPS signal was used for obtaining geographical coordinates and also for synchronizing internal clock with using NTP protocol. A 4G modem provided detailed information about current state such as signal strength, modulation etc. The Raspberry PI computer was installed among other with the MGEN application which was able to continuously send data in the uplink direction to server. GPS coordinates and modem details were collected every second and added to the packet payload sent by the MGEN application, additionally MGEN added precise timestamp for each packet sent. Such hardware equipped and configured was used to sending data when moving through areas covered by 4G. The server which received data was also synchronized with using NTP protocol and was able to determine the network delay, instantaneous and average rate. Traces gathered from real life measurements are presented in the form of CSV or XLS files. The internal script is able to gather required metrics, based on needs of specific scenarios. Information able to retrieve by our approach include:

- Exact timestamps of packet sent and received
- IP addresses of transmitter and receiver
- Traffic type (UDP, TCP, other)
- Packet sequence number (for purpose of further Loss estimation)
- Packet size in Bytes
- Signal Strength (RSSI) in dBm and percent or dB value of Signal Quality (CINR)
- Modulation changes for both Uplink and Downlink direction
- GPS coordinates of mobile nodes

What is important to note, is a fact that data mentioned above are gathered for each subsequent packet, meaning that measurements with higher bitrates provide more accurate data for emulation purposes. For example, in a 5 min window, a 128 Kb/s traffic may transmit approximately 6000 of packets while 1Mb/s traffic ten times more (nearly 60000 packets), therefore significantly more packets provide more detailed and accurate information.

12.3 DELAY MEASUREMENTS

All measurements have been performed with the CBR traffic in uplink direction, and packet size is equal between tests and set to 1370B. This is selected to mimic the size of RTP video packets. The Table 1 shows comparison of average delays and jitter between measurements with IXIA and MGEN. For each measurement the average results are estimated and then compared to estimated average results of the whole test. The ping tool does not allow measurements of jitter at all, and IXIA seems not to support directly the jitter variation. All measurements have been repeated at least 10 times and each flow lasts 15 seconds.

Table 44 Validation results for MGEN (using commercial generator IXIA)

Packet stream	Pkts/sec	OWD _{avg}	OWD (std. dev.)	IPPV _{avg}
Mgen 1024 bitrate	93	39.23	10.00	7.24
IXIA 1024 bitrate	93	36.48	6.68	-
Mgen 512 bitrate	47	45.63	13.60	11.95
IXIA 512 bitrate	47	42.53	10.94	-
Mgen 256 bitrate	23	54.61	10.21	17.28
IXIA 256 bitrate	23	53.08	17.98	-
Mgen 128 bitrate	12	60.14	4.30	5.85
IXIA 128 bitrate	12	59.84	7.94	-
PING (32B)	-	72.61	5.26	-

It can be observed that the measurements performed using IXIA and MGEN in laboratory conditions are close to each other with ca. 3-8% inaccuracy (given same bitrate of a stream for pairwise comparison). The difference between measurement values can be explained by the 30-60us additional delay per packet that is introduced when performing readings with MGEN. Next step was to assess whether measurements using Wireshark provide satisfactory accuracy as compared to IXIA and MGEN assuming identical network setup. Source and destination packets were correlated by using RTP sequence number, and for such a pair of packet delay and jitter were calculated.

Table 45 Comparison of accuracy between MGEN and Wireshark

	OWD _{avg} [ms] (std.dev.)	IPPV _{avg} [ms]
MGEN	43,52 (14,66)	9,2
Wireshark	43,11 (14,66)	9,1

The delay was obtained by subtracting destination timestamp from source timestamp, which is consistent with an approach defined by the RFC2679 with the exception that timestamps were not included in transmitted data but were added by Wireshark. The IPPV metric was calculated in compliance with the RFC3393 and throughput was calculated by summing up all packet lengths in

defined period of time that is compliant with the RFC6349. Measurements collected in Table 2 have been derived from a stream of 2800 thousands packets. The clear conclusion can be made that it has been positively validated that the derivation of packet statistics using data captured either with MGEN or Wireshark exhibits only differences at the level of microseconds. Such inaccuracy is negligible for the IP packet measurements needs.

12.4 TIME SYNCHRONISATION OF MOBILE TERMINALS

Accurate end-to-end measurements in an outdoor environment require a precise time-synchronisation solution. As it was presented in the literature review chapter, there are issues in synchronising end terminals clocks, over wireless channel. To cope with this issue we focused on establishing a reference time source on each end of the connection. Therefore as a final result of author's investigation, the Raspberry Pi (RP) equipped with Adafruit Ultimate GPS MTK3339 receiver was chosen as equipment for synchronising clocks of both measurement end-terminals.

Table 46 Delay measurements for validating time synchronization

Average Delay [ms] each end-point connected to its local time server (on Raspberry Pi)	Average Delay [ms] one of the end points acting as NTP server
43,50	45,63

According to the gathered insights and logs observed GPS synchronisation seems to be a lot more accurate than NTP daemon solution (10 fold improvement in accuracy). Our test proved that such GPS solution is more trustworthy, also important to note is fact, that it will be a lot easier to adopt it in case of mobile tests, where usage of NTP daemon solution will be unavailable or may have suffer significant negative impact of weak, or dynamically changing wireless internet connection.

13 ANNEX B

13.1 TEMPORAL ACTIVITY, SPATIAL ACTIVITY METRICS

It has to be highlighted here that at time of evaluating the proposed solution, the combination of transcoder and the QoE metrics haven't so far been tested together in the realm of QoE based online feedback. So there was no past evidence of utilizing the two (and its efficiency) for this goal which we could base. Some valuable QoE metrics have been identified that support security scenario. For details on how the tests were configured and executed please refer to the Annex D. For additional information about available QoE statistics and its definitions, please refer to Chapter 2. From the prior art study there – the two metrics can be utilized as initial parameters that may help determine proper control actions for the video controller given the target of identifying dynamics of the scene: Temporal Activity (TA) and Spatial Activity (SA). Temporal Activity provides the information about the motion activity in an image. Spatial Activity on the other hand determines the amount of image detail. By analysing both metrics simultaneously at a constant rate, it is possible to monitor what is in the focus of the camera and thus constantly evaluate whether there is a movement in the video feed (which might be the reason to activate other QoE metrics) or not. If the level of motion captured and the need for details in the video stream increase, values of those metrics increase as well. By monitoring the visual activity on the video stream, it is possible to adjust other QoE metrics in order to maintain the highest quality possible of a reference video for a given situation e.g. specific requirements. Values of the most relevant QoE metrics (use-case based) should be mapped to different transcoder optimization parameters e.g. bit rate, FPS, resolution.

13.2 EVALUATION OF QOE METRICS FOR SURVEILLANCE REAL-TIME VIDEO ADAPTATION

From the variety of statistics gathered for delivering surveillance video over challenging wireless channel, three of the most important include: *Blockloss*, *Freezing* and *Blockiness*. Figure 126 shows some of the measured statistics from video tests that were carried out for different bit rates and FPS values and compared to the original video with the most important statistics highlighted. These metrics serve as base reference for performing the tests. Within these set of tests, two following types of video were tested:

- Low quality video with low resolution and destructive compression (TV news) - bitrate: 2.5-3 Mbit/s
- High quality HD video with high resolution and efficient compression algorithms (movie trailer) - bitrate: 0.5-1 Mbit/s

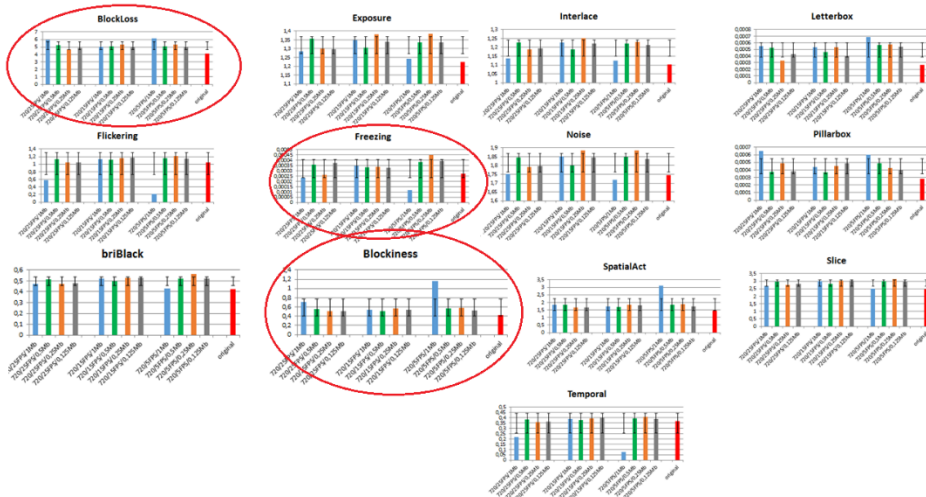


Figure 126 Some of the measured statistics from video tests

For low-quality videos there were no essential differences in results for different transcoder settings as there were no noticeable differences between the best and worst cases when QoE is considered. For high-quality videos, the differences in results were more significant, where difference between best and worst cases fluctuated between 5 to 20 percent in particular metric readings.

During the tests videos were collected for further visual analysis by the tester, where quality of videos based on human perception was correlated to the statistics calculated via QoE Probe [233]. It is worth mentioning, that quality of low-bandwidth video was poor, as there were visible blocks or video errors introduced by encoding process. For high-bandwidth videos however, output quality was good and comparable to the initial video. Therefore to achieve the best results, transcoder bitrate should be set at least to the bitrate of the input video.

14 ANNEX C

14.1 NETWORK EMULATION TOOL

Network emulation tool uses real values of modulation, bitrate and delay from the trace file, it allows defining all crucial parameters described in chapter 4 such as OFDM symbols count, frame duration and various overheads. The application provides a possibility of defining users' activity in certain time, for each user it is possible to define the type of traffic (rtPS, best effort), negotiated minimum and maximum bandwidth, and priority as well as particular activity in certain time i.e. whether user is active or not, the requested bit rate, and loss factor. Basing on the data read from input file and data configured directly on the TBONEX GUI, the application calculates rate, delay and loss for users that would like to send data in given network environment. As the output the application produces two scripts that have to be run on the Linux host. Additionally it also visualizes input and output data on embedded chats. The logical concept of the TBONEX application is presented in the figure Figure 127.

The TC output script contains a set of commands for traffic control. It creates hierarchy of traffic control nodes such as qdisc, class and filters, and attaches the hierarchy structure to the outgoing network interface. For each user it creates the following set of traffic control nodes: HTB (Hierarchical Token Bucket) class that is used to limit bandwidth rate; NETEM qdisc for defining delay and loss for the user's flow; and filter which redirects user's flow to particular class and qdisc basing on the source IP port of transmission. Additionally, for isolating any other traffic from the simulated transmission the script defines qdisc and filter for all remaining traffic flowing through the interface and passes ad it is them without any disturbing. The script manipulates rate, delay and loss for each user by changing parameters of particular htb class or netem qdisc, thus it simulates situation in real world where those parameters change depending on 4G/5G network condition. The other script that is produced by the application is the input script for mgen application. Mgen (Multi-Generator) is open source software used to generate network traffic according to patterns defined in input file.

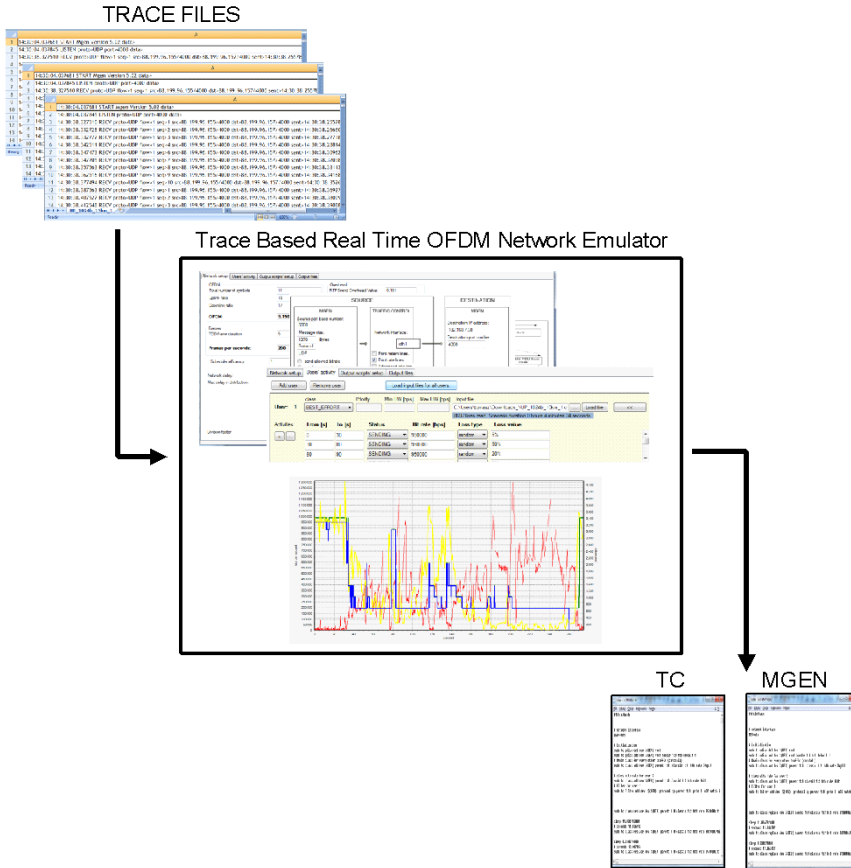


Figure 127 TBONEX logical view

The generated script provides a configuration file for the MGEN, it defines the traffic rate in certain period of times, and it can support two scenarios:

- Scenario#1: Mgen sends data according to the requested video rate.
- Scenario#2: Mgen sends data based on instantaneously emulated channel bitrate (policing mode).

The first scenario (Scenario#1) is simpler and mgen generates traffic with the rate defined as requested rate regardless of the network condition. The main burden of controlling rate takes the first script. The second approach (Scenario#2) simulates a situation when the video stream gets adapted to network condition. In this scenario mgen produces traffic equal to the allowed rate - it means that the same rate that is set by the traffic control is used by the mgen. The second mode entails a need of exact synchronization of traffic control and mgen scripts. The source data rate limited by traffic control and the bitrate produced by mgen should be tightly controlled and changed in exact the same time. The problem was not completely solved yet, despite of the same high microsecond accuracy between the scripts, the traffic control script was executed for a longer time than the mgen script. It happened due to variability of executing traffic control commands. The main concept of the bandwidth limitation

performed by scripts produced by the TBONEX application is presented on the picture below (Figure 128).

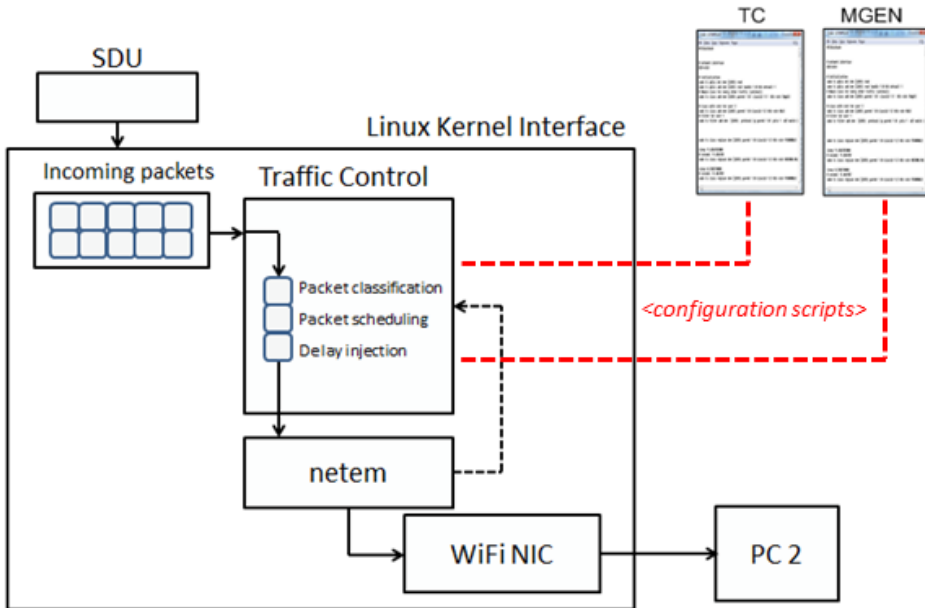


Figure 128 TBONEX – approach to enforcing channel traces at the OS level

In order to use real video stream sent from external device instead of traffic generated locally by MGEN and still have the possibility of using traffic control on outgoing interface there is another dedicated application developed. It receives stream from external host or device and sends them to destination from local port. It is simple application that redirects a stream and allows traffic control to manipulate them on outgoing interface. Additionally it is possible to equip this application with simple shaping mechanism (see Figure 57, Figure 58).

15 ANNEX D

QoE related measurements supporting “security scenario” development are prepared below. In order to perform preliminary tests of the video quality for security scenario the “QoE monitoring” tool was installed in the local testbed.

15.1.1 Software

Regarding the whole testbed, two separate PCs were used: first hosting the transcoder and the second for receiver. Laptops were connected via LAN cable through dedicated router. Following software components were installed on both machines

- Ubuntu 14.04 LTS x64
- ffmpeg codec pack: required to capture, encode and stream video data
- wireshark: used to analyze network traffic as well as to create statistics and graphs.
- as a transcoder, i2cat Media Streamer available at <https://github.com/ua-i2cat/liveMediaStreamer> was used.
- on the receiver side, a dedicated Python script was written to capture video data, pass it to mitsu application in order to get statistics, and finally generate a report.

15.1.2 Performance

In the preliminary tests stage, we considered alternative testbed configurations as well:

- two virtual machines on the same box: one for the transcoder, another for the receiver
- transcoder installed on the virtual machine (guest machine). Receiver installed on the master host.

These two solutions were easier to set-up, but during testing phase we found that offered performance was not acceptable on our testing platform (Intel i3, dual core machine):

- CPU usage was between 200% and 300%
- FPS on the receiver part was below 2 FPS, which was not acceptable
- in the transcoder logs, we could find following error:

```
livemediastreamer output: ERROR [Utils.cpp:478] - Frame
discarded by AVFramedQueue
livemediastreamer output: WARN [Utils.cpp:450] - Your
computer is too slow
```

Switching to two PCs mode solved all the performance problems.

15.1.3 Configuration

The whole video processing chain involves following operations:

- on PC1

- capture video stream from the camera or video file, encode it to desired format and start streaming it using RTP. This task was managed via ffmpeg.
- collect all available RTP streams and transform them into single RTSP stream. This was managed by the transcoder deployed on local Tomcat server.
- on PC2:
 - a dedicated Python script was written in order to perform following steps:
 - invoke cvlc (command-line VLC) in order to capture RTSP data and save it into 20-seconds long MPEG-4 files.
 - For each session (fixed combination of video parameters), 10 video files were recorded in order to average the results.
 - invoke mitsu application in order to process consecutive video files and generate statistics for them.
 - Collect statistics from all processed files and combine them into single CSV file.

For reasons of compatibility CSV file was chosen to be imported into any data analysis tool like MS Excel.

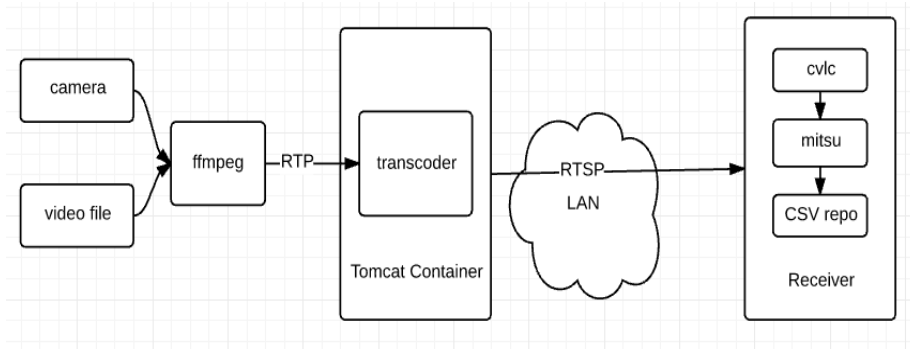


Figure 129

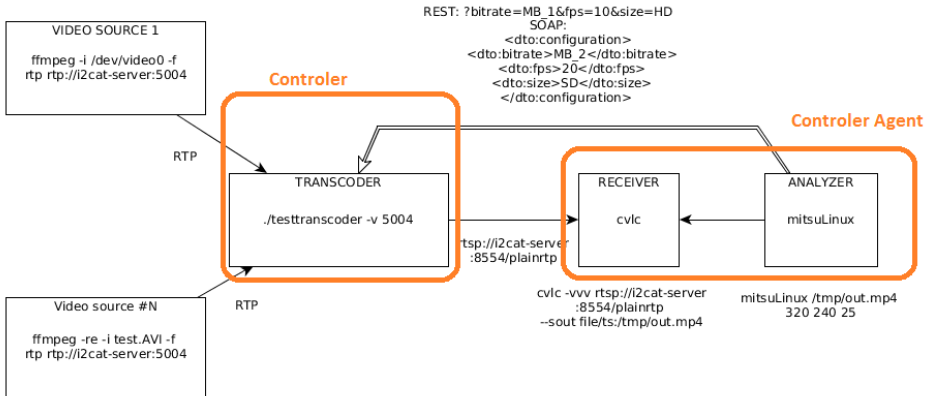


Figure 130 SW configuration

15.1.4 Procedure

Testing procedure used in order to:

- multiple configurations were tested in the testbed. We tried to determine how changing single video parameter can affect the output results. Following video parameters were tested:
 - FPS = { 1, 5, 10, 20, 30 }
 - bitrate = { 0.125, 0.25, 0.5, 1, 2, 4 Mbit/s }
 - Resolution = { 320x240, 640x480, 1366x768 }
- For each combination of tested parameters, 10 short videos were captured on the receiver side.
- Statistics were calculated for each file using mitsu application.
- Once all the videos were processed, all the statistics were dump into a CSV file.

Following parameters were calculated by the mitsu application and stored into the output CSV file: Blackout, BlockLoss, Blockiness, Contrast, Exposure, Flickering, Freezing, Interlace, Letterbox, Noise, Pillarbox, Slice, SpatialAct, Temporal. Sample results for Blockiness are available in table below (Table 47).

Table 47 Blockiness – sample results

Blockiness	Average	Median	Standard deviation
1280_720_15_0125	1,4750	1,4058	0,2288
1280_720_15_025	1,4148	1,3559	0,2004
1280_720_15_05	1,4039	1,3040	0,1929

1280_720_15_1	1,5560	1,5741	0,2391
1280_720_25_0125	1,5460	1,4882	0,2733
1280_720_25_025	1,4098	1,4171	0,0985
1280_720_25_05	1,4907	1,4644	0,1922
1280_720_25_1	1,5685	1,5428	0,1835
1280_720_5_0125	1,5350	1,4577	0,2223
1280_720_5_025	1,5866	1,5690	0,1994
1280_720_5_05	1,6097	1,6497	0,1460
1280_720_5_1	1,4381	1,4076	0,1798
input video	1,1125	1,1116	0,0331

Format of the first column is as follows: WIDTH_HEIGHT_FPS_BITRATE. Statistical functions: Average, Median and Standard deviation were calculated for sequence of all output videos processed with given set of input parameters. Figure below shows some of the measured statistics from video tests that were carried out for different bit rates and FPS values and compared to the original video with the most important statistics highlighted. Within these set of tests, two following types of video were tested:

- Low quality video with low resolution and destructive compression
 - raw DV file: <http://samples.mplayerhq.hu/DV-raw/voxnews.dv> (TV news)
 - Bitrate: 2.5-3 Mbit/s
- High quality HD video with high resolution and efficient compression algorithms.
 - HD video: <https://vimeo.com/120987382> (movie trailer)
 - Bitrate: 0.5-1 Mbit/s

16 ANNEX E

The diagram showing how the simulation environment architecture was prepared in order to deal with parallelization of simulations.

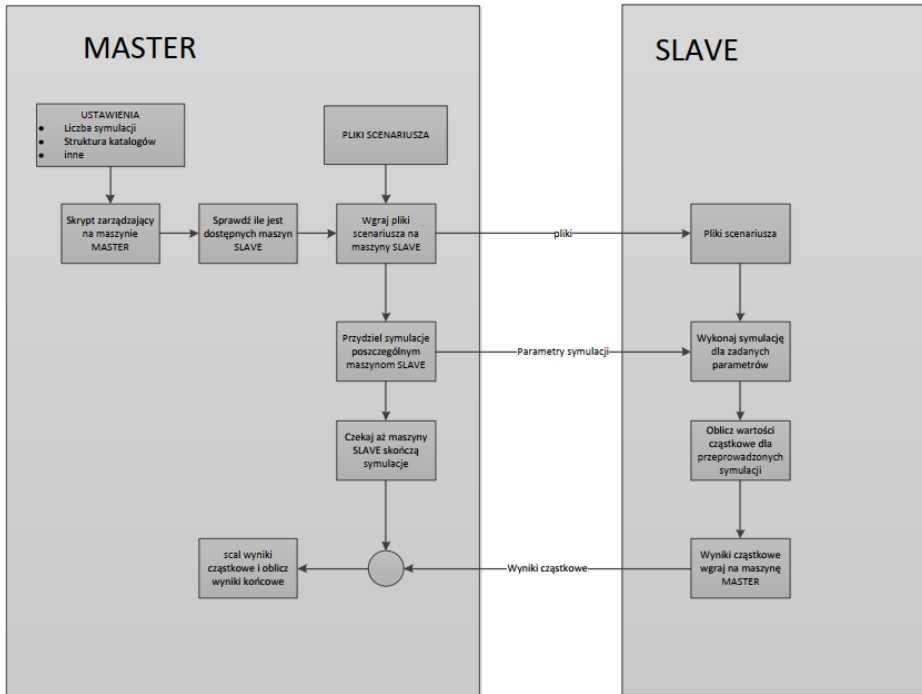


Figure 131 Architecture of the simulation parallelization environment

17 ANNEX F

Test series1 overview. The two different tests were performed:

- **Test A** – 30 minutes tests with the following activities:
 - “No emulation” - during the first 5 minutes of the test there is no emulation, computers are connected directly via Ethernet cable.
 - “Emulation enabled but scenario turned off” - during the next 5 minutes emulator was activated but emulates no scenario. We wanted to see if the emulator itself introduces any disruptions to the video traffic.
 - “Emulation enabled with ramp-up #1” - in the next 4 phases of the test we introduce the cycle of ramp-up and successive ramp-down phases, where the delay is first increased in cycle of 50ms-100ms-200ms-500ms and then decreased in reverse order back to the value of 50ms in the same manner. Each single step lasts for 60 seconds.
 - “Emulation enabled with ramp-up #2” - two series of ramp-up and ramp-down cycles are performed (i.e. the speed of changing steps is 2-times higher than in “ramp-up#1”).
 - “Emulation enabled with ramp-up #3” - four series of ramp-up and ramp-down cycles are performed (i.e. the speed of changing steps is 4-times higher than in “ramp-up#1”).
 - “Emulation enabled with ramp-up #4” - eight series of ramp-up and ramp-down cycles are performed (i.e. the speed of changing steps is 4-times higher than in “ramp-up#1”).

Figures (Figure 132 and Figure 133) present results of above mentioned tests of the Blockiness and Blockloss with the trace divided into “test cases”.

- **Test B** – 20 minute test where each 5 minutes we changed the emulation scenario in the following sequence:
 - Test scenario sequence: UTP Parking > Choszczno Stationary > Choszczno Mobile > UTP Walk
 - *UTP Parking* - it is the low mobility scenario, where 4G modem is deployed inside the car, and the car makes circles on the path which coincides with 100 meter radius circle, nearby to the BS antenna - in the direct vicinity of the BS (around 100 meters from it). There is a direct radio line of sight (LOS).
 - *Choszczno Stationary* - it is the stationary scenario where 4G modem is deployed in stable conditions and doesn't move. Radio channel quality is very good as the modem is in direct exposition of the BS signal (LOS).
 - *Choszczno mobile* - it is average mobility scenario (20-30Km/h) in the suburban district. The 4G modem is deployed in the car, with antennas attached to the roof. The car moves on the square

of streets which are located ca. 2 kilometres from the BS. While driving through the streets the radio modem experiences LOS-NLOS conditions interchangeably.

- *UTP Walk* - here the 4G modem was carried by a person who made a walk around the buildings of the university. At certain locations the modem was almost totally losing the LOS visibility of the BS. The walk was rather slow and in the direct vicinity of the BS.



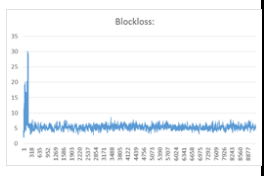
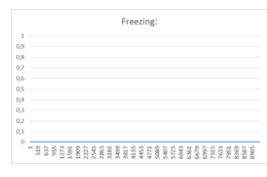
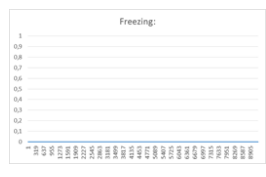
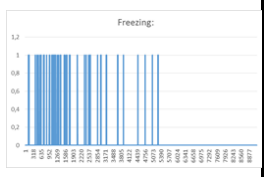

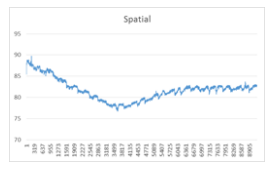
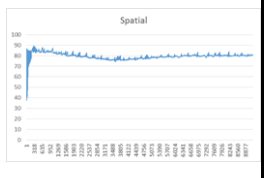
18 ANNEX G

For the test3 we have found that the higher the target rate the more CPU usage of the FFMpeg, which consumed vast majority of the CPU - when measured with the “top” command inside Linux OS. The command used when streaming local file at to the Streaming Streamer was:

```
ffmpeg -re -i 1.mp4 -map 0:0 -vcodec libx264 -bf 0 -f rtp
rtp://localhost:5004
```

When we streamed video via FFMpeg the frame rate presented by FFMpeg started at maximum level and then slowly ramped-down and stabilized at around 15fps. We have defined hypothesis that it means that the computer used (laptop, i5) is not powerful enough. Below three tables are presented with an overview of measurement results for the *laptop case*. It has to be noted that only the metrics with most visible differences between tests were presented. The metrics where there was no difference at all - are removed.

Table 48 Test results

QoE Metric (highest difference only)	Test1	Test2	Test3
	Rate = 1Mb/s		
Blockloss			
Freezing			
SA			

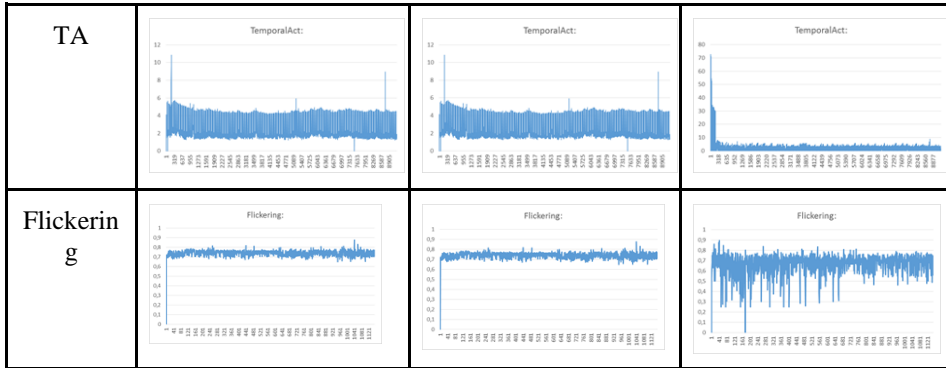
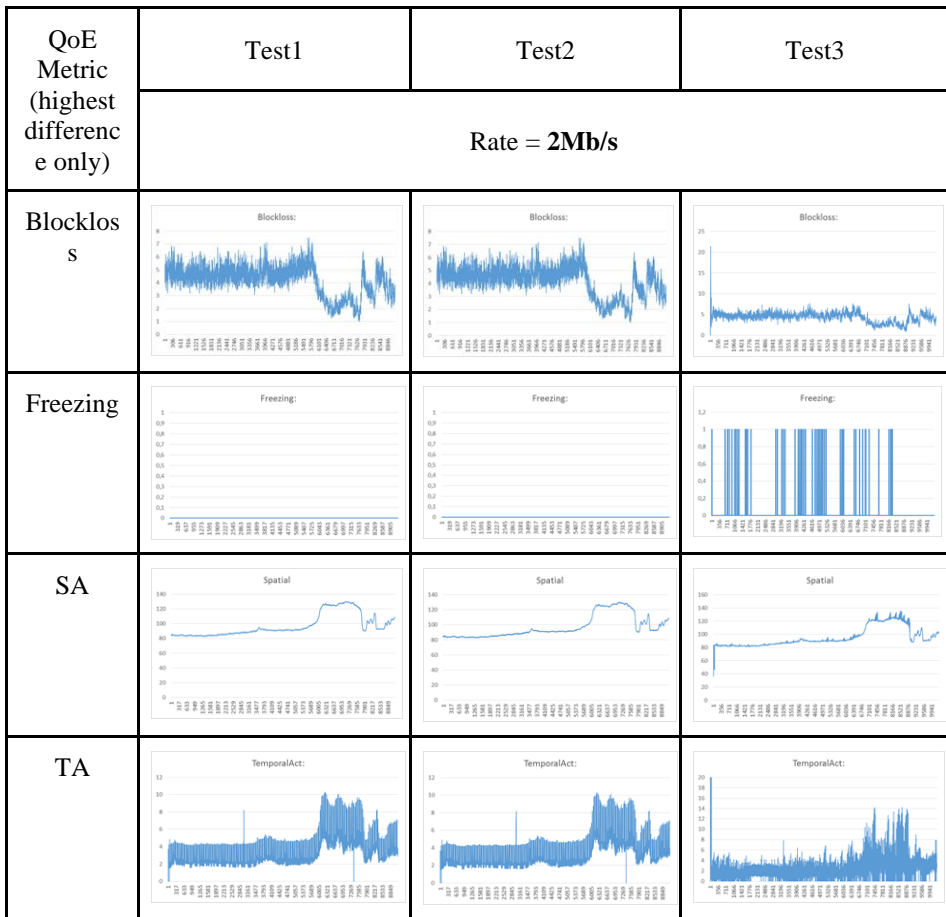


Table 49



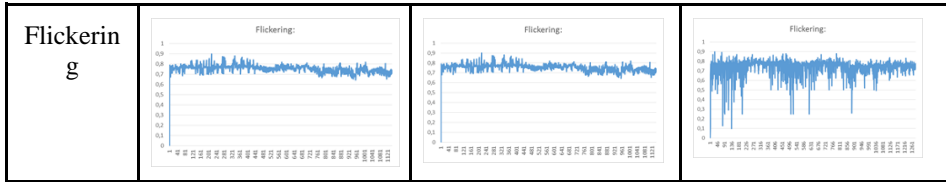


Table 50

QoE Metric (highest difference only)	Test1	Test2	Test3
	Rate = 4Mb/s		
Freezing			
Flickerin g			

19 ANNEX H

In this section an examination and in-depth analysis of available options were the capabilities to dynamically change radio conditions and bandwidth limitation as well as its influence on Linux's kernel is presented. Below most promising options considered are shown:

1. Extracting QoS statistics (OWD, THP, LOSS) directly from csv files and adjusting NETEM setting for each packet – i.e. “brute force” method. Minimizes software demands regarding the process of preparation CSV results but requires very intense controlling process.
2. A simple rate/delay enforcement through thresholding mechanism (engineering approach) - minimizes the number of enforcements while measured delays statistics were equal or below a given level (operatively called "max delay in distributions" threshold).
3. NETEM „custom distribution” approach – using a procedure which enables defining a user-distribution based on the source data file. As a result, one could create a "custom delay" distribution which should be able to recreate delay metrics of measured conditions. Unfortunately this works for delay only.
4. trace based OFDM network emulation “TBONEX” - we assume that TBONEX has:
 - OFDMA scheduler able to assign QoS for different users including:
 - QoS classes of service scheduling
 - priority
 - max/min bandwidth available
 - requested bandwidth
 - chart-driven approach which means that we can define QoS metrics (by customizing users' privileges and activity) based on charts gained from earlier, real life measurements.
5. Using multi-staged custom distributions to recreate a trace - this approach involves thorough statistics analysis of traces and then building mathematical model (based on such distributions). By creating a model, we create a feed generator which forms statistics of QoS metrics variability (OWD, LOSS, THP) stochastically.
6. In-house implementation of HARQ system - it would be ideal approach to control emulator by 1-2 parameters which would be "delay" and "loss". This would automatically affect throughput metrics.

19.1 EVALUATION OF OPTIMAL APPROACHES TO NETWORK DISTURBER

In order to efficiently emulate 4G/5G network traffic, chosen approach specifically requires a set of capabilities from the emulation tool to playback one-way delay (OWD), IP packet loss rate (LOSS) and IP packet throughput (THP) values dynamically from the measurements of the real network. Such requirements cause different challenges and issues that need to be considered to create reliable bandwidth emulation mechanism. Given the requirements mention above related to the ability to dynamically adjust and inject radio conditions including OWD, THP and LOSS, author considered it necessary to create dedicated tool that would be able to set dynamic bit rate option as well as other required properties including delays and packet losses in a dynamic manner as well e.g. per TDMA frame.

20 ANNEX I

20.1 VALIDATING EMULATOR WITH IP CAMERA (SERIES 1)

The tests with IP camera connected in place of the Server node (Figure 55) were performed in this section. The camera provided the 1Mb/s video feed. The tests considered various scenario settings: “no emulation”, “emulation without a replay scenario”, “emulation with delay ramping-up”. The most crucial QoE statistics were evaluated. The detailed steps of Test A were collected in the Annex F. The separate figure for Freezing is missing due to the fact that there was no freezing during the tests.

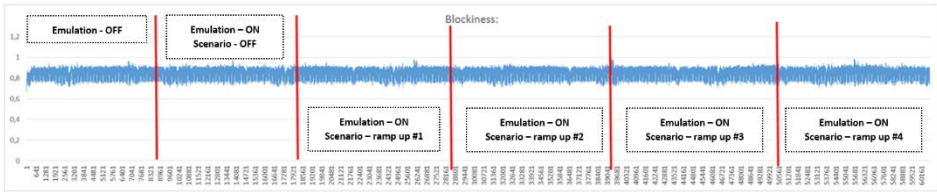


Figure 132 Blockiness metrics comparison between test cases

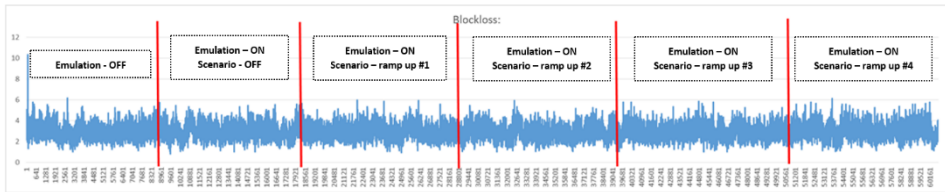


Figure 133 Blockloss metrics comparison between test cases

In next round of tests (Test B) the various scenarios were replayed during 20minute session. This way we are dealing with the different flavours of scenarios (or rather radio conditions) - “good/stable” (UTP Parking, Choszczno Stationary), “more extreme” (Choszczno Mobile), and average (UTP Walk). The detailed settings of the test B are indicated in Annex F.

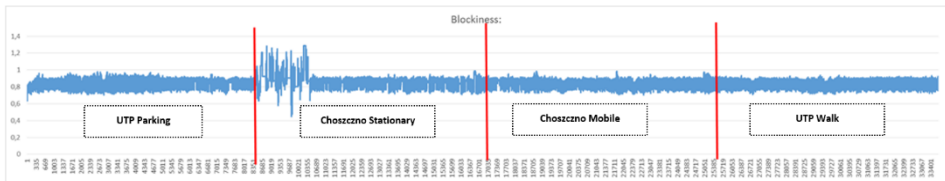


Figure 134 Blockiness metrics comparison between test cases

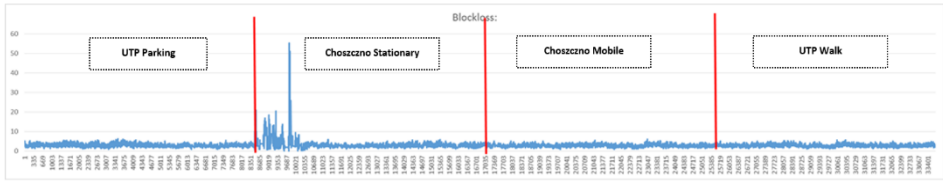


Figure 135 Blockloss metrics comparison between test cases

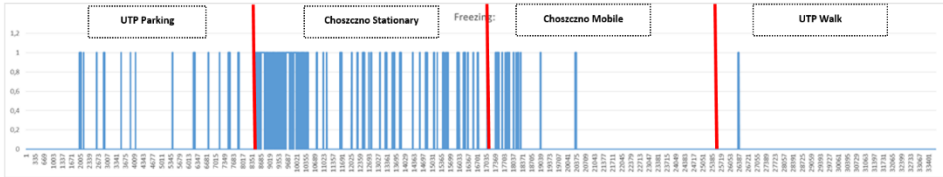


Figure 136 Freezing metric comparison between test cases

It is worth noticing that QoE for all the three parameters Blockiness, Blockloss and Freezing represents low quality, during the second scenario - which should instead provide highest quality of observations. After performing next series of tests (see Section 20.2) but with different tests of the TBONEX it has been figured out that the problem was connected with the “smoothing of delays” which is a feature provided by the emulator in order to improve the injection of delays at the Linux traffic control level. The above mentioned results proven that in reality such feature does not help in replaying the wireless channel dynamics reliably. In fact in this scenario the delays were the “smoothed” because they were most stable, but the smoothing mechanism implementation anyways proved to be too ineffective. After turning it off the freezing was completely removed.

20.2 VALIDATING UDP/TCP USE IN TESTS (SERIES 2)

Presented below are the results from simpler measurements, which aimed at testing the impact of using Server node to transmit videos over UDP and TCP. All measurements lasted 4 minutes, and were made without emulation (marked C), or using a simple script for emulation (labelled D) where we have introduced variable distributions in the form of:

- “no distribution” > $N(50ms, 30)$ > $N(50ms, 50)$ > $N(50ms, 80)$
- where the $N(x,y)$ function denotes the Gaussian function for delay distribution - with x meaning the mean and y - variation in milliseconds.

with minute duration each. The same tests were performed for the UDP and TCP - using VLC to replay videos on the receiver side. It can be seen right away that in the tests with no emulation there was no freezing at all (thus freezing figures are not presented). Also it has been found that the QoE difference between scenarios with UDP and TCP is negligible. The latter fact was also confirmed by

observing video playback during tests.

Table 51 Summary of tests validating the influence of delay distribution

Protocol	Blockiness	Blockloss	Freeze	Protocol	Blockiness	Blockloss	Freeze
TCP-C	0,88	2,479	0	TCP-C	0,818	1,307	0
TCP-D	0,878	2,639	0,043	TCP-D	0,817	1,752	0,183
UDP-C	0,876	2,467	0	UDP-C	0,818	1,395	0
UDP-D	0,879	2,549	0,044	UDP-D	0,801	3,599	0,023

(A) Test1

(B) Test2

Scenario	Average		
	Blockiness	Blockloss	Freeze
1Mb/s	0,742	2,100	0
2Mb/s	0,818	1,307	0

Test3 verifying the difference in QoS for the two stream rates: 1Mb/s, 2Mb/s

20.2.1 Validation of the “delay smoothing” feature of TBONEX (Series 3)

In this test we have repeated the settings from series 2 but with the focus on delay smoothing to validate how it influences the QoE readings at the receiver. Tests were compacted in the table below. The averaged results clearly show that turning this feature on, complicates the issues with packets reordering and eventually the player is not able to properly decode the video stream. This results in at least freezing to become apparent.

Table 52 Test4 - Emulation with/without the delay smoothing option enabled

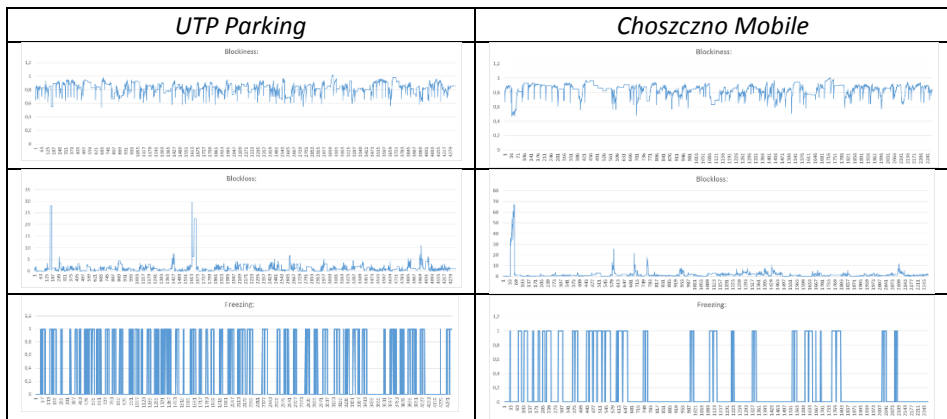
Scenario	Average		
	Blockiness	Blockloss	Freeze

Emulation with delay smoothing	0,833	1,408	0,104
Emulation without delay smoothing	0,829	1,217	0

20.2.2 Validating emulation with both rate and delay enabled

The next step in validating the TBONEX emulation, the complete set of actuations, including delays (as in previous tests) and rate, were enforced together. To perform tests the two traces were used - UTP Parking and Choszczno Mobile, so the average variation and extreme link variations (see Table 33). This time focus was on the comparison of QoE metrics between the two - different - scenarios but for the full emulation.

Table 53 Comparison of QoE between (with delay smoothing)



It can easily be seen in the table Table 53, that results for Choszczno Mobile resemble the case of UTP Parking, the only difference is in the *number of freeze events*. The reason here is that due to more variable channel in the case of Choszczno the delay smoothing mechanism takes almost no effect due to intensity of “actual” distribution. This way this case does not exhibit such intense freezing as in the case of “seemingly” better wireless link in the UTP Parking.

Table 54 Comparison of QoE statistics

Scenario	Blockiness	Blockloss	Freeze
UTP	0,824	1,296	0,384
Choszczno	0,835	1,728	0,243

The table above shows that freeze events are present in both scenarios but their

average is higher in case of UTP Parking.

20.3 VALIDATING THE RESOURCE CONSUMPTION OF VIDEO PROCESSING AT THE VIDEO STREAMING SERVER

In order to identify the sources of video disruption at the transmitter node we have performed number of tests that aimed at identifying the source of disturbance. The components we have evaluated were: VLC player, TR Streamer, FFMpeg. In order to compare the QoE of the video captured utilizing various combinations of the tools in chain we have utilized QoE Probe SW. In order to figure out which one is the source of problems we have designed three tests:

- Test1: Source_video > QoE_Probe
- Test2: Source_video > VLC > QoE_Probe
- Test3: Source_video > FFMpeg > TR Streamer > QoE_Probe

The source video we have used was captured using external IP camera but with different target rates: 1Mb/s, 2Mb/s, 4Mb/s. The video was captured with camera pointed at local street with some minor car traffic on it. The computer we have used was laptop with i5 CPU and desktop with i7 CPU. Original frame rate of the video is 30 fps. The detailed measurements behind this section are collected in the Annex G. To summarize the findings collected in the annex, it can be concluded that:

- Test3 is the most challenging one - Freezing increases as well as Flickering
- The higher the rate the more Freezing is introduced - this shows the lack of processing resources at the host computer
- The higher the rate the less metric is different between tests.

After notifying these differences it was decided to compare various metrics on two computers: laptop (i5) and desktop (i7). The averaged results are shown in the figures: Figure 137, Figure 138. The metrics where there was no difference regardless of the computer type are excluded for brevity.

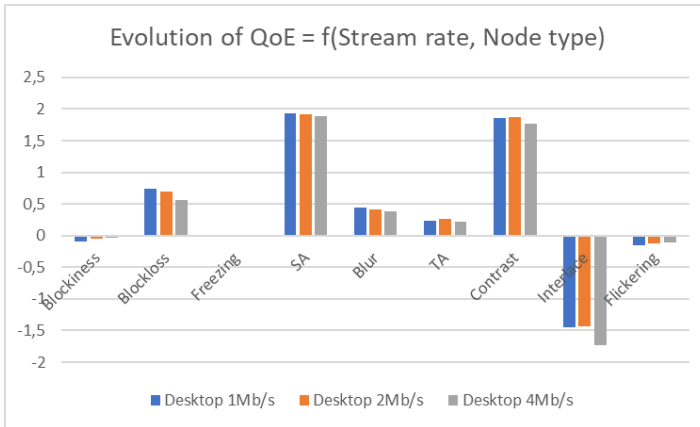


Figure 137 QoE metrics for various bitrates and computing power of the node (Desktop)
- LINK

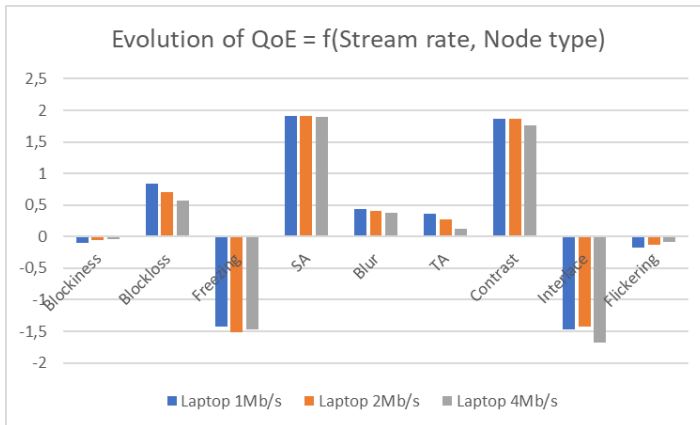


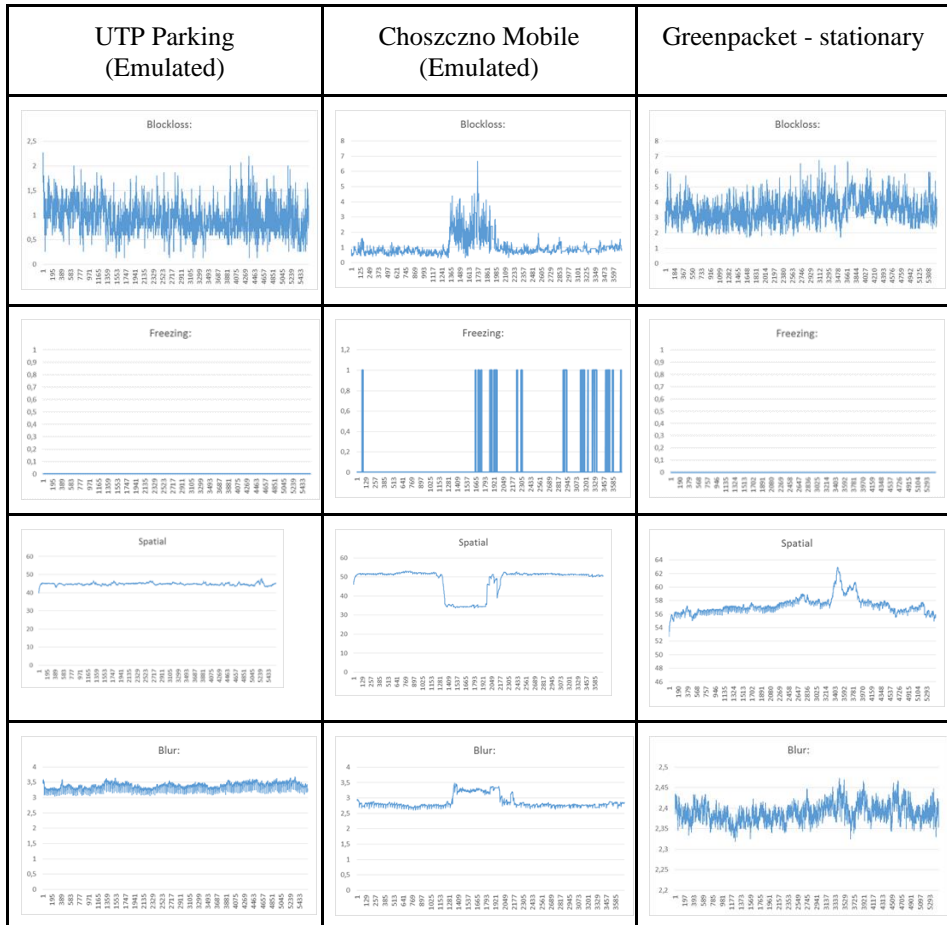
Figure 138 QoE metrics for various bitrates and computing power of the node (Laptop)

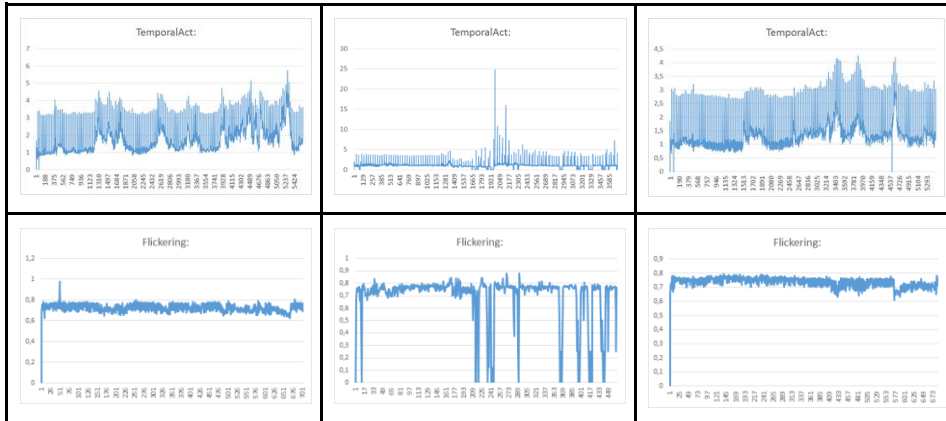
It can easily be seen that (a) 1Mb/s video was worse regarding the blockloss for both computers. Freezing was not identified for stronger machine. The higher the video rate the less blockloss. Blockiness increases (improves) more significantly with the video rate, and only minimally for more powerful machine. In addition to these averaged results it has to be noted that the 4Mb/s video on Laptop was slowed-down due to lack of appropriate level of resources. The human operator also highlights that using the Desktop computer the overall quality perceived was significantly improved compared to the same tests with laptop (although the averaged values of the metrics may not reflect it that clearly). In addition utilizing the Desktop for 4Mb/s video removed the extreme consumption of CPU (from 400% in case of laptop to just 180% for desktop).

20.4 VALIDATING THE INFLUENCE OF TCP USE (INSTEAD OF UDP) WITH FULL EMULATION

In these tests we were interested in the influence on video quality (QoE) when using TCP protocol in delivery of video streams. The test assumed video streamed from camera (same as tests above) with the 1Mb/s average rate. Video was shoot over the outdoor location with low spatial activity. The emulator was used with both parameters enforced: **delay and rate**. Due to previous findings the “delay smoothing” was disabled, not to disturb quality. Two wireless scenarios were used in emulated mode (UTP and Choszczno), in addition for comparison of delivering the same video but through the real 4G network with high quality channel was performed. The results are presented in Figure 68 - the metrics for which results doesn't differ significantly are removed for brevity. It should be noticed that only 6 metrics are reasonably different between scenarios.

Table 55 Comparison of emulator with field test results (Greenpacket)





The interesting findings which are important for the evaluation of the emulator are that:

- Behavior of QoE metrics between “UTP Parking” and the use of real modem - Greenpacket is very similar. Which should be the case as both cases were captured at similarly good radio conditions.
- The freezing (after disabling the “delay smoothing” in TBONEX) is not happening in both scenarios: UTP, Greenpacket, which means that reality and emulation produce satisfactory results. Similarly for Flickering the Greenpacket (real network measurement) and the UTP Parking trace (emulation) behave almost the same with respect to QoE metrics. Meanwhile the “Choszczno mobile” is visibly worse for the quality of the video.
- What is interesting (and unsatisfactory for real modem) is that Blockloss experienced with video delivered over Greenpacket is much worse (5:1 increase) than the one tested with emulated scenario.

20.5 VALIDATING MCATS

In the example case where above script would be used to adjust resolution of the transcoded image (i.e. “size” parameter) the following effects to the transcoder output rate are observed.

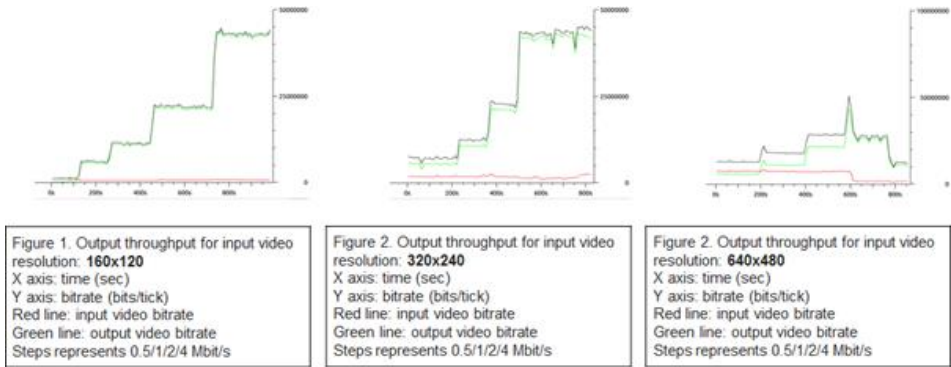


Figure 139 Comparison of video bitrate adaptation based on the controller logic

It can be seen that the green line “reacts” to the requests from controller that is triggering Transcoder to assure particular value of output data rate of the video it takes on input. This proves that when instructed by the monitoring of the “Radio stats” element which can similarly consume the “outputs of the emulator” and requests optimizations from the transcoder to keep the level of QoE at targeted level. The transcoder methods that are utilized within the framework are described in Annex D.