Resource Allocation Techniques for Multicast Streaming over Cellular Mobile Networks

A thesis submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

by

Sadaf ul Zuhra
(Roll No. 144070007)

Under the guidance of:
Prof. Abhay Karandikar
and
Prof. Prasanna Chaporkar

Department of Electrical Engineering
Indian Institute of Technology Bombay
Powai, Mumbai 400076
February 2020

To my parents, Dr Maqsood Ali Shah and Dr. Zuhra Jabeen

Thesis Approval

The thesis titled

Resource Allocation Techniques for Multicast Streaming over Cellular Mobile Networks

by

Sadaf ul Zuhra

(Roll No. 144070007)

is approved for the degree of

Doctor of Philosophy

D. Das

Examiner

Examiner

Advisor

Co Advisor

O. P. Ramani

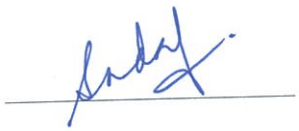
Chairman

Date: 06/02/2020

Place: Mumbai

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.



Sadaf ul Zuhra

144070007

Date: 12/02/2020

Abstract

With an increasing popularity of online streaming platforms, the amount of video content being streamed over mobile cellular networks has seen an enormous increase over the past decade. Videos are projected to account for 79% of the total mobile data traffic by the year 2022. Multicast transmissions provide an excellent means of catering to this bandwidth intensive video streaming traffic. In this thesis, we address various problems related to resource allocation in cellular multicast. For effective use of multicast transmissions, we need to appropriately group users and efficiently allocate resources to multicast streams. Users grouped for multicast transmissions must experience similar channel conditions in addition to requiring the same content. Variations of channel states of users across time and frequency make grouping a complicated problem. We prove that the optimal grouping problem is NP-hard, and hence, no polynomial-time algorithms exist for determining the optimal grouping. We propose a heuristic grouping scheme that divides users into multicast groups based on their average channel states.

Multicast streaming may involve services like online movie premieres that require high quality video content to be delivered to the users. For such services, we consider a lossless multicast system that serves each user in each time slot. We formulate the resource allocation problem for such a lossless multicast system to meet the rate requirements of all users in minimum possible spectrum resources. We prove that this optimal resource allocation problem is NP-hard. Therefore, we propose two efficient heuristic schemes for solving this problem. We also design a randomized algorithm based on Simulated Annealing that works iteratively to estimate the optimal resource allocation and provides a benchmark for performance evaluation of the resource allocation heuristics. Our simulations indicate that the proposed heuristics provide solutions close to the optimal.

The performance of a lossless multicast system that serves all users in each time

slot is always dependent on the weakest user in the system. This leads to a loss in overall system performance and dissatisfaction of users experiencing good channels. We leverage the loss tolerant nature of video streams to overcome these issues. Due to the loss tolerant nature of videos, we can selectively drop packets from video streams without any significant degradation in the quality experienced by end users. For streaming of live events such as sports events, news feeds, factors such as delay take precedence over video quality. We propose using loss tolerant multicasting to serve such applications. We convert the problem of resource allocation in such a loss tolerant multicast system to the problem of stabilizing a constructed virtual queueing system. We propose two loss optimal resource allocation policies for this system. Through extensive simulations, we show that the proposed policies perform significantly better than policies from the existing literature.

Due to the wide variety of services available nowadays, a cell has to cater to a heterogeneous mix of users and services having diverse QoS requirements. The existing literature lacks a generalized resource allocation algorithm that can adapt to optimize any system parameter to meet the QoS requirements of any service. We propose such a generalized auction based resource allocation algorithm that can allocate resources to a diverse set of services simultaneously. We prove that the proposed algorithm is strategy-proof. Hence, it successfully elicits the actual valuations of users for the system resources and maximizes the social utility of the system.

We also propose the use of Multi-Connectivity in multicast transmissions. Multi-connectivity has a potential to improve the performance of video multicast transmissions significantly. We propose the procedures and signaling exchange required for establishing multi-connectivity in cellular multicast. We prove that the resource allocation problem that maximizes the number of users served by multi-connectivity multicast is NP-hard. We propose a greedy approximation algorithm for this resource allocation problem that provides an approximation ratio of $(1 - \frac{1}{e})$. No polynomial-time algorithm can provide a better approximation for this problem.

Since video streaming services are the primary focus of this thesis, we have made use of traces from actual video streams to generate realistic video traffic patterns in our simulations.

Contents

Abstract **vii**

Acknowledgments **xiii**

List of Acronyms **xv**

List of Symbols **xix**

List of Tables **xxi**

List of Figures **xxiii**

1 Introduction **1**

1.1 Overview of Resource Allocation in LTE 4

1.2 Overview of Multicast in LTE 5

1.3 Challenges in Multicast Grouping and Resource Allocation 7

1.3.1 Grouping 8

1.3.2 Resource Allocation 10

1.4 Motivation for the Thesis 10

1.5 Contributions and Organization 13

2 Multicast Streaming: Literature and Open Problems **17**

2.1 Multicast Group Formation 18

2.2 Resource Allocation in Multicast 19

2.2.1 Opportunistic Multicast Scheduling 19

2.2.2 Joint Optimization for Unicast and Multicast 20

2.2.3 Multicasting of DASH and SVC Content 21

2.2.4 Auction Based Multicast Scheduling and Pricing 23
2.2.5 Broadcasting 24
2.3 Loss Tolerant Multicast Streaming 25
2.4 Network Coding for Multicast 25
2.5 Multi-Connectivity 26
2.6 Open Research Problems 27

3 Grouping and Resource Allocation for Lossless Multicast Transmissions 31
3.1 System Model 33
3.2 Problem Definition 35
3.2.1 Problem 1: Optimal Resource Allocation $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ 35
3.2.2 Problem 2: Optimal Grouping $\mathbf{C}^{\star}$ 39
3.3 Randomized Algorithm for Optimal Resource Allocation 42
3.3.1 DTMC Construction 45
3.3.2 Performance comparison of the RS and the BLP 54
3.4 Heuristic Schemes for Resource Allocation 55
3.4.1 Greedy Allocation 55
3.4.2 LP-relaxation Based Allocation 56
3.5 Heuristic Scheme for Grouping 59
3.5.1 Hybrid Grouping Policy 59
3.6 Simulations 61
3.6.1 Results 63
3.7 Generalizations 70
3.8 Conclusions 72

4 Resource Allocation for Loss Tolerant Multicast Video Streaming 73
4.1 System Model and Problem Formulation 75
4.1.1 System Model 75
4.1.2 Problem Definition 77
4.2 Queueing System for Resource Allocation in Loss Tolerant MBMS Systems 78
4.2.1 Construction 78

4.2.2 Feasible Region of the Optimal Resource Allocation Policy and Stability Region of the Queueing System . . . 80
4.3 Proposed Resource Allocation Algorithms . . . 87
4.3.1 Loss Optimal Resource Allocation . . . 87
4.3.2 Priority Loss Optimal Resource Allocation . . . 90
4.3.3 Modified Exponential (Queue length) Rule (Γ_E) . . . 94
4.3.4 Computational Complexity . . . 95
4.4 Polynomial-time Implementation of LORA, p-LORA and Modified EXP-Q 95
4.5 Simulation Results . . . 97
4.6 Conclusions . . . 103

5 Resource Allocation and Pricing for Heterogeneous User Demands 105
5.1 System Model . . . 106
5.2 Problem Formulation . . . 108
5.3 A VCG Based Mechanism for Generalized Resource Allocation and Pricing 109
5.3.1 Computational Complexity of $\Gamma^\star$ . . . 113
5.4 MWBM Implementation of $\Gamma^\star$ . . . 113
5.5 Simulations . . . 115
5.6 Conclusions . . . 117

6 Multi-Connectivity Multicast Streaming 119
6.1 MBMS and Multi-Connectivity . . . 121
6.2 Procedures for Establishing Multi-Connectivity in Multicast Transmissions 123
6.3 Resource Allocation in MC Multicast . . . 127
6.3.1 System Model . . . 127
6.3.2 Problem Formulation . . . 128
6.4 $\mathbf{K}^\star$ is NP-hard . . . 129
6.5 Approximation Algorithm for $\mathbf{K}^\star$ . . . 130
6.6 Distributed versus Centralized Allocation . . . 133
6.6.1 Distributed Greedy Allocation . . . 134
6.7 Simulations . . . 135
6.8 Conclusions . . . 142

7 Summary of Results and Future Directions **143**

7.1 Summary of Results 143

7.2 Future Research Directions 148

Acknowledgments

It takes a village, they say. To raise a child, yes, but even more so, to get through an engineering doctorate. So, I have a lot of people I need to thank and acknowledge. Please bear with me as I do so because this is extremely important to me. If you do decide to read this, I must ask you to stick with me till the end because each and every one of these people have contributed to bringing this work to fruition.

I would like to begin by thanking my advisor, Prof. Abhay Karandikar, for being an incredible mentor and a father figure to me. His vision, knowledge, and expertise have been instrumental in defining and shaping this work. His guidance during the formulation, research, and writing have been exceedingly useful. He has been a constant source of inspiration for me, not only because of his towering intellect but also his kindness, humility, and enthusiasm. He has provided ample support and encouragement whenever I needed it. He has taught me to hold myself to the highest standards of integrity in whatever I do. I am grateful for the wisdom and knowledge I have received from him.

My sincere thanks to my co-advisor Prof. Prasanna Chaporkar, for his guidance and encouragement. All the work in this thesis is a result of hours of discussions and brainstorming with him. He has always encouraged me to push my abilities as a researcher, and I am very grateful for it. I would like to express my deepest gratitude to him for providing support and help every step of the way. I will always take pride in being known as the student of Prof. Karandikar and Prof. Chaporkar.

I would like to thank my Research Progress Committee members, Prof. Jayakrishnan Nair and Prof. Saravanan Vijayakumaran, for their invaluable inputs throughout my Ph.D. Through their questions and feedback, they have often given new directions to my work, which has been instrumental in shaping this thesis. I wholeheartedly appreciate all their help and guidance.

I wish to show my gratitude to Mr. Pranav Jha for all the help he has provided. His technical insights have been invaluable in bringing this work to completion. I would also like to pay my regards to all the professors and HOD of the EE department. I consider it a great honor to have been taught by and interacted with such accomplished people.

I am grateful to Sonal, Margaret, Beena, and Mrs. Sangeeta for all their help. Our lives are made easier because of their efficient assistance with all the paperwork. I am thankful to Aditya Ji and Rajesh Ji for providing help with logistics. My sincere thanks to the incredibly efficient EE office staff, particularly Mr. Santosh and Mrs. Madhu, whose promptness and readiness to help have made this journey smooth.

I would like to thank so many of my friends and colleagues who not only helped me cope up with the challenges of this degree but made it fun. My heartfelt gratitude to Irshad, Shuvolina, Meghna, Shubham, Rashmi, and Parvathi, who have stood by my side every single day. Reaching the finish line would have been impossible without their continuous encouragement and kindness. I owe a great deal to our discussions over hot cups of 'adrak wali chai'. I am thankful to them for believing in me and lending me courage when I failed to find any within myself.

I would like to thank my friends and lab mates for the many fruitful discussions over the years. I am indebted to Arghyadip for always being there to lend help and support. I would like to thank Akshatha, Shashi, Indranil, Sweety, Indu, Annapurna, Pradnya for creating and nurturing an environment of intellectual growth and positivity.

I am forever indebted to my wonderful parents and sister, who taught me to value education above all else and to hold myself to the highest standards of integrity and honesty in my work. They have inspired me to be a lifelong learner and encouraged me to commence this journey in the first place. Above all, I want to thank ALLAH for blessing me with a life that enabled this journey and for all these people who made it possible.

Lastly, I wish to thank my excellent institute IIT Bombay and the beautiful city of Bombay with all my heart. It has been an experience of a lifetime.

Sadaf ul Zuhra
February 2020

List of Acronyms

4G	Fourth Generation
5G	Fifth Generation
3GPP	Third Generation Partnership Project
a.s.	almost surely
AVC	Advanced Video Coding
BCCH	Broadcast Control Channel
BM-SC	Broadcast-Multicast Service Centre
BS	Base Station
BLP	Binary Linear Program
CGA	Centralized Greedy Approximation
CP	Cyclic Prefix
CQI	Channel Quality Indicator
CSI	Channel State Information
DASH	Dynamic Adaptive Streaming over HTTP
DC	Dual Connectivity
DCI	Downlink Control Information
DG	Distributed Greedy
DTMC	Discrete Time Markov Chain
eMBMS	Evolved Multimedia Broadcast Multicast Services
eNodeB/eNB	Evolved NodeB
EPC	Evolved Packet Core
FEC	Forward Error Correction
HD	High Definition
HSPA	High Speed Packet Access

i.i.d	Independent and Identically Distributed
IP	Internet Protocol
ISP	Internet Service Provider
LORA	Loss Optimal Resource Allocation
LTE	Long Term Evolution
LTE-A	Long Term Evolution - Advanced
LP	Linear Programming
LPr	LP-relaxation Based Allocation
MBMS	Multimedia Broadcast Multicast Services
MBMS-GW	Multimedia Broadcast Multicast Services Gateway
MBSFN	Multimedia Broadcast Multicast Service Single Frequency Network
MC	Multi-Connectivity
MCCH	Multicast Control Channel
MCMC	Markov Chain Monte Carlo
MCP	Maximum Coverage Problem
MCS	Modulation and Coding Scheme
MDP	Markov Decision Process
MDT	Mobile Data Traffic
MIB	Master Information Block
MTCH	Multicast Traffic Channel
MWBM	Maximum Weight Bipartite Matching
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OMS	Opportunistic Multicast Scheduling
p-LORA	Priority Loss Optimal Resource Allocation
PRB	Physical Resource Block
QoE	Quality of Experience
QoS	Quality of Service
RAT	Radio Access Technology
RNLC	Random Network Linear Coding
RS	Randomized Scheme

RRC	Radio Resource Control
SA	Simulated Annealing
SDN	Software Define Networking
SFN	Single Frequency Network
SIB	System Information Block
SNR	Signal to Noise Ratio
SVC	Scalable Video Coding
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution
TPM	Transition Probability Matrix
UE	User Equipment
UEP	Unequal Error Protection
UT	User Terminal
VCG	Vickrey-Clarke-Groves
VoIP	Voice over Internet Protocol
w.h.p.	with high probability
w.p.	with probability
WiMAX	Worldwide Interoperability for Microwave Access

List of Symbols

$\|.\|$	Cardinality of a set
M	Number of multicast UEs
L	Number of multicast groups
N	Number of PRBs in a sub-frame
$[n]$	$\{1, 2, \ldots, n\}$
$\mathcal{N}$	Set of available PRBs in a sub-frame
$\mathcal{L}$	Set of multicast groups
$h_{iu}[t]$	Channel gain of UE u on PRB i in sub-frame t
$\overline{h}_{iu}$	Average channel gain of UE u in PRB i
$H_{iu}[t]$	Fast fading component of the channel gain of UE u in PRB i in sub-frame t
$r_{iu}[t]$	Maximum rate supportable by UE u on i^{th} PRB in sub-frame t
Δ	Grouping policy
$\Delta^\star$	The optimal grouping policy
G_i^Δ	i^{th} group under grouping policy Δ
Γ	Resource allocation policy
$\overline{V}_{i\Gamma}^\Delta$	Set of PRBs assigned to G_i^Δ under policy Γ
R	Rate requirement of the multicast UEs
$S_\Gamma^\Delta[t]$	Number of PRBs left unutilized under Δ in sub-frame t using policy Γ
S_Γ^Δ	Average number of PRBs left unutilized under Δ using policy Γ
$x_{ij}[t]$	Indicator random variable that equals 1 when PRB j is assigned to group i in sub-frame t
$\mathbf{B}_{\mathbf{\Delta}}^\star$	Optimal resource allocation problem for lossless multicast transmission
$\mathbf{B}_{\mathbf{D}}^\star$	Decision problem corresponding to $\mathbf{B}_{\mathbf{\Delta}}^\star$
$\mathbf{C}^\star$	Optimal grouping problem

r_{max} Maximum rate achievable in a PRB

$\ell_k^{\Gamma}[t]$ Loss indicator for UE k under policy Γ in sub-frame t

$\tilde{\ell}_k$ Loss tolerance of UE k

$\tilde{\boldsymbol{\ell}}$ Loss tolerance vector of the system

$\mathcal{L}^{\Gamma}$ Feasible region of policy Γ

$\mathcal{L}$ Feasible region of MBMS system

$Q_k[t]$ Token queue length of UE k in sub-frame t

$A_k[t]$ Arrival process of queue k in sub-frame t

$D_k[t]$ Departure process of queue k in sub-frame t

$c_k[t]$ Priority weight of queue k in sub-frame t

$\mu_k^{\Gamma}[t]$ Service rate of token queue k in sub-frame t

$\boldsymbol{\lambda}$ System arrival rate vector

$\boldsymbol{\mu^{\Gamma}[t]}$ System service rate vector in sub-frame t under policy Γ

$\mathcal{S}$ Stability region of queueing system

$\mathbf{Q[t]}$ State of the queueing system in sub-frame t

Γ_0 Loss optimal resource allocation policy

Γ_P Priority loss optimal resource allocation policy

Γ_E Modified EXP-Q rule

$v_k[t]$ Valuation of UE k for being scheduled in sub-frame t

$b_k[t]$ Bid submitted by UE k for being scheduled in sub-frame t

$p_k^{\Gamma}[t]$ Price to be paid by UE k for being scheduled in sub-frame t under policy Γ

$u_k^{\Gamma}[t]$ Utility obtained by UE k in sub-frame t under policy Γ

$i(k)$ Index of the group to which UE k belongs

U_{jc} The set of users that would be successfully served if PRB j is allocated to the multicast service in cell c

$\mathbf{K}^{\star}$ Optimal resource allocation problem in MC multicast

List of Tables

3.1 Performance comparison of RS and BLP 54
3.2 Time taken in seconds to run RS and LPr 58
3.3 System Simulation parameters [1] 62
3.4 Average number of groups formed 64
4.1 System Simulation parameters [1] 98
5.1 System Simulation parameters [1] 115
6.1 System Simulation parameters [1, 2] 136

List of Figures

1.1 Multicast transmissions in a cell 2
1.2 Resource block structure in LTE 6
1.3 MBMS architecture 7

3.1 Number of PRBs saved under LPr and RS. 57
3.2 Variation of the reward of the state of RS with the increasing number of iterations. 58
3.3 Number of PRBs saved under various policies 64
3.4 Number of infeasible cases under various 65
3.5 Number of PRBs saved under various resource allocation policies for $R = 2$ Mbps 65
3.6 Number of infeasible cases under various resource allocation policies for $R = 2$ Mbps 66
3.7 Number of PRBs saved for different UE placements under the Greedy scheme 67
3.8 Number of PRBs saved for different UE placements under the LPr scheme 67
3.9 Histogram of number of PRBs saved for a real-time video stream under the Greedy scheme 68
3.10 Histogram of number of PRBs saved for a real-time video stream under the LPr scheme 68
3.11 Comparison of the average system throughput under LPr and PF 69
3.12 Comparison of the percent unsatisfied groups under LPr and PF 70

4.1 Virtual queueing system model 79
4.2 Bipartite graph between multicast groups and PRBs 96
4.3 Tolerable loss versus loss encountered using LORA 99

4.4 Tolerable loss versus loss encountered using p-LORA 99
4.5 Tolerable loss versus loss encountered using EXP-Q 100
4.6 Comparison of average losses in LORA, EXP-Q and p-LORA schemes . . . 101
4.7 PSNR degradation of different videos 101
4.8 Loss pattern of a UE (losses per sub-frames) 102

5.1 Bipartite graph between multicast groups and PRBs 114
5.2 Tolerable loss versus loss encountered 116
5.3 Average loss pattern over time 116

6.1 Procedure for enabling multi-connectivity multicast for UEs in RRC idle mode and single connected UEs. 124
6.2 Procedure for enabling multi-connectivity multicast for dual connected UEs.126
6.3 A snapshot of the simulation scenario 135
6.4 Average umber of packets received successfully under MC using centralized and distributed allocation 137
6.5 Average number of packets received successfully under greedy approximation algorithm as a function of increasing number of users 137
6.6 Average number of packets received successfully under greedy approximation algorithm as a function of cell radius 138
6.7 Average number unserved users under greedy approximation algorithm as a function of increasing number of users 139
6.8 Average number unserved users under greedy approximation algorithm as a function of cell radius 139
6.9 Average number of packets received successfully under MC and SC multicast for a real-time video stream 140
6.10 Average number of unserved UEs under MC and SC multicast for a real-time video stream 140
6.11 Average number of packets received successfully under MC multicast and MBSFN for a real-time video stream 141
6.12 Average number unserved users under MC multicast and MBSFN for a real-time video stream 141

Chapter 1

Introduction

The unprecedented growth of mobile data traffic in the last decade has been the main driver of technological advancements in mobile telecommunication. The global Mobile Data Traffic (MDT) is expected to grow at a compound annual growth rate of around 46 percent from 2017 to 2022, reaching 77.5 ExaBytes (EB) per month by 2022 [3]. By 2022, more than 90 % of the MDT will emanate from smartphones [3]. In India alone, the yearly MDT increased from 0.83 EB in 2014 to 46.4 EB in 2018 [4], witnessing a 58 fold increase in a span of four years. Video traffic is the largest contributor to this massive amount of mobile data. Videos are projected to account for 79% of the total MDT by the year 2022 [3]. The amount of mobile video traffic will reach 69 EB per month in 2022 from 8.5 EB per month in 2016 [5]. With the widespread deployment of Fourth Generation (4G) Long Term Evolution (LTE) worldwide, the number of high-speed mobile connections has seen an enormous increase. This has also contributed to an unparalleled amount of videos being sent over the Internet every day.

The explosion of video traffic has essentially transitioned us from the age of downloads to an age of streaming. This paradigm shift has been primarily driven by the growing popularity of platforms like Netflix, YouTube, Hotstar, Hulu, Amazon Prime Video. Prevalence of such streaming services has led to a fundamental shift in the way users consume online video content. Users increasingly prefer streaming content over cellular networks on the go on their mobile devices like smartphones and tablets. This often involves online streaming of television (TV) programs, live streaming of major world events, sports matches, software updates, news feeds. All these applications require transmitting

the same content to a large audience simultaneously.

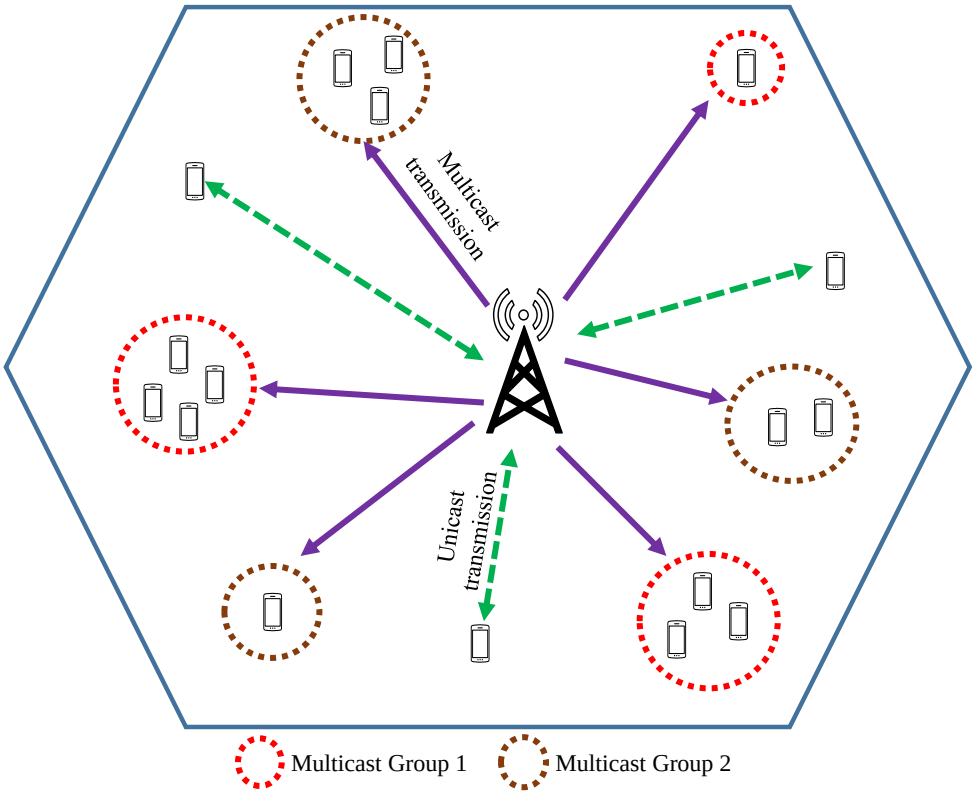


Figure 1.1: Multicast transmissions in a cell

Current cellular communications are primarily based on one-to-one or unicast transmissions. In unicast, the base station communicates with each user separately, using different resources for each one of them. Using unicast transmissions for the above mentioned applications and assigning orthogonal resources to all users receiving the same content consumes a substantial fraction of the limited amount of spectrum available for use by cellular systems. This has created an immediate need for techniques that can better utilize the system bandwidth and accommodate this surge in video traffic within the available spectrum resources. Multicast transmission is one such technique that can significantly ease the burden on cellular resources. Multicast refers to one-to-many transmissions in which several users receive the same content over shared spectrum resources. Figure 1.1 illustrates the use of multicast transmissions in a cell. Using multicast instead of unicast for serving streaming applications enables efficient bandwidth usage as it can accommodate more users and services within the available resources.

Using multicast saves valuable resources which enable the network to support users in numbers otherwise impossible to handle using unicast. Let us take an example to understand this. Consider a cell with 50 users requiring the same content and 50 channels available for allocation in a time slot. This is an everyday practical use case when streaming live events. Let the minimum rate required by the users be 10^3 bits/slot. Consider that all users are experiencing similar channel states so that we can transmit over 10^3 bits in each channel. If unicast transmission is used here, one channel will be allocated to each user to satisfy the required rate, and all the available channels will be used. On the other hand, if we make a single multicast group of all the users, we can provide the required rate in a single channel to all the users. Multicast, therefore, uses 49 fewer channels than unicast, which can be used to support more traffic. Instead of 50 users, if we had 60 users in this example, unicast transmission would become infeasible, and 10 users would have to be blocked. However, with multicast, we can still provide the required rate to all users in a single channel. This example illustrates how multicast can support many more users than unicast transmissions. Multicast has recently been garnering much attention in the research community as well as the industry. As of September 2018, KT, Verizon, Telstra and Reliance (Jio) have already deployed LTE multicast services, and 41 operators have invested in LTE multicast [6] in the form of trials and deployments worldwide.

We consider the use of multicast transmissions primarily for video streaming in LTE and Fifth Generation (5G) cellular networks. In Sections 1.1 and 1.2, we discuss certain aspects of LTE related to resource allocation and multicast, respectively. In Section 1.3, we discuss the challenges presented by optimal grouping and resource allocation problems in cellular multicast transmissions. We then discuss the motivation of the thesis in Section 1.4. The organization and the main contributions of this thesis are presented in Section 1.5.

Remark 1. *The work done in this thesis was started in* 2014 *when LTE systems were being deployed worldwide but as of writing this thesis, work on 5G technology and standards is in advanced stage. While we discuss most of the problems in this thesis in the context of an LTE system, all the results and algorithms proposed are also equally applicable to the 5G cellular systems. We discuss the necessary generalizations of the proposed solutions wherever needed.*

1.1 Overview of Resource Allocation in LTE

LTE is part of the Fourth Generation (4G) of wireless communications. LTE uses Orthogonal Frequency Division Multiple Access (OFDMA) for downlink transmissions and Single-carrier FDMA (SC-FDMA) for uplink transmissions. In LTE, the base station is given the name of evolved NodeB (eNB) and any device that enables the end users to communicate with the LTE system is called a User Equipment (UE). The bandwidth of LTE systems is variable ranging from 1.4 to 20 MHz. The available system bandwidth is divided on a time and frequency scale. On the temporal scale, the available bandwidth is divided into radio frames. A radio frame in LTE spans 10 ms and consists of 10 sub-frames of 1 ms each. A sub-frame is further composed of two slots of 0.5 ms each, and each slot consists of 7 OFDM symbols. On the frequency scale, a sub-frame is partitioned into blocks of 12 sub-carriers spanning 180 kHz with a sub-carrier spacing of 15 kHz. The resource block formed by 12 sub-carriers and spanning over one slot (0.5 ms) is termed as a Physical Resource Block (PRB). This resource structure is illustrated in Figure 1.2. Resource allocation in LTE is done on the scale of PRBs. The number of PRBs in a sub-frame ranges from 6 in a 1.4 MHz LTE system to 100 in a 20 MHz system.

The amount of data that can be transmitted in a PRB is determined by the Modulation and Coding Scheme (MCS) used. In LTE, UEs report their channel states to the eNB in the form of a 4 bit value known as the Channel Quality Indicator (CQI). This 4 bit indicator can take 15 distinct integral values. CQI is directly proportional to the channel gain of a UE, i.e., a higher CQI indicates a better channel. A higher CQI also means that the eNB can use a better modulation scheme for transmitting data to that UE. This relationship between CQI and MCS is defined in Third Generation Partnership Project (3GPP) standards for LTE [7]. These CQI to MCS mappings determine the modulation scheme and code rate that can be used while transmitting data to a UE. The process of mapping is carried out as follows. We first determine the SNR of the UE which is mapped to the CQI using the CQI table specified in 3GPP R1-081483 [8]. The corresponding MCS index is then determined based on the spectral efficiency using the MCS table also given in 3GPP R1-081483 [8]. Finally, the MCS index is mapped to the Transport Block Size (TBS) index and the TBS index gives us the TBS using the mappings specified in Tables $7.1.7.1-1$ and $7.1.7.2.1-1$ of 3GPP TS 36.213 [7] respectively. Thus, the amount of

data that can be sent in one PRB is a function of the CQI of the user. The channel gain of a UE and hence its CQI can vary from one sub-frame to another. Due to frequency selective fading, the CQI of a user also varies across PRBs in a sub-frame. Therefore, the rate at which data can be transmitted to a UE in a particular sub-frame is different for different PRBs.

Allocation of resources in LTE is done once every sub-frame (1 ms). The resource allocation information is contained in a Downlink Control Information (DCI) which is conveyed to the UEs over the Physical Downlink Control Channel (PDCCH). DCI informs UEs, among other things, of which PRBs carry their data and what kind of modulation has been used to send the data. The UE then uses this information to decode the data sent by the eNB successfully.

1.2 Overview of Multicast in LTE

The provisions for multicast and broadcast services in LTE are known as Multimedia Broadcast Multicast Services (MBMS). MBMS was introduced in Release 9 of the 3GPP standards for LTE [9]. Multicast in MBMS functions like a subscription service. It consists of eight phases, namely, subscription, service announcement, joining, session start, MBMS notification, data transfer, session stop, and leaving. Of these, subscription, joining, and leaving are up to the user. The service announcement informs users of the available MBMS services. The user can then get associated with a particular service via subscription and join to indicate that it is interested in receiving the MBMS content.

To support MBMS, three new network elements have been added to the LTE architecture, Broadcast Multicast Service Centre (BM-SC), MBMS GateWay (MBMS-GW) and Multicell/Multicast Coordination Entity (MCE) [10]. The positioning of these elements in the architecture is shown in Figure 1.3. BM-SC serves as an interface between core network and multicast/broadcast content providers. It is responsible for transporting MBMS data into the core network, managing group memberships and subscriptions and charging for MBMS sessions [9]. MCE is responsible for allocating radio resources to the eNBs [9] for Multimedia Broadcast Single Frequency Network (MBSFN) operations. MBMS-GW uses IP multicast to forward the MBMS session data to the eNBs. The eNBs

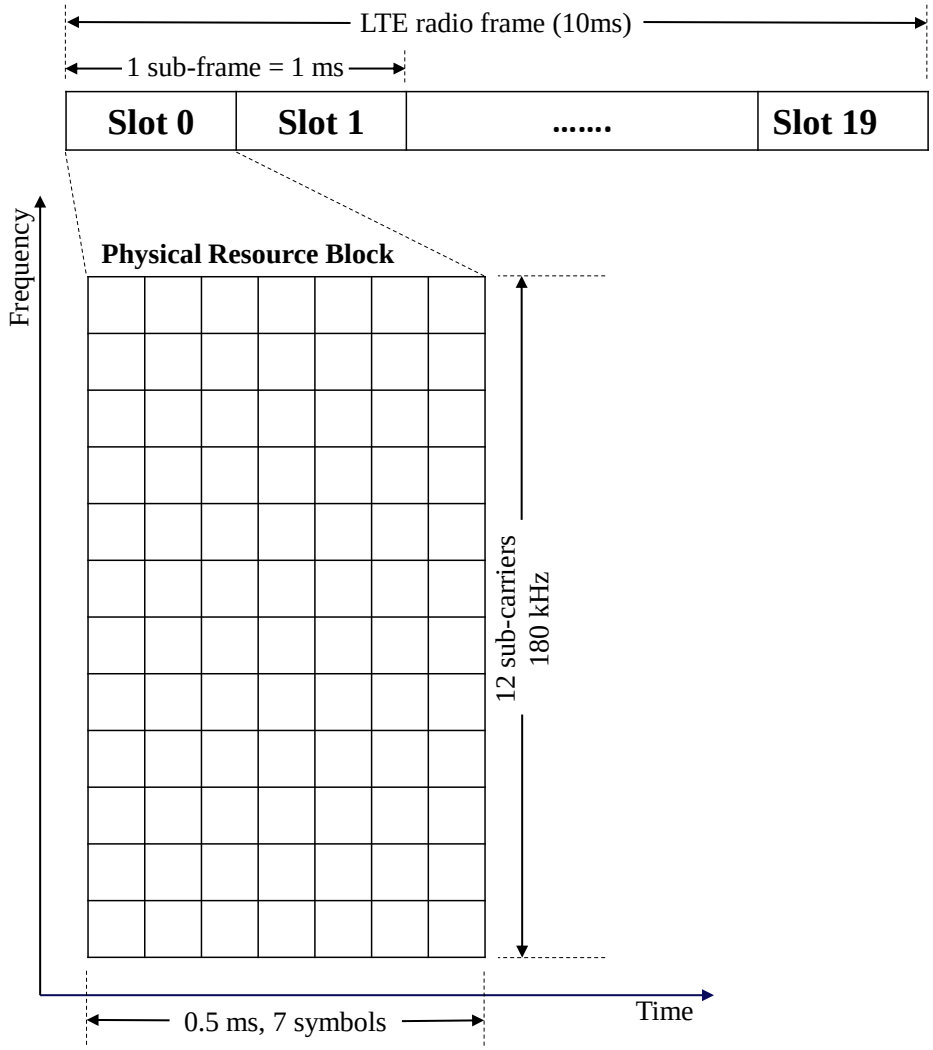


Figure 1.2: Resource block structure in LTE

can then transmit the data to the UEs via wireless multicast/broadcast.

MBMS defines two modes of operation, namely, MBSFN and Single Cell Point-to-Multipoint (SC-PTM) transmissions. In MBSFN operations, several eNBs in an MBSFN area transmit the same content in strict synchronization. Owing to tight synchronization between transmissions, content from different eNBs is perceived by the UEs as multipath transmissions from a single source. This results in improved spectral efficiency, especially at the cell edge. The signals from the neighboring eNBs, which would act as interference in regular operation, interfere constructively with the useful signal and reinforce it. The reinforced signal results in higher SNR and hence improves the spectral efficiency of the system. SC-PTM mode of MBMS transmissions involves multicasting/broadcasting of

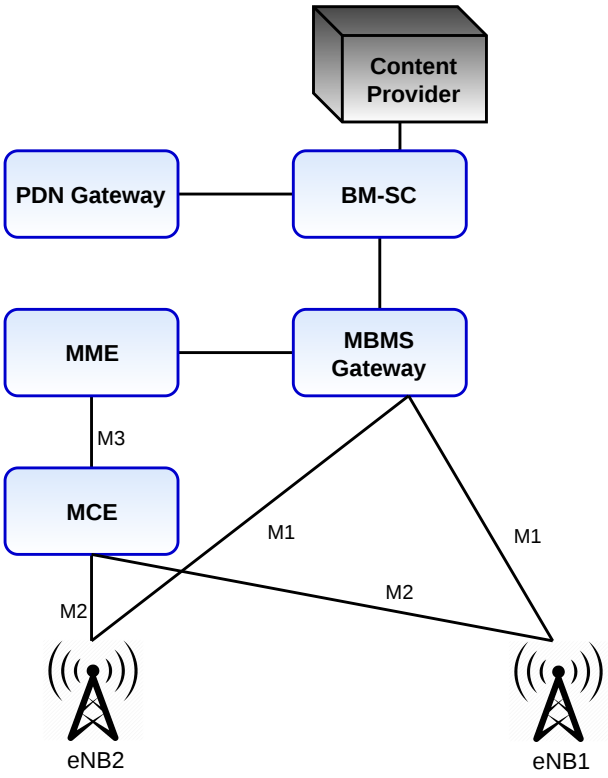


Figure 1.3: MBMS architecture

content within an individual cell separately.

MBMS is an idle mode operation [11] which means that there is no need for a Radio Resource Control (RRC) connection to be established for a UE to receive MBMS services. The control information required by a UE for receiving MBMS services is contained in SystemInformationBlockType13 (SIB13) which is transmitted by the eNB over Broadcast Control Channel (BCCH). SIB13 provides the control information needed by the UE to read Multicast Control Channel (MCCH). Using the information conveyed over MCCH, the UE can then procure Multicast Traffic Channel (MTCH) of the service that it wants to receive.

1.3 Challenges in Multicast Grouping and Resource Allocation

In this thesis, we address various problems related to grouping and resource allocation in multicast transmissions. For successfully using multicast transmissions in a cellular

system, two main challenges need to be addressed. The first is the problem of determining how to divide UEs into multicast groups. The set of UEs grouped together to be served on common spectrum resources form a multicast group. The UEs in a multicast group are treated as a single entity by the eNB and are served using the same PRBs. Which UEs are grouped together is dictated by the criteria used for grouping. One obvious requirement is for all the UEs in a group to require the same content. However, as we shall see later, grouping all the UEs who need the same content together may lead to a degraded system performance due to varied channel gains experienced by them. Therefore, channel gains of the UEs also need to be considered during group formation. Variation of the channel gains of users over time and different frequencies makes optimal group formation a complex problem. The second challenge in multicast transmissions is that of allocating resources to the multicast groups. Unlike unicast, where PRBs are allocated to individual users, in multicast, PRBs are allocated to groups of users. Different users typically experience different channel states in a PRB. As a result, while assigning PRBs to a multicast group, the channel states of all the users in it need to be taken into consideration. This makes the resource allocation problem for multicast significantly more complicated than that for unicast transmissions. We discuss the challenges associated with these two problems in greater detail in the following sections.

1.3.1 Grouping

We begin by discussing how grouping plays a crucial role in multicast transmissions. When a PRB is allotted to a set of users, the rate at which reliable transmission can take place corresponds to the user with the worst channel state in the group. Due to this dependence on channels, creating groups at random can lead to degraded system performance. Let us consider an example to illustrate this. Say we have 10 PRBs available for allocation in a sub-frame and two users, U_1 and U_2 in the cell who require the same content. Let the rate required by each user be 10^3 bits/sub-frame. Consider a channel state where U_1 has a good channel in all odd numbered PRBs in which as many as 10^3 bits can be transmitted at a time. In the rest of the PRBs, U_1 can get a maximum of 100 bits each. U_2 has a good channel in even-numbered PRBs and can receive 10^3 bits in each of them and 100 bits in odd numbered PRBs. Now, if we choose to group these users for multicast transmission,

we need to transmit data at the rate of the weakest user in the group. In this case, we can transmit at most 100 bits in each PRB. Therefore, to satisfy the required rate, we need to use all 10 PRBs. On the other hand, if we use unicast transmissions, U_1 can be allotted PRB 1, U_2 can be allotted PRB 2 and 10^3 bits can be transmitted in each of these PRBs. In this case, the required rate for both users is satisfied in just 2 PRBs, 8 fewer than the multicast scenario. This example shows that appropriate grouping of users is essential for obtaining any benefit whatsoever from multicast operations.

As is clear from the above examples, the channel states of users need to be taken into account while dividing the multicast UEs into groups. The main challenge in grouping UEs based on their channel gains is that, due to fast fading, channel gains experienced by UEs keep on changing. As a result, grouping done based on channel states in one sub-frame may not be optimal in subsequent sub-frames. Grouping UEs based on their instantaneous Signal-to-Noise Ratio (SNR) in every sub-frame is also not feasible as it leads to increased control overhead due to frequent changes in grouping. Since each multicast group is treated as a unique entity by the eNB, each group is assigned a unique MBMS Radio Network Temporary Identifier (M-RNTI). M-RNTI of a group is used for scrambling its DCI which carries the resource allocation information in LTE [12]. If grouping is changed every sub-frame, a new M-RNTI has to be assigned and conveyed to UEs every sub-frame, leading to increased control overhead. Therefore, grouping policies need to achieve a balance between efficiency and robustness of grouping. Grouping policies also need to answer critical questions like the number of groups to be formed or the maximum number of UEs to be placed in a group. Creating a lesser number of groups means more UEs in a single group which may result in lesser number of PRBs being used. However, as the number of UEs in a group increases, the probability that at least one UE is in deep fade also increases, leading to degraded system performance. On the other hand, a larger number of groups may require more resources. Thus, there is a tradeoff between the number and size of multicast groups which needs to be taken into account while grouping.

1.3.2 Resource Allocation

Resource allocation problem in a multicast system determines which PRB should be allocated to which multicast group. A multicast group comprises several UEs with different channel states and different achievable rates in a PRB. Hence, the rate at which data can be successfully transmitted in a PRB depends on the channel states of all the users in a group. Unlike unicast, there is no unique CQI value for a group. All UEs in a group experience different channel states and hence have different CQI values, but the data in a PRB can only be sent at one rate using a particular MCS. Therefore, the resource allocation policy also needs to define what this rate and MCS of transmission should be. For allocating PRBs, a resource allocation policy needs a representative CQI value for a group in each PRB of a sub-frame. The most common practice is to use the CQI value corresponding to the UE with the least channel gain in the group. This is done to ensure that all users in a group can decode the transmitted content successfully. While choosing a robust CQI ensures that the entire group is served, it leads to degradation in the overall system performance. The weakest UE in a group limits the throughput of other users who may be experiencing much better channel states, leading to user dissatisfaction. A resource allocation policy for multicast transmissions, therefore, needs to balance between robustness and improving system performance parameters such as the overall system throughput.

1.4 Motivation for the Thesis

In this work, we study various aspects of grouping and resource allocation for multicast transmissions. As discussed in Section 1.3.1, it is essential to use appropriate criteria for dividing multicast users into groups. In addition to their content requirements, the channel states of users also need to be taken into account while grouping. Since the channel states of UEs can change every sub-frame, the ideal thing to do would be to group users with similar channels together in every sub-frame. However, doing this every sub-frame would result in a prohibitive amount of additional control overhead. The alternative is to use the knowledge of the average channel state of the users to make grouping decisions that remain viable for a reasonable duration. Knowing the average SNR of users and

the statistical properties of the fading encountered, we can make reasonable predictions about the channel states of users. These predictions can then be used to create a grouping that remains feasible for a long enough period. Thus, grouping policies can be designed based on the average channel states of users that can result in better user satisfaction and overall system performance.

Multicast transmissions are uniquely suited for serving video streaming services which are often requested by a large audience simultaneously. Video streaming could involve real-time streaming for events such as a sport event, live telecasts of award shows, political rallies, a live news feed, or real-time telecast of any other important event. For such services, performance factors like delay and buffering time are much more important than quality of the video i.e., some degradation in video quality may be acceptable if delay and buffering time are kept small. On the other hand, for non-real-time streaming such as the premiere of a movie or a series pilot, the video quality takes precedence over other parameters. The resource allocation policies for these two types of streaming services should satisfy their unique performance requirements. For non-real-time services that prioritize video quality, multicast streaming has to be lossless. All UEs subscribed to such services should receive all video packets to ensure the highest viewing quality of the video stream. Such lossless transmission requires a robust resource allocation scheme that transmits the streaming content at a rate decodable by the weakest user in a multicast group. Therefore, the resource allocation policy is constrained to serve all the users throughout the duration of the session. For efficient resource utilization in such a system, we must ensure that the rate requirements of the multicast users are satisfied in the minimum possible number of PRBs. This problem is of great practical importance because it minimizes the impact of multicast operations on other services that may be simultaneously going on in a cell.

Serving each UE in every sub-frame makes the performance of multicast operations dependent on the weakest users in the system. As a result, if a UE experiences poor channel states for an extended period, the eNB has to continually allocate more PRBs to the corresponding multicast group and transmit content at a lower MCS. This decreases the overall system throughput and impacts the performance of other services in the cell. Moreover, the users with good channel states are constantly forced to settle for lower rates

despite their high CQI values, making them dissatisfied with the services. Under such conditions, users with a good channel may choose to leave the MBMS session and opt for a dedicated unicast stream instead, resulting in even more load on the system resources. The answer to overcoming this bottleneck lies in exploiting some unique properties of video streams. Videos are inherently tolerant to a certain amount of packet loss [13], meaning that the quality perceived by the end user does not undergo any degradation despite the loss of some packets. We can take advantage of this property for designing resource allocation algorithms for real-time video streaming services that allow for some amount of packet loss. The eNB can choose to selectively drop some packets of the weakest users to improve the overall system performance. The system is no longer constrained to serve every user in every sub-frame. Several factors need to be considered while designing a resource allocation policy for such a system. The policy should be able to contain the packet losses of users within certain allowable limits, and it should also ensure that no user starves for long periods. We design such resource allocation policies in this work.

An LTE cell typically caters to several different kinds of services simultaneously. The users in a cell could be using Voice over Internet Protocol (VoIP), downloading files over a unicast link, streaming real-time and non-real-time multicast videos or simply browsing the Internet. Each of these services has a different Quality of Service (QoS) requirement and each service needs a resource allocation algorithm suited to its unique requirements. For instance, even the two forms of multicast streaming require different kinds of allocation algorithms, as discussed above. So, while catering to all these different services at a time, which resource allocation algorithm should we use? Is it even possible to satisfy the varying QoS requirements of all the services using a single algorithm? The existing literature lacks a unified and flexible resource allocation policy that can meet the requirements of different kinds of services. Considering the plethora of services provided by today's cellular networks, there is a need for designing such a unified resource allocation algorithm. We address this requirement by designing a generalized policy that can be used to cater to the requirements of a heterogeneous mix of users and services.

The MBSFN mode of operation in LTE was designed for achieving higher spectral efficiency, especially at the cell edge. However, it has not seen much usage by the cellular operators. One of the reasons for this is the rigidity of operation of MBSFNs. MBSFN

requires all eNBs within an MBSFN area to transmit the same multicast stream over the same resources in strict time synchronization. It also requires using extended Cyclic Prefix (CP) to ensure that the users can combine multiple copies of the content received by them. Using extended CPs leads to a decrease in the system throughput. We propose the use of Multi-Connectivity (MC) with MBMS to overcome these issues. As discussed in Section 1.2, the MBMS stream content is sent to all the eNBs by the MBMS-GW. Since the content is already available in all the cells in an MBSFN area, using MC, a user can independently receive the same stream from multiple eNBs in its vicinity. Like MBSFNs, MC multicast also results in users receiving multiple copies of the same content which can then be combined by them, resulting in improved performance. Besides, without the need for synchronization, eNBs can individually optimize resource allocation in their respective cells and use the most suitable PRBs for transmitting the MBMS content. The resulting frequency diversity increases the probability of the MBMS content being successfully delivered to the users. Using MC multicast in place of MBSFNs also eliminates the control overhead involved in synchronizing multiple eNBs in the MBSFN area. MC has never been studied in the context of multicast transmissions before. Therefore, there is a need for investigating the impact of MC on the performance of multicast transmissions and designing suitable resource allocation algorithms for it. We address both these problems in this thesis.

1.5 Contributions and Organization

In this thesis, we focus on methods for improving video streaming using multicast transmissions and multi-connectivity. The related literature and open issues in this field are discussed in Chapter 2. Chapters 3 through 6 discuss the main contributions of this thesis. The chapter-wise contributions of this work are summarized below:

1. In Chapter 3, we address the problem of grouping and resource allocation in lossless multicast transmissions to minimize the resource utilization of multicast applications. A large portion of multicast literature like [14] and [15] claim that the grouping and resource allocation problems are ‘hard to solve’ or ‘infeasible’. However, none of these papers present any mathematical proof of hardness of these problems.

In this work, we present the proof of NP-hardness of both the optimal grouping as well as the optimal resource allocation problems. Since we prove both these problems to be NP-hard, no polynomial-time algorithms exist for determining their optimal solutions. Therefore, we use the following means to solve these problems:

(a) We devise a simulated annealing based randomized scheme for estimating the optimal resource allocation. The randomized scheme works iteratively to converge to the optimal solution with high probability. The output of this scheme acts as a benchmark for evaluating the performance of heuristic resource allocation schemes.

(b) We propose two heuristic schemes for resource allocation to multicast groups. Through extensive simulations, we show that these schemes provide a significant performance improvement over unicast transmissions. We also propose a heuristic scheme for multicast group formation. We compare the performance of the proposed policies with the existing state of the art through rigorous simulations. Using traces from various video streams [16, 17], we evaluate the performance of the proposed heuristics specifically for streaming video content. The proposed policies succeed in meeting the requirements of video traffic in far fewer PRBs than the existing state of the art.

2. In Chapter 4, we propose a model for a loss tolerant MBMS system. The system allows for a certain fraction of packet losses in MBMS streams being transmitted to multicast UEs as long as the losses stay within a predefined threshold. The loss thresholds of UEs can be a function of the channel quality experienced as well as the video content being streamed by them. Allowing for these losses can help control congestion in the network during peak traffic hours. Our main contributions to resource allocation in loss tolerant video streaming are:

(a) We convert the problem of resource allocation in a loss tolerant MBMS system to a problem of stabilizing a virtual queueing system that models the loss tolerant MBMS system. We prove that stabilizing the token queues in this queueing system ensures that the losses encountered by all UEs stay below their allowed thresholds.

(b) We propose an online Loss Optimal Resource Allocation (LORA) policy for resource allocation in the loss tolerant MBMS network. The loss tolerance of each UE is taken into consideration by this policy. We prove that the policy is throughput optimal. We also propose an online Priority Loss Optimal Resource Allocation (p-LORA) policy that progressively increases the priority (up to a certain limit) of a UE for being scheduled every time it is not served. p-LORA improves upon LORA by ensuring that users are not starved for long periods at a stretch. It regulates the pattern of packet loss in addition to the amount of loss encountered.

(c) The Exponential (Queue length) (EXP-Q) rule [18] is a well-known throughput optimal resource allocation policy that is defined for a single time varying channel shared by multiple flows. We generalize this EXP-Q rule for use in the multi-channel multicast scenario under consideration. Using extensive simulations, we compare the performance of the proposed schemes to that of the EXP-Q rule [18]. Since the proposed loss tolerant system is specifically designed for video streaming, we use traces from actual videos [16,17] to generate realistic data traffic for these simulations.

3. The resource allocation algorithms proposed in Chapters 3 and 4 are aimed at optimizing specific objective functions. However, present day networks cater to a heterogeneous mix of services and user devices with different QoS requirements. This presents the need for a generalized allocation algorithm that can be used irrespective of the performance parameters being optimized and the types of traffic being served. We design such an auction based algorithm in Chapter 5. The proposed algorithm succeeds in meeting the requirements of a heterogeneous mix of users without any prior knowledge of their QoS requirements. The allocation decisions made by the proposed algorithm are only based on the bids submitted by the users. We prove that the algorithm is strategy-proof which means that the users have no incentive in lying about their actual valuations for being scheduled. This ensures the maximization of social welfare under the proposed policy.

4. In Chapter 6, we investigate the use of multi-connectivity for multicast transmis-

sions. We prove that the optimal resource allocation problem in MC multicast is an NP-hard problem. Therefore, we propose a greedy approximation algorithm and prove that it provides an approximation ratio of $(1 - 1/e)$. We also prove that this is the best approximation ratio possible for the given problem. Through extensive simulations, we show that the use of MC provides huge performance benefits over a single connected system. Since multi-connectivity is not defined to be used with multicast in the current 3GPP standards, we also propose procedures and signaling exchanges needed for establishing multi-connectivity in MBMS systems.

Chapter 7 concludes the thesis along with a discussion on possible directions for further research.

Chapter 2

Multicast Streaming: Literature and Open Problems

Wireless multicast provides an excellent means of efficiently serving the ever-increasing video streaming traffic over cellular mobile networks. Using multicast enables the network to serve any number of users in resources that would be needed by a single user with unicast transmissions. As discussed in the previous chapter, there is a requirement for efficient grouping and resource allocation algorithms that can help integrate multicast seamlessly with the existing cellular mobile operations. Developing such algorithms and solving the optimal resource allocation problem for various forms of multicast transmission is the primary focus of this thesis. Towards this end, we first discuss the existing literature related to multicast group formation and resource allocation in Sections 2.1 and 2.2 respectively. As outlined in Chapter 1, the inherent loss tolerant nature of video streams can play a pivotal role in changing the way videos are streamed online. Utilizing this property for selectively dropping some packets can combat network congestion due to excessive mobile video streaming. To this end, we design resource allocation algorithms for loss tolerant video streaming. Some limited literature is available in this area which we discuss in Section 2.3. While multi-connectivity is considered a key enabler for unicast communications in LTE and 5G, its use in multicast has not been adequately considered. Use of multi-connectivity with multicast has a potential to significantly improve the spectral efficiency of multicast transmissions and provide advantages of MBSFN transmissions without their stringent synchronization requirements. We discuss some literature relevant to this area in Section 2.5.

2.1 Multicast Group Formation

In the existing framework for MBMS, all users that subscribe to the same MBMS stream are treated as a single group. However, as discussed in the previous chapter, this can lead to degraded system performance and user dissatisfaction. Dividing the users into smaller groups can help tackle these problems. In [19], the authors deal with the grouping problem for MBMS in High Speed Packet Access (HSPA) networks. They propose a grouping policy that minimizes a 'global dissatisfaction index'. This global dissatisfaction index accounts for the difference in the maximum data rates achievable by UEs and the rates assigned to them. The authors show, using simulations, that their proposed policy performs better in terms of UE satisfaction compared to MBMS transmission without grouping. In [20], the same authors investigate the effect of pedestrian mobility on the performance of the grouping policy proposed in [19]. The authors conclude that pedestrian mobility does not impact the performance of the proposed grouping policy.

In [21], the authors propose subgrouping and resource allocation for multicast in LTE Advanced (LTE-A) systems. Extensions of LTE multicast subgroup formation are presented for use in LTE-A systems. They propose a radio resource management scheme that achieves a trade-off between efficiency and fairness. For resource allocation, they make use of the bargaining solutions proposed in [22]. It is shown that, due to carrier aggregation in LTE-A systems, the overall system throughput is significantly increased, but the relative performance of the algorithms investigated remains the same. Extensions of bargaining solutions proposed in [22] to multi-carrier systems like LTE-A have been studied in [23]. In [24], the authors have extended the work from [22] to exploit frequency selectivity for improving the spectral efficiency of multicast in LTE. In [25], the authors explore the use of multicast in heterogeneous networks. The grouping cum resource allocation problem in [25] aims at maximizing the system throughput while meeting the rate of all users. In [26], users are divided into groups such that the system capacity is maximized while optimizing different cost functions. The cost functions used are system throughput, fairness, and user satisfaction.

Most of the papers mentioned in this section assume that the entire set of PRBs can be used for catering to multicast transmissions. In practice, however, an eNB has to support multiple other services alongside multicast sessions. Therefore, we formulate the

grouping problem in this work with the aim of satisfying the rate requirements multicast UEs in the minimum possible number of PRBs.

2.2 Resource Allocation in Multicast

The existing research on resource allocation for multicast transmissions can be classified into several categories based on the objective of the allocation problem, the techniques used for allocation, or the types of videos being streamed. Therefore, we study the relevant literature in this section under the following five categories.

2.2.1 Opportunistic Multicast Scheduling

Opportunistic scheduling schemes, as the name suggests, are throughput maximizing schemes that schedule UEs with the best channel states in a sub-frame. In [27], the authors present an optimized version of Opportunistic Multicast Scheduling (OMS) that achieves a balance between multicast gain and multi-user diversity. A fraction of UEs with the best channel gains are scheduled in each time slot. Use of opportunistic multicasting for Single Frequency Networks (SFNs) has been studied in [28]. The authors focus on maximizing the spectral efficiency of the SFNs by opportunistically scheduling UEs who report higher Channel Quality Indicator (CQI) values.

In [29], the authors propose a Frequency Domain Packet Scheduler (FDPS) for MBMS that maximizes the minimum rate achievable by UEs in a PRB. It uses a conservative approach in that it only minimizes the performance loss caused by the worst PRB assignment. Moreover, the performance of the proposed policy has only been compared to a blind FDPS policy that uses a static allocation that doesn't change over time which is not a good benchmark for comparison.

In [14], the use of a genetic algorithm for resource allocation is proposed for OFDMA multicast followed by power allocation based on the technique proposed in [30]. The resource allocation problem in [14] aims to maximize the total throughput subject to power and fairness constraints. The authors, however, do not subgroup the UEs based on their channel states. All UEs receiving the same content are put into a single multicast group. Recently, there has also been some work on multicast transmissions in 5G satellite

systems. In [31] and [32], the authors propose solutions for radio resource management and subgrouping for multicast over 5G satellite systems. The optimization problems formulated seek to maximize the aggregate data rate of the system. These papers also assume a single CQI value corresponding to a multicast UE which means that all PRBs in a sub-frame are equivalent for a UE. Maximizing the aggregate data rate is also the objective function of [22] in which game-theoretic bargaining solutions are used for grouping and resource allocation of multicast UEs.

Most of the literature considers only wideband CQI (i.e., a single CQI value for the entire available bandwidth) for grouping and resource allocation in multicast transmission. One work that explores the use of subband CQI values in multicast resource allocation is [33]. The objective function here is still the maximization of the aggregate data rate as in [34], [31] and [32]. However, with the consideration of different subband CQI values, a closed-form solution for subgroup formation as given in [34] no longer remains feasible. While all the papers mentioned in this section seek to maximize the aggregate data rate in some way, in this work, we focus on providing a certain rate to each multicast UE based on the service that it is subscribed to. While allocating resources, we also take into consideration, the variation of CQIs of UEs across different PRBs in a sub-frame.

2.2.2 Joint Optimization for Unicast and Multicast

This section summarizes the literature that deals with the problems of joint resource allocation to unicast and multicast UEs. In [35] and [15] joint delivery of unicast and multicast/broadcast transmission in LTE and OFDMA systems has been addressed. Policies proposed in [15] guarantee a certain rate to all the multicast UEs and make use of unicast transmission for serving UEs with the worst CQI values. In [15], the broadcast services are also made available through unicast channels to minimize the probability of users being in an outage. In [35], the performance of streaming over MBSFNs and file delivery over evolved MBMS (eMBMS) has been evaluated through simulations. Performance indicators like outage probability, coverage and maximum supportable MCS have been used to asses the feasibility of various MBMS configurations from the perspective of the service providers.

In [36], authors deal with fair and optimal resource allocation in eMBMS. It is

assumed that the video content is simultaneously available through unicast as well as eMBMS, and the primary problem seeks to jointly optimize over the grouping of UEs and allocation of resources to unicast and eMBMS. The resource allocation scheme proposed in the paper allocates resources to groups proportional to the number of UEs in the group. However, while allocating resources, the varying channel states of UEs over different PRBs have not been considered. In [37], authors consider the problem of determining throughput maximizing resource allocation for an MBSFN area. They propose a joint multicast/unicast allocation scheme that maximizes the total throughput while guaranteeing a certain bit-rate to all the users.

None of these papers consider the varying channel states of UEs over different PRBs while allocating resources. In all the problems addressed in this work, however, we account for the fact that the CQIs of UEs may vary across sub-frames and also across PRBs of a sub-frame. Due to these channel variations, all PRBs are not equivalent for a UE. This makes the resource allocation problems considerably harder.

2.2.3 Multicasting of DASH and SVC Content

Dynamic Adaptive Streaming over HTTP (DASH) [38] is a streaming technique that stores several different encoded bit-rates of a video and UEs are given a specific bit-rate based on their channel states. In [39], the authors have used convex optimization to obtain an optimal solution for multicasting DASH and Scalable Video Coding (SVC) streaming content over LTE. The problem optimizes the resource allocation, the MCS, and the Forward Error Correction (FEC) code rates used. File Delivery over Unidirectional Transport (FLUTE) [40] protocol has been used for sending DASH content over an eMBMS system. UEs are grouped based on their distances from the eNB. UEs closer to the eNB receive better quality videos than the ones farther away from it.

In [41], the authors use a pricing based scheme for allocating resources to multicast groups streaming SVC video content. Users are divided into three multicast groups based on the price they pay. UEs that pay the most receive the maximum number of enhancement layers, and the base layer is provided to all. Allocation of resources is done based on a multicast transmission score that is a function of the CQI values, past throughput, the number of UEs in a group, and the price paid by that group. In [42], the authors

investigate the use of Random Network Linear Coding (RNLC) for improving the performance of multicast services. They use two different forms of RNLC for multicasting H.264/SVC videos in a generic cellular system. The resource allocation problem formulated by the authors aims at minimizing the number of coded packets required to be transmitted for successfully delivering all the layers of the SVC video streams and providing the required QoS guarantees to at least a certain fraction of the users. Since the resource allocation optimization problems are hard to solve, the authors have provided efficient heuristics that provide solutions close to the optimal. The performance of the proposed schemes has been studied on an LTE single-cell eMBMS system. The authors in [43] deal with optimizing the delivery of network coded scalable video content using eMBMS. They have made use of Unequal Error Protection (UEP) for ensuring the reliability of multilayer video transmissions. They propose a UEP resource allocation model that maximizes the profit to cost ratio of the system. The system profit is defined by the number of video layers that the UEs can recover with a given probability, and the cost captures the number of transport blocks used in transmitting the video content. It is shown that the proposed UEP resource allocation model provides much better coverage than conventional multi-rate transmission [44].

Problem of resource allocation for MBMS Operation On-Demand for SVC video streams has been studied in [45] and [46]. The authors propose resource allocation schemes that maximize Quality of Experience (QoE) instead of QoS. Power-efficient streaming of high-quality SVC encoded videos via MBMS has been examined in [47]. The UEs are grouped based on the content, the quality of content requested, and their physical proximity. The algorithms proposed in [47] minimize the power consumption by sending traffic in intermittent bursts, allowing UEs to sleep in between bursts. In [48], authors make use of a multi-criteria decision-making tool called Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [49] for grouping and resource allocation in SVC multicast video streaming. Grouping and resource allocation decisions are based on three criteria, maximizing the throughput, maintaining fairness, and minimizing the dissatisfaction of groups. TOPSIS has also been used in [50] for comparing the performance of various multicast resource allocation schemes based on their aggregate data rate, fairness, and spectral efficiency.

Even though SVC provides an exciting new method of video encoding with various benefits, H.264/AVC [51] continues to be the preferred method of encoding videos over the Internet. Most of the popular streaming platforms like Netflix [52] and YouTube [53] use non-layered coding formats like H.264/AVC and VP9 to encode their videos. Therefore, in this thesis, we focus on multicast streaming of non-layered videos. As we shall discuss in Chapter 3, the algorithms proposed by us can also be used for transmitting the base layer while streaming layered videos. Enhancement layers can then be opportunistically transmitted to users with good channels [54, 55].

2.2.4 Auction Based Multicast Scheduling and Pricing

Auction mechanisms and other game-theoretic tools are being increasingly used for addressing various allocation problems in communication networks [56–64]. Auctions have also been used in the literature for spectrum allocation [65], channel allocation in vehicular networks [66] and resource allocation in Device-to-Device (D2D) multicast [67]. In [63], the authors study optimal pricing for SVC multicasting systems with stochastic user arrivals. The interaction between subscribers and service providers is modeled as a game. Multi-dimensional Markov Decision Process (MDP) has been used to model the behavior of the users and determine their equilibrium actions. Pricing response of the service providers is then determined to maximize their revenue. In [23], the authors propose a game-theoretic bargaining solution for multicast resource allocation in multi-carrier systems like LTE-A. The proposed technique seeks to balance fairness among users with system efficiency.

In [56, 57], the authors have proposed multi-dimensional auction mechanisms for crowdsourcing in an adaptive bitrate [68] video streaming framework. Crowdsourcing enables users to form cooperative groups to share their resources and download video fragments for each other. The auction mechanisms determine which segment a user should download and at what bitrate and how the user should be compensated for downloading a fragment for other users. In [59], the authors propose an auction-based subcarrier allocation for SVC video transmission to maximize the net revenue gained by the system. Allocation for SVC video transmission in 4G WiMAX has been discussed in [60] and [61] using Vickrey-Clarke-Groves (VCG) auction mechanism. In [62], a social utility

maximizing mechanism has been proposed for multi-rate multicast over the Internet. It is assumed that the valuations of players are known to each other but are not known to the central allocating entity.

All these auction based algorithms are designed to optimize a specific system parameter. Such algorithms cannot cater to a variety of services having different QoS requirements. The auction based algorithm proposed by us in Chapter 5 can be used for optimizing any system parameter. It can, therefore, serve a heterogeneous mix of users and services simultaneously.

2.2.5 Broadcasting

In [69], the authors propose a scheduling scheme for eMBMS broadcast that is focused on reducing the average latency of broadcast services in the system. The broadcast content is divided into two categories based on popularity. The proposed scheme starts transmission in unicast mode and gradually moves to the broadcast mode as the number of UEs increases. The less popular content queue is served using a stretch scheduler [70] while the more popular content queue is served using a round-robin scheduler. The policy takes into account the impact of UE impatience reflected in departure and request repetition. In [71], the authors deal with efficient broadcasting in LTE eMBMS. The problem of broadcasting over Single Frequency Networks (SFN) is divided into two sub-problems. The first sub-problem determines the optimal SFN configuration for all the broadcast sessions active in the involved eNBs. The problem seeks to achieve a balance between the diversity gain obtained by creating a large SFN and the multiplexing gains that can be obtained by splitting the area into smaller SFNs. The second sub-problem deals with resource allocation to the broadcast and unicast sessions in the cells. The SFN configuration problem has been solved using heuristics for the minimum p-cut problem [72] and the solution for the resource allocation problem uses a water filling form of the proportional fair scheduling [73,74]. The proposed broadcasting mechanism has been given the name of Broadcast over LTE (BoLTE). The authors have evaluated the performance of BoLTE using a WiMAX testbed.

We propose the use of multi-connectivity multicast as a replacement for MBSFN operations. MBSFNs require strict synchronization between eNBs in an MBSFN area and

an extended cyclic prefix. The need for strict synchronization leads to significant control overheads and the extended cyclic prefixes lead to reduction in the system throughput. The proposed multi-connectivity multicast overcomes these issues while providing better performance than MBSFNs.

2.3 Loss Tolerant Multicast Streaming

Owing to the interdependence of frames, video streams can tolerate some packet loss without significantly affecting the quality perceived by the users. Present day decoders can conceal as much as 40% packet loss in videos [13]. In the existing literature, no resource allocation policy takes advantage of this unique loss tolerant nature of video streams for optimizing resource allocation for multicast services. However, various forms of source coding have been developed that make the video streams more resilient to losses [75–77]. In [76, 77], the authors design variations of Distributed Source Coding (DSC) frames for reducing error propagation, facilitating view switching and minimizing the effect of packet loss in Interactive Multiview Video Streaming (IMVS). IMVS enables users to switch between different views of a video stream. The authors design Drift Elimination DSC (DE-DSC) frames that halt error propagation within the video stream due to the dependence of frames on their predecessors. They also design uDSC frames that facilitate view switching in addition to preventing the propagation of errors in IMVS. Once the IMVS streams are encoded, the packetization and ordering of frames are done to maximize the expected number of correctly decoded frames at the receiver.

The existing literature does not take advantage of the loss resilience of video streams to improve resource utilization in multicast video streaming systems. We leverage the loss tolerance of video streams in this work to design loss optimal resource allocation policies for multicast streaming systems. These policies allow for some controlled packet losses while providing users with the required video quality.

2.4 Network Coding for Multicast

The use of network coding in multicast has been studied in [78–88]. In [78], the authors propose an online coding and queue update algorithm known as *drop-when-seen* that

drops packets at the sender queue once all users have seen the packet. They use acknowledgments on degrees of freedom instead of the actual decoding of packets. The resulting algorithm provides significantly lower queue sizes than *drop-when-decoded* algorithms. A feedback based adaptive network coding has been proposed in [79, 86] to minimize decoding and delivery delays in each transmission for applications that require in order delivery of packets. The authors show that by taking channel memory into account in network coding decoding can lead to considerably lesser decoding delays.

A distributed random linear network coding for transmission and compression in multicast has been proposed in [80]. The coding approach asymptotically achieves the capacity of multicast networks with network coding given in [81]. Throughput-smoothness trade-offs in multicast streaming applications have been studied in [82] and [83]. The authors consider streaming content that requires strict in order delivery, over an erasure channel. The authors propose a variety of coding schemes that can be tuned to operate at a point that achieves a suitable trade-off between the two parameters. The problem of minimizing the playback delay in streaming over an erasure channel has been studied in [84]. The authors analyze the expected playback delay with and without feedback and show that they achieve the same asymptotic value as the bandwidth approaches infinity. The authors in [87] study codes that can be tuned to obtain a trade-off between correcting burst and isolated erasures. They propose a new class of codes called embedded-random linear codes that achieve a balance between correcting these two types of erasures. The extension of low delay erasure correction codes to multicast streaming has been studied in [88]. They propose a new set of codes for the low delay regime that are shown to be optimal in a subset of the regime.

2.5 Multi-Connectivity

In addition to multicast, we can make use of multi-connectivity for further enhancing video streaming in LTE and 5G. Multi-connectivity allows devices to receive content from multiple sources and over multiple Radio Access Technologies (RATs) simultaneously. In the current state of literature and standards, the only form of multi-connectivity that exists is Dual Connectivity (DC). DC capable devices can connect to at most two base

stations at a time. In LTE, DC allows a user to connect to a primary macro base station and a secondary micro or femto base station [89]. In 5G, DC refers to a UE being connected to a primary LTE eNB and a secondary gNodeB (gNB) or vice versa [90]. DC is expected to be a key enabler in 5G wireless networks [91]. The high data rate, ultra-reliable low latency, and high mobility requirements of 5G necessitate the reduction of radio link failures due to mobility. Use of DC makes it possible to avoid such failures and ensure seamless connectivity for mobile users [92].

Even though the use of DC has been extensively studied by the research community in the past few years for throughput and handover improvement [89, 93–95], its use in multicast transmissions has not received much attention. We investigate the use of multi-connectivity in multicast transmissions in this work. Our work reveals that the use of multi-connectivity significantly improves the performance of multicast transmissions.

2.6 Open Research Problems

In this chapter, we have discussed some of the existing perspectives on grouping and resource allocation for multicast transmissions in cellular mobile networks. The resource allocation policies in the literature focus on maximizing the system throughput, maintaining fairness, or minimizing the dissatisfaction of users. We have also discussed various methods proposed in the literature for halting error propagation and loss mitigation in video streams. While methods like source coding have been previously investigated for curtailing the effects of packet loss in video streaming, we leverage the inherent loss tolerance of videos to reduce their resource consumption without compromising on the quality of streaming. In this thesis, we look to minimize the number of resources used for serving the multicast streams while meeting their respective QoS requirements. We also investigate the use of multi-connectivity with multicast transmissions and show that it can have enormous benefits for improving the quality of multicast operations.

The current literature does not address the problem of satisfying the rate requirement of all multicast UEs while minimizing the PRB utilization of multicast transmissions. This is a critical problem because the practical success of multicast services strongly depends on how well they can co-exist with the other vast number of services supported

by LTE and the next-generation 5G networks [96]. While MBMS services are ideal for real-time streaming applications, their resource utilization has to be such that sufficient resources are available for the non-real-time applications being simultaneously provided in the networks. Since the resources used for providing an MBMS service are essentially disseminating common content to the MBMS UEs, over-provisioning of resources for multicast must be avoided. The multicast UEs could simultaneously be using unicast services along with other UEs in the cell, which may not be involved in the ongoing MBMS sessions. Minimizing the resources used by the MBMS services ensures that the impact of multicast services on the rest of the operations in the network is minimized.

A common shortcoming in the existing resource allocation literature is that the channel states of the UEs are often assumed to be the same in all subcarriers/PRBs. This makes the identity of the PRBs irrelevant, and the problem reduces to determining the number of resources to be allocated to a group. In practice, however, the channel gain of a UE can be different in each PRB of a sub-frame which makes the resource allocation problem considerably harder. Throughout this work, we take the channel variations of UEs over different PRBs in a sub-frame into consideration. Therefore, the resource allocation policies designed specify the identities of the PRBs and not just the number of PRBs to be assigned to a group/user.

None of the resource allocation algorithms in the literature employ the loss tolerance of videos for optimizing the allocation of resources to multicast video streams. In Chapter 4, we design throughput optimal resource allocation policies for video streaming services that use this unique property to optimize PRB allocation and improve the overall system performance. The current literature also lacks in generalized allocation algorithms that can be effectively used irrespective of the nature of services and allocation objectives. The algorithms in the existing literature are built around a certain objective function such as maximizing the system throughput [14, 37], ensuring fairness [23], or maximizing revenue [59]. No single algorithm can be used for any objective function and any range of services that might have completely different service requirements. In this work, we design such a generalized auction-based algorithm in Chapter 5. The proposed algorithm can be used to serve a heterogeneous mix of users and streaming services for optimizing multiple objectives.

The use of multi-connectivity for multicast transmissions also remains unexplored in the current standards and literature. We discuss in Chapter 6 how multi-connectivity has the potential to enhance the performance of wireless multicast greatly. Multi-connectivity multicast also provides a simplified and flexible alternative to the concept of single frequency networks in MBMS. We assess the impact of multi-connectivity on the performance of multicast, define procedures for establishing multi-connectivity in LTE multicast, and discuss the corresponding enhancements required in the existing 3GPP standards.

We begin by first examining the problem of minimizing the resources used by multicast services while meeting their specific rate requirements in the next chapter. We construct the grouping and the resource allocation problems with this objective and propose various efficient algorithms for obtaining their solutions.

Chapter 3

Grouping and Resource Allocation for Lossless Multicast Transmissions

In this chapter, we investigate the problems of optimal grouping and optimal resource allocation in a lossless eMBMS system. We consider a constant rate model that requires a certain rate (and hence a certain QoS) to be provided to all the users receiving multicast services in minimum possible number of PRBs. Minimizing the number of PRBs used by the multicast services is extremely important from a practical standpoint because the success of multicast operations strongly depends on how well they can co-exist with the multitude of other services supported by the current LTE and next generation 5G networks [96]. While eMBMS services are suitable for streaming, their resource utilization has to be such that sufficient resources are available for non real time applications being simultaneously provided in the cells. Minimizing the resources used by eMBMS services ensures that their impact on other services is minimized. The optimal resource allocation problem that minimizes the number of PRBs used subject to satisfying the rate requirements of all the users is a Binary Linear Program (BLP). BLPs are inherently hard to solve and require significant computational power even for small input sizes. The optimal resource allocation problem is in fact an NP-hard [97] problem. The optimal grouping problem aims to divide the users into groups such that, for any resource allocation policy, the average number of PRBs saved per sub-frame is maximized. As discussed in Chapter 1, there is a trade-off involved between creating more groups and increasing the number of users contained in a group. The grouping policy needs to strike a balance

between these factors. The optimal grouping problem is also NP-hard.

In most of the existing literature, the rate achievable by a UE is assumed to be the same in all PRBs of a sub-frame. This assumption greatly simplifies the resource allocation problem as the identities of the PRBs are no longer significant. In practice, however, the channel quality experienced by a UE is different for different frequency channels resulting in varying channel gains over different PRBs. In this work, we take these channel variations into account. Also, a large portion of the literature including [14] and [15] claims that the multicast grouping and resource allocation problems are 'hard to solve' or 'infeasible'. However, none of these provide any mathematical proof of hardness of these problems. In this work we provide proofs of NP-hardness of both these problems. Since both the problems we are trying to address here are NP-hard, we cannot determine their optimal solutions in polynomial-time unless P=NP [97]. Therefore, we have to rely on randomized searches and efficient heuristics for obtaining solutions close to the optimal. We devise a randomized scheme for estimating the optimal resource allocation. It uses Simulated Annealing (SA) to converge to the optimal solution with high probability. The randomized scheme works in an iterative manner, exploring a large number of possible allocations to converge to the optimal solution. As a result, it takes a long time to run and is unsuitable for practical use. However, the output of this scheme provides a benchmark for evaluating the performance of heuristic resource allocation schemes. We design two efficient online heuristic resource allocation policies for use in practical systems. These policies are specially suitable for streaming services. Our policies ensure that all the users receive the video streams at a constant rate. The rate of transmission required by a steaming service depends on the kind and quality of the video being streamed. In order to see a consistent quality of streaming, the users must be served at a certain fixed rate. The resource allocation policies proposed in this chapter achieve this objective. We also design a two stage grouping policy that makes use of the knowledge of the average Signal to Noise Ratio (SNR) of UEs to divide them into multicast groups. The performance of the proposed grouping and resource allocation policies is compared with existing state of the art through extensive simulations. Our simulation results clearly indicate the superiority of the proposed policies for use in multicast transmission.

The rest of this chapter is organized as follows. In Section 3.1, we discuss the system

model used. The optimal resource allocation and grouping problems are formally stated and proved to be NP-hard in Section 3.2. The SA based randomized scheme and related results are presented in Section 3.3. In Sections 3.4 and 3.5, we present the proposed heuristic schemes for resource allocation and grouping respectively. The simulation results are presented and analyzed in Section 3.6. We then discuss various generalizations of the proposed policies in Section 3.7 and conclude in Section 3.8.

3.1 System Model

We consider an LTE cell with M UEs. All UEs are subscribed to the same eMBMS service and are required to be served at a rate of R bits/sec. The required rate can be provided to each UE by allotting one or more PRBs to its group in each sub-frame. We denote the number of PRBs in a sub-frame by N. Let $[n] = \{1, \ldots, n\}$ and let $|A|$ denote the cardinality of a set A. Thus, $[M]$ and $[N]$ denote the set of multicast UEs and the set of PRBs in a sub-frame, respectively. We assume that the channels between eNB and the UEs are location and time varying. Thus, each UE has different channel gains in different PRBs and also across different sub-frames. We assume block fading channel model, and hence the channel gain of a UE is assumed to remain the same during a sub-frame. Though we do not consider mobility explicitly, our approach can be extended to cases where UE positions evolve at a slower time scale than the sub-frame duration. Let $h_{iu}[t]$ denote the channel gain for UE u on i^{th} PRB in sub-frame t. $h_{iu}[t] = \overline{h}_{iu} + H_{iu}[t]$, is made up of 2 components. $\overline{h}_{iu}$ denotes the average channel gain which accounts for path loss and shadowing and is invariant across sub-frames. $H_{iu}[t]$ is the fast-fading component that varies across sub-frames. $H_{iu}[t]$'s are independent and identically distributed (i.i.d) exponential random variables.

We assume that the eNB has full knowledge of the Channel State Information (CSI) of all the UEs. This not a restricting assumption in the current state of the LTE systems where CQI can be periodically fed back to the eNB by the users [7]. Corresponding to the channel gain, the eNB assigns the maximum supportable rate, $r_{iu}[t]$ bits/sec for UE u on i^{th} PRB in sub-frame t. Note that $r_{iu}[t]$ is determined by the Modulation and Coding Scheme (MCS) used, and thus can take finitely many values (15 as per current standards

for LTE [7]). Next, we discuss grouping.

Since all multicast UEs want the same content in each sub-frame, the UEs can be grouped together and served on common PRBs. We denote the number of groups formed by L. A grouping strategy Δ is defined as follows:

Definition 1. *A grouping strategy Δ, defines a partition $\{G_1^\Delta, \ldots, G_L^\Delta\}$ of $[M]$, where $G_i^\Delta \subseteq [M]$ is referred to as the i^{th} group.*

Note that $L \leq M$. For $L = M$, we have the unicast case. Henceforth, unicast is not dealt with separately. Throughout this chapter, we assume that groups once defined at the beginning of an eMBMS session cannot be changed during the session. This is done to avoid excessive control overhead that may result due to rapid changes in grouping. One can relax this assumption and allow for grouping to be potentially changed every K sub-frames, where K is large. This will allow the scheme to adapt in case of mobile networks. The minimum supportable rate for a group G_j on i^{th} PRB in sub-frame t ($r_{ij}^\Delta[t]$) is equal to the minimum of the rates achievable by its constituent members, i.e., $r_{ij}^\Delta[t] = \min_{u \in G_i^\Delta}\{r_{iu}[t]\}$. This ensures that the content received by a group can be successfully decoded by all its members. If we transmit content to a group at rates higher than this, the weakest UE in the group will not be able to decode the received content successfully. Once $r_{ij}^\Delta[t]$'s are obtained, we need to decide how to allot resources to each group so that the total number of PRBs used is minimized subject to giving each group at least the minimum required rate R. This is a resource allocation problem. The formal definition of a resource allocation policy is stated below.

Definition 2. *For a given grouping Δ a resource allocation policy Γ defines an assignment of PRBs to the L multicast groups, $\{\overline{V}_{1\Gamma}^\Delta, \ldots, \overline{V}_{L\Gamma}^\Delta\}$, where, $\overline{V}_{i\Gamma}^\Delta$ is the set of PRBs assigned to group i by resource allocation policy Γ under grouping Δ. The allocation Γ should be such that $\overline{V}_{i\Gamma}^\Delta \cap \overline{V}_{j\Gamma}^\Delta = \phi$ whenever $i \neq j$ and $\bigcup_{i=1}^L \overline{V}_{i\Gamma}^\Delta \subseteq [N]$.*

Resource allocation policy Γ is said to be feasible if $\sum_{j \in \overline{V}_{i\Gamma}}^\Delta r_{ij}^\Delta[t] \geq R$ for every $i \in [L]$. The other parameter used by us to characterize a resource allocation policy is the number of PRBs left unused after resource allocation in a sub-frame t, $S_\Gamma^\Delta[t] = N - |\bigcup_{i=1}^L \overline{V}_{i\Gamma}^\Delta|$. We shall now formally state our resource allocation and grouping problems.

3.2 Problem Definition

3.2.1 Problem 1: Optimal Resource Allocation $\mathbf{B}^{\star}_{\mathbf{\Delta}}$

Consider a fixed grouping policy Δ, and define indicators in sub-frame t as follows:

$$x_{ij}[t] = \begin{cases} 1, & \text{if PRB } j \text{ is assigned to group } i \\ 0, & \text{otherwise.} \end{cases}$$

The optimal resource allocation can then be obtained as a solution to the following BLP for every t:

$$\begin{aligned} (\mathbf{B}^{\star}_{\mathbf{\Delta}}): \quad & \min \sum_{\mathbf{j}\in[\mathbf{N}]} \sum_{\mathbf{i}\in[\mathbf{L}]} \mathbf{x_{ij}}[\mathbf{t}], \\ \text{subject to:} \quad & \sum_{j\in[N]} x_{ij}[t] r^{\Delta}_{ij}[t] \;\geq\; R, \quad \forall\, i \in [L], \qquad (3.1) \\ & \sum_{i\in[L]} x_{ij}[t] \;\leq\; 1, \quad \forall\, j \in [N]. \qquad (3.2) \end{aligned}$$

The objective function of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$seeks to minimize the number of PRBs used in sub-frame t. Constraint (3.1) guarantees that the rate given to each group is at least equal to the required rate R and (3.2) ensures that each PRB is given to at most one group. Note that $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ gives the optimal resource allocation for any grouping Δ. Next, we establish the hardness of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$.

3.2.1.1 $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ is NP-hard

Since $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ is an optimization problem, in order to prove that it is NP-hard, we must show the corresponding decision problem to be NP-complete. The decision problem corresponding to $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ (denoted by $\mathbf{B}^{\star}_{\mathbf{D}}$) is defined as follows:

$\mathbf{B}^{\star}_{\mathbf{D}}$: Does there exist an assignment of binary variables $\{x_{ij}\}_{i,j}$, $i \in [L]$ and $j \in [N]$ such that (3.1) and (3.2) of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ are satisfied?

$\mathbf{B}^{\star}_{\mathbf{D}}$ determines whether or not there exists a feasible solution of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$. In order to prove that $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ is an NP-hard problem, it is sufficient to show that $\mathbf{B}^{\star}_{\mathbf{D}}$ is NP-complete. We prove the NP-completeness of $\mathbf{B}^{\star}_{\mathbf{D}}$ by reduction from a version of the 3-partition problem (3P) defined below [98]:

- **Input:** A set Y, of $P = 3m$ positive integers, $\{\rho_1, \rho_2, \ldots, \rho_P\}$ such that $\frac{B}{4} < \rho_j < \frac{B}{2}$ for every $\rho_j \in$ Y and $\sum_{j=1}^{P} \rho_j = mB$.
- **Problem:** Can we obtain a disjoint partition of Y, $\{Y_1, Y_2, \ldots, Y_m\}$ such that $\sum_{\rho_k \in Y_i} \rho_k = B$ and $|Y_i| = 3$ for every $Y_i, i \in \{1, 2, \ldots, m\}$ and $\bigcup_{i=1}^{m} Y_i =$ Y?
- **Output:** If the problem is feasible, the output is a suitable partition of Y, else, the output states that the problem is infeasible.

The 3P problem is known to be NP-complete [98]. We now show the NP-completeness of $\mathbf{B^{\star}_{D}}$ by reduction from 3P.

Algorithm 1: Pseudo-code for reducing 3P to $\mathbf{B^{\star}_{D}}$

Input: 3-partition problem with set Y, of $P = 3m$ positive integers, $\{\rho_1, \rho_2, \ldots, \rho_P\}$ such that $\frac{B}{4} < \rho_j < \frac{B}{2}\ \forall\ \rho_j \in$ Y and $\sum_{j=1}^{P} \rho_j = mB$

Output: An instance of $\mathbf{B^{\star}_{D}}$with

1 $L \leftarrow m$

2 $N \leftarrow P$

3 $R \leftarrow B$

4 $r_{ij} = r_j \leftarrow \rho_k\ \forall\ j \in \{1, 2, \ldots, P\}\ ,\ i \in \{1, 2, \ldots, m\}$

Theorem 1. $\mathbf{B^{\star}_{D}}$ *is an NP-complete problem.*

Proof. In order to prove that $\mathbf{B^{\star}_{D}}$ is NP-complete, we first need to show that $\mathbf{B^{\star}_{D}}$ belongs to the class NP. Given a certificate for $\mathbf{B^{\star}_{D}}$, we can verify in polynomial-time whether or not it is a solution by checking if it satisfies the requirements stated in constraints (3.1) and (3.2) of $\mathbf{B^{\star}_{\Delta}}$. This can be done in $\mathcal{O}(LN)$ computations. Therefore, $\mathbf{B^{\star}_{D}} \in$ NP.

Having proved that $\mathbf{B^{\star}_{D}} \in$ NP, we now need to reduce 3P to an instance of $\mathbf{B^{\star}_{D}}$ in polynomial-time. The pseudo-code for the algorithm used for the said reduction is presented in Algorithm 1. For the purpose of this reduction, we assume that all the UEs experience the same channel conditions in a particular PRB of a sub-frame, i.e. $r_{ij} = r_j$ for every $i \in [L]$. We put L (number of groups) $= m$, N (number of PRBs) $= P$, R (rate requirement of UEs per sub-frame duration) $= R$ and r_k (rates of the groups corresponding to different PRBs of a sub-frame) $= \rho_k$ for every $k \in \{1, 2, \ldots, P\}$.

Note that, to define an instance of $\mathbf{B^{\star}_{D}}$, we need to define the number of groups, number of available PRBs, rate requirement of groups (R) and the rates that can be achieved by the groups in every PRB. These are defined in lines 1 through 4 of Algorithm 1 respectively. The reduction in Algorithm 1 can be accomplished in $\mathcal{O}(N)$ computations.

We now show that a solution for $\mathbf{B^{\star}_{\Delta}}$ gives us a solution for 3P as well. Assume that there exists a polynomial-time algorithm for solving $\mathbf{B^{\star}_{\Delta}}$. If we try to solve $\mathbf{B^{\star}_{\Delta}}$ using this algorithm, it will either give us a feasible solution or tell us that $\mathbf{B^{\star}_{\Delta}}$ is infeasible. We will now show how each of these outputs can be mapped to a corresponding solution for 3P.

Say that the algorithm gives us a feasible solution for $\mathbf{B^{\star}_{\Delta}}$. Let the feasible solution be a matrix of binary values $[\tilde{x}_{ij}]_{i,j}$ for $i \in [L]$ and $j \in [N]$. The corresponding solution for 3P can be obtained from this solution in polynomial-time as follows:

For every $i \in [m]$, $Y_i = \{\rho_j : \tilde{x}_{ij} = 1\}$.

The solution thus obtained is a feasible solution for 3P. To prove this, we need to show that:

- The solution results in a disjoint partition of Y, $\{Y_1, Y_2, \ldots, Y_m\}$.
- $\sum_{\rho_j \in Y_i} \rho_j = B$, for every i.
- $|Y_i| = 3$ for every i.

We shall prove these by contradiction as follows:

1. Let's first show that the resulting solution is a disjoint partition on Y. Suppose not. Then, one of the following two things must happen:

   (a) there exists Y_i and $Y_{i'}$ such that $Y_i \cap Y_{i'} \neq \phi$ or,

   (b) there exists some j' such that $\rho_{j'} \notin \bigcup_i Y_i$.

   If 1a is true and there exist Y_i and $Y_{i'}$ such that $Y_i \cap Y_{i'} \neq \phi$, it means that:

   $$\exists\ j \in [P] \text{ such that}, \tilde{x}_{ij} = 1 \text{ and } \tilde{x}_{i'j} = 1,$$
   $$\implies \sum_l \tilde{x}_{lj} \geq 2,$$

   which violates constraint (3.2) of $\mathbf{B^{\star}_{\Delta}}$. This means that $[\tilde{x}_{ij}]_{i,j}$ is not a feasible solution of $\mathbf{B^{\star}_{\Delta}}$ which is a contradiction. Therefore, $Y_i \cap Y_{i'} = \phi$ for every i and $i' \in [m]$.

If 1b is true and there exists $j' \in [P]$, such that $\rho_{j'} \notin \bigcup_i Y_i$, it means that $\tilde{x}_{ij'} = 0$ for every i. But, we have a feasible solution of $\mathbf{B^\star_D}$ which guarantees that the rate requirement of every group is satisfied. So,

$$\sum_{\rho_j \in Y_i} \rho_j \geq B, \ \forall \ i \in [m],$$
$$\implies \sum_{j=1, j\neq j'}^{P} \rho_j \geq mB, \implies \sum_{j=1}^{P} \rho_j > mB,$$

which is a contradiction. Hence, 1b cannot be true.

Therefore, the resulting solution will be a partition on Y.

2. We now show that $\sum_{\rho_j \in Y_i} \rho_j = B$, for every i. Suppose this is not true. Since $[\tilde{x}_{ij}]_{i,j}$ is a feasible solution of $\mathbf{B^\star_\Delta}$, we have, $\sum_{\rho_j \in Y_i} \rho_j \geq B$, for every $i \in [m]$. Let us say that at least one of these is a strict inequality. That is, there exists $i' \in [m]$ such that $\sum_{\rho_j \in Y_{i'}} \rho_j > B$. This implies that $\sum_{j=1}^{P} \rho_j > mB$, which is a contradiction. Therefore, we will have $\sum_{\rho_j \in Y_i} \rho_k = B$, for every i.

3. Next, we prove that $|Y_i| = 3$ for every Y_i. Let us suppose, for the sake of contradiction, that one subset, $Y_{i'}$ has less than 3 elements. Since the rate requirement of every group is B, we have, $\sum_{\rho_j \in Y_{i'}} \rho_j \geq B$. Also, from the problem definition of 3P, we have, $\rho_j < \frac{B}{2}$. Since $Y_{i'}$ can have a maximum of 2 members, we get, $\sum_{\rho_j \in Y_{i'}} \rho_j < B$ which is in contradiction to $\sum_{\rho_j \in Y_{i'}} \rho_j \geq B$ above. Thus, $Y_{i'}$ cannot have less than 3 elements. Therefore, $|Y_i| = 3$ for every $Y_i, \ i \in [m]$.

We have now established that a feasible solution for $\mathbf{B^\star_\Delta}$ also gives us a feasible solution for 3P. All that is left to complete the proof, is to show that, if $\mathbf{B^\star_\Delta}$ turns out to be infeasible, then, 3P has to be infeasible as well. We prove this by contradiction as follows:

Let's assume that 3P has a feasible solution even when $\mathbf{B^\star_\Delta}$ is infeasible. This means that, there exists a disjoint partition of Y, $\{Y_1, \ldots, Y_m\}$ such that, $\sum_{\rho_j \in Y_i} \rho_j = B$ and $|Y_i| = 3$ for every $Y_i, \ i \in [m]$. This solution can be mapped to a corresponding solution for $\mathbf{B^\star_\Delta}$ as follows:

$$x_{ij} = \begin{cases} 1, & \text{if } \rho_j \in Y_i, \\ 0, & \text{otherwise.} \end{cases}$$

So, for every i, we have:

$$\sum_{j=1}^{N} x_{ij} r_{ij} = \sum_{\rho_j \in Y_i} r_j = \sum_{\rho_j \in Y_i} \rho_j = B = R.$$

Also, since Y_i's form a disjoint partition of Y, we will have, $\sum_{i=1}^{N} x_{ij} \leq 1$ for every j. This means that $[x_{ij}]_{i,j}$ is a feasible solution for $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ which is a contradiction. Therefore, 3P has to be infeasible every time $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ is infeasible.

Thus, a polynomial-time solution for $\mathbf{B}^{\star}_{\mathbf{D}}$ results in a polynomial-time solution for 3P as well which is not possible unless P = NP. Therefore, there is no polynomial-time algorithm for solving the optimal resource allocation problem. Hence, $\mathbf{B}^{\star}_{\mathbf{D}}$ is an NP-complete problem. □

Lemma 1. *Optimization* $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ *is NP-hard.*

Proof. The proof follows from Theorem 1. Since the decision version of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ is NP-complete, $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ is an NP-hard problem. □

3.2.2 Problem 2: Optimal Grouping $\mathbf{C}^{\star}$

Recall that $S^{\Delta}_{\Gamma}[t]$ denotes the number of PRBs left unutilized under grouping policy Δ in sub-frame t when using resource allocation scheme Γ. Note that these PRBs can be used for other services in the system. Define,

$$\overline{S}^{\Delta}_{\Gamma} = \liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} S^{\Delta}_{\Gamma}[t]. \tag{3.3}$$

Thus, $\overline{S}^{\Delta}_{\Gamma}$ is the average number of unutilized PRBs per sub-frame under grouping policy Δ and resource allocation policy Γ. The optimal grouping problem can be defined for any Γ. The definition of the optimal grouping problem is stated below:

($\mathbf{C}^{\star}$) : Determine the optimal grouping policy $\Delta^{\star}$ such that $\overline{S}^{\Delta^{\star}}_{\Gamma} \geq \overline{S}^{\Delta}_{\Gamma}$ for every Δ.

We note that determining $\overline{S}^{\Delta}_{\Gamma}$ for a general grouping Δ and resource allocation Γ itself is a very hard, if not an impossible problem. The value of $\overline{S}^{\Delta}_{\Gamma}$ depends on the combined channel states of all the UEs in various sub-frames. We show in the following result that the problem of determining $\Delta^{\star}$ for a given Γ is NP-hard.

3.2.2.1 $\mathbf{C}^\star$ is NP-hard

Before addressing the hardness of the optimal grouping problem, we wish to point out that, given a grouping policy, Δ, calculating $\overline{S}^{\Delta}$ in polynomial-time is itself hard. Computing $\overline{S}^{\Delta}$ is non-trivial even when the channels are independent across UEs. We prove the NP-hardness of $\mathbf{C}^\star$ by reduction from the Set Cover problem which is an NP-complete problem [99] and is defined as follows [99]:

- **Input:** Set Cover takes as input a universe, $\mathcal{U} = \{u_1, \ldots, u_m\}$ containing m elements and a set $\mathcal{S} = \{S_1, \ldots, S_n\}$ of subsets of $\mathcal{U}$ such that $\bigcup_{j=1}^{n} S_j = \mathcal{U}$.
- **Problem:** Any collection of subsets from $\mathcal{S}$ form a set cover if their union is equal to the universe. The Set Cover problem is required to determine the smallest such collection of subsets.
- **Output:** The output is the smallest collection of subsets from $\mathcal{S}$ that form a set cover.

We now show the NP-hardness of $\mathbf{C}^\star$ by reduction from Set Cover.

Lemma 2. *For a fixed Γ, the problem of determining $\Delta^\star$ is NP-hard.*

Proof. In order to prove that $\mathbf{C}^\star$ is NP-hard, we first need to show that $\mathbf{C}^\star$ belongs to the class NP. Given a certificate for $\mathbf{C}^\star$, we can verify in polynomial-time whether or not it is a solution by checking if it satisfies the requirements stated in Definition 1. This can be done in $\mathcal{O}(L^2)$ computations. Therefore, $\mathbf{C}^\star \in$ NP.

We now prove that $\mathbf{C}^\star$ is NP-hard by reducing Set Cover to an instance of $\mathbf{C}^\star$. The pseudo-code for the algorithm used for this reduction is presented in Algorithm 2. The reduction can be accomplished in $\mathcal{O}(MN)$ computations. We define the total number of multicast UEs to be m and the number of PRBs in a sub-frame to be n. The k^{th} UE in $\mathbf{C}^\star$ maps to the variable u_k in Set Cover. Let r_{max} denote the maximum rate achievable in any PRB. The rate achievable by a UE k in PRB j, r_{kj} is defined to be equal to r_{max} if $u_k \in S_j$ and equal to 0 otherwise. We define the rate requirement of the groups to be $R < r_{max}$.

Let us now assume that there exists a polynomial-time algorithm for solving $\mathbf{C}^\star$. Using this algorithm to solve $\mathbf{C}^\star$ will output some grouping $\{G_1, \ldots, G_l\}$. We now show

how to map this output to a solution for Set Cover in polynomial-time. Since the rate achievable by a UE in any PRB can either be r_{max} or 0, all the UEs that are grouped together will be able to achieve r_{max} in some PRB and the number of PRBs needed to satisfy the groups will be exactly l because 1 PRB will be sufficient for providing the required rate. Let the n_i^{th} PRB be that PRB for group G_i. Hence, $u_k \in G_i, r_{kn_i} = r_{max} \implies u_k \in S_{n_i}$. Therefore, the corresponding solution for the Set Cover problem is $\{S_{n_1}, \ldots, S_{n_l}\}$. By the definition of a grouping we have $\cup_{i=1}^{l} G_i = [m] \implies \cup_{i=1}^{l} S_{n_i} = \mathcal{U}$. Therefore, the resulting solution is a valid set cover of $\mathcal{U}$.

We now show that this is indeed the smallest such collection that covers the universe $\mathcal{U}$. Suppose that this is not true. Then, there exists a collection of subsets from $\mathcal{S}$ smaller than l that forms a set cover. Let's denote this optimal solution as $\mathcal{S}' = \{S_{n'_1}, \ldots, S_{n'_z}\}, z < l$. We can then construct the following grouping from this set cover:

$$G_1 = S_{n'_1},\ G_2 = S_{n'_2} \setminus S_{n'_1},\ G_3 = S_{n'_3} \setminus \bigcup_{j=1}^{2} S_{n'_j}, \ldots,\ G_z = S_{n'_z} \setminus \bigcup_{j=1}^{k} S_{n'_j}.$$

Since $\mathcal{S}'$ is a set cover for $\mathcal{U}$, we have $\bigcup_{i=1}^{z} G_i = [M]$ and by construction of the groups, $\bigcap_{i=1}^{z} = \phi$. Hence, $\{G_1, \ldots, G_z\}$ is a valid grouping. Moreover, the number of PRBs needed to satisfy the UEs under this grouping will be $z < l$ which is a contradiction to the grouping $\{G_1, \ldots, G_l\}$ being the optimal solution of $\mathbf{C}^\star$. Therefore, $\{S_{n_1}, \ldots, S_{n_l}\}$ is the optimal solution of the Set Cover problem.

Thus, a polynomial-time solution for $\mathbf{C}^\star$ results in a polynomial-time solution for Set Cover as well which is not possible unless P = NP. Therefore, there is no polynomial-time algorithm for solving $\mathbf{C}^\star$ i.e. $\mathbf{C}^\star$ is an NP-hard problem. □

Algorithm 2: Pseudo-code for reducing Set Cover to $\mathbf{C}^\star$

Input: Set Cover problem with a universe $\mathcal{U} = \{u_1, \ldots, u_m\}$ of m variables and set $\mathcal{S} = \{S_1, \ldots, S_n\}$ of subsets of $\mathcal{U}$ such that $\bigcup_{i=1}^{n} S_i = \mathcal{U}$,

Output: An instance of $\mathbf{C}^\star$

1 $M \leftarrow m$, $N \leftarrow n$, k^{th} UE $\leftarrow u_k$

2 $$r_{kj} = \begin{cases} r_{max}, & \text{if } u_k \in S_j, \\ 0, & \text{otherwise.} \end{cases}$$

Since we have proved that both optimal grouping and optimal resource allocation problems are NP-hard, no polynomial-time algorithms exist for determining their optimal solutions. We can, however, use some intelligent heuristic schemes to obtain near optimal solutions. In the following section, we formulate an iterative randomized scheme for estimating the optimal resource allocation.

3.3 Randomized Algorithm for Optimal Resource Allocation

As stated in the previous section, no polynomial-time algorithm exists for determining the optimal resource allocation for the system under consideration. We can, however, estimate the optimal solution using randomized algorithms that iteratively explore the possible solutions to finally converge to the optimum. The randomized scheme proposed here serves dual purpose, 1) it provides near optimal solutions in much lesser computational power than that required to solve $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ and, 2) its output can be used as a benchmark for evaluating the performance of heuristic resource allocation schemes. We now describe the randomized algorithm.

The allocation of resources in LTE is done in every sub-frame. So, for brevity, we fix a sub-frame t and omit it from notations in this section. Grouping strategy Δ impacts resource allocation via r_{ij}^{Δ}, which is the rate achievable by group i in PRB j under Δ. Here, we deal with resource allocation for any given Δ. So, we omit Δ from the notations as well for better readability.

We refer to the randomized resource allocation algorithm proposed in this section as the Randomized Scheme (RS). The RS is based on Simulated Annealing (SA), a well known Markov Chain Monte Carlo (MCMC) technique [100]. SA is a randomized algorithm used for obtaining the optimal solution of an optimization problem. In SA, we construct a Markov chain on the states of the problem under consideration and transition between the states to ultimately converge to at the global optimum with high probability. Here, states correspond to possible resource allocations to the multicast groups. Therefore, a state, s_d of the Discrete Time Markov Chain (DTMC) is a possible distribution of PRBs, $\{\overline{V}_{0d}, \overline{V}_{1d} \ldots, \overline{V}_{Ld}\}$ where $\overline{V}_{id}$ is the set of PRBs assigned to group G_i, $i \in [L]$

in state s_d. G_0 is a dummy group that is assigned all the unused PRBs. The state space χ corresponds to all possible PRB allocations to groups. Let ℓ_{di} denote the total rate achieved by the i^{th} group in allocation s_d. Thus, $\ell_{di} = \sum_{j \in \overline{V}_{id}} r_{ij}$. Moreover, let q_d denote $|\{i : \ell_{di} \geq R\}|$, i.e. q_d is the number of satisfied groups in allocation s_d. In SA, each state has an associated reward that defines how good the state is. For our DTMC, we define the real valued reward function E as follows:

$$E(s_d) = |\overline{V}_{0d}| - \sum_{i=1}^{L} [R - \ell_{di}]^+ + q_d, \tag{3.4}$$

where $[y]^+ = \max\{y, 0\}$ and $|\overline{V}_{0d}|$ is the number of unused PRBs in state s_d. The reward function is a monotonically increasing function of the number of satisfied groups and the number of unused PRBs. It also decreases proportionally with the difference between the required and achieved rates of the groups. Thus, intuitively, maximizing E will maximize the number of unused PRBs while satisfying all the groups. We prove this formally in the next result.

Lemma 3. *Let $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ has a feasible solution and $s_{d^\star} \in \arg\max_{s_d} E(s_d)$. Define $x^{\star}_{ij} = 1$ if $j \in \overline{V}_{id^\star}$ and 0 otherwise. Then, $\{x^{\star}_{ij}\}_{i,j}$ is the optimal solution of the BLP $\mathbf{B}^{\star}_{\mathbf{\Delta}}$.*

Proof. We have $s_{d^\star} \in \arg\max_{s_d} E(s_d)$ i.e. $E(s_{d^\star}) \geq E(s_d)$ for every $s_d \in \chi$. The solution for the BLP $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ corresponding to the state $s_{d^\star}$, $\{x^{\star}_{ij}\}_{i,j}$ is obtained as follows:

$$x^{\star}_{ij} = \begin{cases} 1, & \forall\ j \in \overline{V}_{id^\star}, \\ 0, & \text{otherwise.} \end{cases}$$

In LTE, the rates achievable in a PRB are discrete and can take 15 different values corresponding to the 15 CQI values [7]. We denote the minimum rate that can be provided in a single PRB by r_{min}. Since the value of $E(.)$ depends on the value of R, two cases arise:

- $R \leq r_{min}$: In this case, we can satisfy all groups by allocating a single PRB to every group. This is a trivial case and so, it is sufficient to consider the case with $R > r_{min}$.
- $R > r_{min}$: Before proving that $\{x^{\star}_{ij}\}_{i,j}$ is the optimal solution of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$, we will first show that $\{x^{\star}_{ij}\}_{i,j}$ is a feasible solution of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$. Suppose $\{x^{\star}_{ij}\}_{i,j}$ is not a feasible solution

of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$. This means, that there exists $i \in [L]$ such that $\sum_{j=1}^{N} x^{\star}_{ij} r_{ij} < R$. Then the reward of $s_{d^{\star}}$ will be:

$$E(s_{d^{\star}}) = \left(N - \sum_{i\in[L]} \sum_{j\in[N]} x^{\star}_{ij} \right) - \sum_{i=1}^{L} [R - \ell_{d^{\star}i}]^{+} + q_{d^{\star}}. \tag{3.5}$$

Note that $q_{d^{\star}} < L$ because $\{x^{\star}_{ij}\}_{i,j}$ is infeasible. Depending on the value of $\sum_{i\in[L]} \sum_{j\in[N]} x^{\star}_{ij}$, two cases arise:

1. $\sum_{i\in[L]} \sum_{j\in[N]} x^{\star}_{ij} < N$: For this case, consider a state s_d obtained from $s_{d^{\star}}$ by allotting one of the PRBs, $j' \in \overline{V}_{0d^{\star}}$ to one of the unsatisfied groups i'. On allocating j' to i', one of two things can happen:

   - Rate requirement of the group i' is satisfied: This means that $q_d = q_{d^{\star}} + 1$. The reward of the resulting s_d will be:

     $$E(s_d) = E(s_{d^{\star}}) + (R - \ell_{d^{\star}i'}).$$

     Since group i' was unsatisfied in state $s_{d^{\star}}$, $(R - \ell_{d^{\star}i'}) > 0$. Therefore, $E(s_d) > E(s_{d^{\star}})$ which is a contradiction because $E(s_{d^{\star}}) \geq E(s_d)$ for every $s_d \in \chi$.

   - Rate requirement of the group i' is not satisfied: In this case, the reward of the state s_d will be:

     $$E(s_d) = E(s_{d^{\star}}) - 1 + (\ell_{di'} - \ell_{d^{\star}i'}).$$

     Here, $(\ell_{di'} - \ell_{d^{\star}i'})$ is the additional rate provided to group i' by the PRB j' which is why it can be no less than r_{min}. Since $r_{min} > 1$, $E(s_d) > E(s_{d^{\star}})$ which is a contradiction.

2. $\sum_{i\in[L]} \sum_{j\in[N]} x^{\star}_{ij} = N$: Here, the reward of $s_{d^{\star}}$ is:

   $$E(s_{d^{\star}}) = q_{d^{\star}} - \sum_{i=1}^{L} [R - \ell_{d^{\star}i}]^{+}.$$

   Since $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ is feasible, let $s_{d'}$ be a state corresponding to a feasible solution $\{x_{ij}\}_{i,j}$. The reward of $s_{d'}$ will be:

   $$E(s_{d'}) = \left(N - \sum_{i\in[L]} \sum_{j\in[N]} x_{ij} \right) + L > E_{s_{d^{\star}}},$$

which is a contradiction.

Therefore, $\{x^{\star}_{ij}\}_{i,j}$ has to be a feasible solution of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$. All we need to complete the proof is to show that $\{x^{\star}_{ij}\}_{i,j}$ is also an optimal solution of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$. We show this as follows:

Suppose $\{x^{\star}_{ij}\}_{i,j}$ is not an optimal solution of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$. Let's denote the optimal solution of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ by $\{\overline{x}_{ij}\}_{i,j}$ and the corresponding resource allocation state by $s_{\overline{d}}$. Since $\{x^{\star}_{ij}\}_{i,j}$ is not the optimal solution, we will have, $\sum_{i\in[L]}\sum_{j\in[N]} x^{\star}_{ij} > \sum_{i\in[L]}\sum_{j\in[N]} \overline{x}_{ij}$ The reward of $s_{\overline{d}}$ will be:

$$E(s_{\overline{d}}) = \left(N - \sum_{i\in[L]}\sum_{j\in[N]} \overline{x}_{ij}\right) + L,$$

$$\implies E(s_{\overline{d}}) > \left(N - \sum_{i\in[L]}\sum_{j\in[N]} x^{\star}_{ij}\right) + L = E(s_{d^{\star}}),$$

which is a contradiction. Therefore, $\{x^{\star}_{ij}\}_{i,j}$ is an optimal solution of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$. □

Thus, determining a state that maximizes the reward function is equivalent to determining the optimal solution of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$. Note that the proposed approach uses a DTMC on χ where $|\chi| = (L+1)^N$. Recall that L denotes the number of groups and N denotes the number of PRBs available in a sub-frame. Hence, the Transition Probability Matrix (TPM) corresponding to the DTMC will have dimensions exponential in N. So, for guaranteeing computational feasibility of the proposed approach, one must ensure that the TPM need not be stored, rather, given the current state, transition probability to the neighboring states can be determined in time polynomial in system parameters. Next, we elaborate how such a DTMC can be constructed.

3.3.1 DTMC Construction

Let $E^{\star}$ denote the maximum value of the reward function $E(\cdot)$ defined in (3.4), i.e. $E^{\star} = \max_{s_d} E(s_d)$. Suppose we construct a DTMC $\{X_n\}_{n\geq 1}$ on χ such that

$$\lim_{n\to\infty} P(E(X_n) = E^{\star}) = 1.$$

If we simulate this DTMC for a large enough time, say τ, the probability that the state of the DTMC at τ yields the optimal resource allocation should be very close to one. Towards this end, we first define a time homogeneous DTMC $\{X^T_n\}_{n\geq 1}$ on χ. We

will subsequently define the DTMC $\{X_n\}_{n\geq 1}$ with parameter T varying as a function of n. As we will see in the following sections, transition probabilities of the constructed DTMC are a function of T. Therefore, variation of T as a function of n makes $\{X_n\}_{n\geq 1}$ non time homogeneous. For defining the DTMC $\{X_n^T\}_{n\geq 1}$, it is enough to specify its TPM, which we do next.

3.3.1.1 Neighboring States

Consider any state $s_d \in \chi$. A state $s_{d'}$ is a neighbor of s_d if it can be obtained from s_d using one of the following actions:

- *Swap* (A_1): Swap takes two PRBs j_1 and j_2 from groups i_1 and i_2 respectively and assigns j_1 to i_2 and j_2 to i_1. Only allocation to groups i_1 and i_2 are changed through this action. Mathematically, $s_{d'}$ is obtained from s_d using swap if:

  1. $j_1 \in \overline{V}_{i_1 d}$ and $j_2 \in \overline{V}_{i_2 d}$,
  2. $\overline{V}_{id'} = \overline{V}_{id}$ for all $i \neq i_1, i_2$ and
  3. $\overline{V}_{i_1 d'} = (\overline{V}_{i_1 d} \setminus \{j_1\}) \cup \{j_2\}$, $\overline{V}_{i_2 d'} = (\overline{V}_{i_2 d} \setminus \{j_2\}) \cup \{j_1\}$.

- *Drop* (A_2): The drop action takes a PRB j_1 from a group i_1 ($i_1 \neq 0$) and assigns it to group G_0. Here, only allocation of groups i_1 and 0 is changed by dropping the PRB j_1. Mathematically, $s_{d'}$ is obtained from s_d using drop if:

  1. $j_1 \in \overline{V}_{i_1 d}$,
  2. $\overline{V}_{id'} = \overline{V}_{id}$ for all $i \neq i_1, 0$ and
  3. $\overline{V}_{i_1 d'} = \overline{V}_{i_1 d} \setminus \{j_1\}$, $\overline{V}_{0d'} = \overline{V}_{0d} \cup \{j_1\}$.

- *Add* (A_3): The add action takes a PRB j_1 from $\overline{V}_{0d}$ and assigns it to a group $i_1 \neq 0$. Here, only allocation of groups i_1 and 0 is changed by assigning the PRB j_1 to group i_1. Mathematically, $s_{d'}$ is obtained from s_d using add if:

  1. $j_1 \in \overline{V}_{0d}$,
  2. $\overline{V}_{id'} = \overline{V}_{id}$ for all $i \neq i_1, 0$ and
  3. $\overline{V}_{i_1 d'} = \overline{V}_{i_1 d} \cup \{j_1\}$, $\overline{V}_{0d'} = \overline{V}_{0d} \setminus \{j_1\}$.

Note that the neighboring relation defined here is symmetric in nature. This is proved in the following result.

Lemma 4. *The neighboring relation of the DTMC $\{X_n^T\}_{n\geq 1}$ is symmetric. Moreover, if transition from s_d to $s_{d'}$ occurs due to a swap action, then transition from $s_{d'}$ to s_d can also take place using a swap action only. Similarly, if transition to s_d from $s_{d'}$ occurs due to add (drop, respectively), the transition from $s_{d'}$ to s_d can only result from drop (add, respectively).*

Proof. To prove the required result, we need to show that if a state $s_{d'}$ is a neighbor of the state s_d, then, s_d is also a neighbor of $s_{d'}$. Since neighbors are defined using three different actions, we consider the cases corresponding to each action separately:

- **Swap:** Consider that $s_{d'}$ is obtained from s_d by swapping PRBs j_1 and j_2 belonging to groups i_1 and i_2 respectively. Then, from the definition of the swap action, $\overline{V}_{id'} = \overline{V}_{id}$ for all $i \neq i_1, i_2$, $\overline{V}_{i_1d'} = (\overline{V}_{i_1d} \backslash \{j_1\}) \cup \{j_2\}$ and $\overline{V}_{i_2d'} = (\overline{V}_{i_2d} \backslash \{j_2\}) \cup \{j_1\}$. Now, let us see if state s_d can be obtained from $s_{d'}$. Say PRBs j_1 and j_2 are picked for swapping in $s_{d'}$. Note that in $s_{d'}$, $j_1 \in \overline{V}_{i_2d'}$ and $j_2 \in \overline{V}_{i_1d'}$. For the resulting state $s_{d''}$, we have:

$$\begin{aligned}
\overline{V}_{id''} &= \overline{V}_{id'} = \overline{V}_{id}, \forall\ i \neq i_1, i_2, \\
\overline{V}_{i_1d''} &= (\overline{V}_{i_1d'} \setminus \{j_2\}) \cup \{j_1\} = \overline{V}_{i_1d}, \\
\overline{V}_{i_2d''} &= (\overline{V}_{i_2d'} \setminus \{j_1\}) \cup \{j_2\} = \overline{V}_{i_2d}.
\end{aligned}$$

  Therefore, $\overline{V}_{id''} = \overline{V}_{id}$ for all i which implies that $s_{d''} \equiv s_d$. So, s_d is also a neighbor of $s_{d'}$ and can be obtained from $s_{d'}$ using a swap action only.

- **Add:** Consider that $s_{d'}$ is obtained from s_d by adding PRB j_1 to group i_1. Then, from the definition of the add action, $\overline{V}_{id'} = \overline{V}_{id}$ for all $i \neq i_1, 0$, $\overline{V}_{i_1d'} = \overline{V}_{i_1d} \cup \{j_1\}$ and $\overline{V}_{0d'} = \overline{V}_{0d} \backslash \{j_1\}$. Now, let us see if state s_d can be obtained from $s_{d'}$. Say PRB j_1 is picked for a drop action in $s_{d'}$. Note that in $s_{d'}$, $j_1 \in \overline{V}_{i_1d'}$. For the resulting state $s_{d''}$, we have:

$$\begin{aligned}
\overline{V}_{id''} &= \overline{V}_{id'}, \forall\ i \neq i_1, 0, \\
\overline{V}_{i_1d''} &= \overline{V}_{i_1d'} \setminus \{j_1\} = \overline{V}_{i_1d}, \\
\overline{V}_{0d''} &= \overline{V}_{0d'} \cup \{j_1\} = \overline{V}_{0d}.
\end{aligned}$$

Therefore, $\overline{V}_{id''} = \overline{V}_{id}$ for all i which implies that $s_{d''} \equiv s_d$. So, s_d is also a neighbor of $s_{d'}$ and can be obtained from $s_{d'}$ using a drop action only.

- **Drop:** The proof for drop action is very similar to that for add. It can be shown in the same manner that if $s_{d'}$ is obtained from s_d using a drop action, s_d can be obtained from $s_{d'}$ using an add action and so s_d is also a neighbor of $s_{d'}$.

□

We now define the transition probability matrix.

3.3.1.2 Transition Probability Matrix

Let $p_{dd'}$ denote the probability that the DTMC transitions to $s_{d'}$ in the next step from the current state s_d. The transition happens in two steps. 1) In state s_d, we first randomly choose one of the three actions A_1(swap), A_2(add) or A_3 (drop) and then randomly choose a neighboring state s_{d_p} that can be obtained from s_d by performing the chosen action. The state s_{d_p} is referred to as the proposed future state. 2) Based on the reward values $E(s_d)$ and $E(s_{d_p})$, the proposed transition from s_d to s_{d_p} is either accepted, i.e. $s_{d'} = s_{d_p}$ or rejected, i.e. $s_{d'} = s_d$. We discuss these steps in detail below.

- *Step 1*: In this step, one of the three actions is picked. Since different actions lead to different sets of potential neighboring states, we use $s_{d_{A_i}}$ to denote a state that can be obtained from s_d by performing action $A_i, i \in \{1, 2, 3\}$. Probability of picking every action is different. Action A_1 is picked with probability (w.p.)

$$\beta_{dd_{A_1}} = \frac{1}{3}.$$

A_2 is picked w.p.

$$\beta_{dd_{A_2}} = \frac{2}{3} \times \frac{N - |\overline{V}_{0d}|}{L(|\overline{V}_{0d}| + 1) + (N - (|\overline{V}_{0d}| + 1))}.$$

A_3 is picked w.p.

$$\beta_{dd_{A_3}} = \frac{2}{3} \times \frac{L|\overline{V}_{0d}|}{L|\overline{V}_{0d}| + (N - |\overline{V}_{0d}|)}.$$

With the remaining probability, the state of the DTMC remains unchanged. A_3 corresponds to the add action and so, is chosen with a probability directly proportional to the number of unused PRBs ($|\overline{V}_{0d}|$) and the number of multicast groups

(L). Therefore, for a large number of groups and unused PRBs, the algorithm is more likely to choose the add action. Similarly, for a greater number of used PRBs $(N - |\overline{V}_{0d}|)$, the algorithm is more likely to choose the drop action.

Now we explain how one of the neighboring states is chosen for potential transition given the chosen action. If the chosen action is A_1, the two PRBs to be swapped, j_1 and j_2 are chosen uniformly at random from $[N]$. The swap of j_1 and j_2 is then performed as discussed in Section 3.3.1.1. For A_2, the PRB to be dropped, j_1 is picked uniformly at random from $[N] \setminus \overline{V}_{0d}$ and dropped as discussed in Section 3.3.1.1. Similarly for A_3, a group i_1 is picked uniformly at random from $[L]$ and a PRB to be added to it, j_1 is chosen uniformly at random from $\overline{V}_{0d}$. The addition of j_1 to i_1 is then done as discussed in Section 3.3.1.1. In the next step, we discuss how the exact values of the transition probabilities are determined.

- *Step 2*: Let $s_{d'}$ denote the state chosen for transition using the procedure in *Step 1*. If $s_{d'}$ has reward greater than or equal to that of s_d, the DTMC transitions to $s_{d'}$. Otherwise, the transition to $s_{d'}$ takes place w.p. $e^{(-(E(s_d)-E(s_{d'}))/T)}$. Thus, the probability that the DTMC will transition to $s_{d'}$ is

$$\alpha_{dd'} = \min\left(1, e^{(-(E(s_d)-E(s_{d'}))/T)}\right).$$

Here, T is a parameter commonly known as 'temperature' in SA [100]. For a fixed $T > 0$, $\{X_n^T\}_{n\geq 1}$ denotes the corresponding time homogeneous DTMC.

$s_{d_{A_1}}, s_{d_{A_2}}$ and $s_{d_{A_3}}$ denote the states resulting from s_d due to actions A_1, A_2 and A_3 respectively. Then the corresponding transition probabilities take the following form :

$$p_{dd_{A_1}} = \beta_{dd_{A_1}} \times \frac{1}{N(N-1)} \times \alpha_{dd_{A_1}}, \tag{3.6}$$

$$p_{dd_{A_2}} = \beta_{dd_{A_2}} \times \frac{1}{N - |\overline{V}_{0d}|} \times \alpha_{dd_{A_2}}, \tag{3.7}$$

$$p_{dd_{A_3}} = \beta_{dd_{A_3}} \times \frac{1}{|\overline{V}_{0d}|}\frac{1}{L} \times \alpha_{dd_{A_3}}, \tag{3.8}$$

$$p_{dd'} = 0, \text{if } s_{d'} \text{ is not a neighbor of } s_d. \tag{3.9}$$

Note that (3.6), (3.7), (3.8) and (3.9) completely describe the TPM. $p_{dd_{A_1}}$ is the probability of transitioning from s_d to $s_{d_{A_1}}$. In (3.6), $\beta_{dd_{A_1}}$ is the probability of picking

action A_1, the second term, $\frac{1}{N(N-1)}$ accounts for choosing any 2 PRBs for swapping and $\alpha_{dd_{A_1}}$ is the probability with which the DTMC transitions to the resulting state $s_{d_{A_1}}$. Thus, $p_{dd_{A_1}}$ is the overall probability of transitioning to state $s_{d_{A_1}}$ from s_d. Similarly, in (3.7) and (3.8), $\beta_{dd_{A_2}}$ and $\beta_{dd_{A_3}}$ are the probabilities of picking actions A_2 and A_3 respectively, $\frac{1}{N-|\overline{V}_{0d}|}$ is the probability of choosing one of the allocated PRBs for dropping, $\frac{1}{|\overline{V}_{0d}|}\frac{1}{L}$ is the probability of choosing a PRB for addition from $\overline{V}_{0d}$ times the probability of picking a group to which the PRB can be assigned, $\alpha_{dd_{A_2}}$ and $\alpha_{dd_{A_3}}$ are the probabilities with which the DTMC transitions to the resulting states $s_{d_{A_2}}$ and $s_{d_{A_3}}$ respectively. In (3.9), $p_{dd'} = 0$ because the DTMC cannot jump from s_d to a state that is not a neighbor of s_d.

Algorithm 3: Algorithm for the Randomized Scheme

Input: Rates $r_{ij} \forall\ i \in [L]$ and $j \in [N]$, $max\_iter = 10^5$

Initialize: s_0, initial random allocation state

$s_d \leftarrow s_0$

for $n = 1 : max\_iter$ **do**

  $s_{d'} \leftarrow s_d$

  $T \leftarrow \frac{1}{\log(n)}$

  Pick action A_1, A_2 or A_3 w.p. $\beta_{dd_{A_1}}$, $\beta_{dd_{A_2}}$ and $\beta_{dd_{A_3}}$ respectively

  if action=A_1 **then**

    Pick any two PRBs, $j_1, j_2 \in [N]$. Say, $j_1 \in \overline{V}_{i_1d'}$ & $j_2 \in \overline{V}_{i_2d'}$

    $\overline{V}_{i_1d'} = \overline{V}_{i_1d'} \setminus \{j_1\} \cup \{j_2\}$, $\overline{V}_{i_2d'} = \overline{V}_{i_2d'} \setminus \{j_2\} \cup \{j_1\}$

  else if action=A_2 **then**

    Pick a PRB, $j \in \cup_{i=1}^{L} \overline{V}_{id'}$. Say, $j \in \overline{V}_{id'}$

    $\overline{V}_{id'} = \overline{V}_{id'} \setminus \{j\}$, $\overline{V}_{0d'} = \overline{V}_{0d'} \cup \{j\}$

  else

    Pick any $j \in \overline{V}_{0d'}$ and any $i \in \{1, 2, \ldots, L\}$

    $\overline{V}_{id'} = \overline{V}_{id'} \cup \{j\}$, $\overline{V}_{0d'} = \overline{V}_{0d'} \setminus \{j\}$

  end

  $s_d \leftarrow s_{d'}$, if $E(s_{d'}) \geq E(s_d)$

  $s_d \leftarrow s_{d'}$ w.p. $e^{(-(E(s_d)-E(s_{d'}))/T)}$, otherwise

end

s_d is the proposed resource allocation

In the randomized scheme here, we aim to simulate the constructed DTMC with these transition probabilities. The steps involved in the randomized scheme are presented in the form of a pseudo-code in Algorithm 3. Note that the TPM of the DTMC is not being stored in this algorithm and the transition probabilities defined above can be determined in polynomial-time. Thus, the TPM satisfies all the conditions stated earlier for computational feasibility of the algorithm. In the next result, we prove certain important properties of the constructed DTMC.

Lemma 5. *The constructed time homogeneous DTMC* $\{X_n^T\}_{n\geq 1}$ *is finite, aperiodic and irreducible for every* $T \in (0, \infty)$.

Proof.

- Finite: The DTMC is finite because the total number of possible resource allocation states is $(L+1)^N$.
- Aperiodic: The DTMC has self loops because there is a positive probability of remaining in the same state at a transition epoch. Hence, the DTMC is aperiodic.
- Irreducible: The DTMC can transition from any state s_d to any other state $s_{d'}$ by first dropping all the used PRBs into G_0 by choosing the drop action repeatedly. Then, the PRBs can be added one by one according to the assignment in state $s_{d'}$ by choosing the add action repeatedly. Thus, there exists at least one finite length path from any state s_d to any other state $s_{d'}$. Therefore, the DTMC is irreducible.

□

Having established that the constructed DTMC is finite, aperiodic and irreducible, it is guaranteed to have a unique steady state distribution. In the following result, we determine this steady state distribution.

Theorem 2. *For any fixed* $T > 0$, *the steady state distribution of the DTMC* $\{X_n^T\}_{n\geq 1}$ *is given by*

$$\pi_d^T = \frac{e^{E(s_d)/T}}{\sum_{s_d} e^{E(s_d)/T}} \forall\ s_d \in \chi. \tag{3.10}$$

Proof. To prove the required, we show that the transition probabilities in (3.10) satisfy $\pi_d^T p_{dd'} = \pi_{d'}^T p_{d'd}$ for every $s_d, s_{d'}$. This will imply that the DTMC is reversible [101] and has steady state distribution $\pi_d^T = \frac{e^{E(s_d)/T}}{\sum_{s_d} e^{E(s_d)/T}}, \forall\ s_d \in \chi$.

Suppose s_d and $s_{d'}$ are not neighboring states, then $p_{dd'} = p_{d'd} = 0$. Hence, the required follows trivially. Thus, it suffices to consider the case when s_d and $s_{d'}$ are neighbors. If s_d and $s_{d'}$ are neighbors, there are three possibilities, that $s_{d'}$ is obtained from s_d by 1) swap action, 2) drop action or 3) add action. We consider each case separately:

- **Swap:** If the transition from s_d to $s_{d'}$ occurs due to a swap action, then $p_{dd'}$ and $p_{d'd}$ take the form given by (3.6). For $E(s_d) \geq E(s_{d'})$ we have:

$$\begin{aligned}&\frac{e^{E(s_d)/T}}{\sum_{d\in\chi} e^{E(s_d)/T}} \frac{1}{3} \frac{1}{N(N-1)} e^{-(E(s_d)-E(s_{d'}))/T} \\ =&\frac{e^{E(s_{d'})/T}}{\sum_{d\in\chi} e^{E(s_d)/T}} \frac{1}{3} \frac{1}{N(N-1)},\end{aligned}$$

  which is true. Therefore, the given π_d^T satisfies $\pi_d^T p_{dd'} = \pi_{d'}^T p_{d'd}$ for the swap action. This can be similarly shown for $E(s_d) < E(s_{d'})$ as well.

- **Add:** If the transition from s_d to $s_{d'}$ occurs due to an add action, $p_{dd'}$ and $p_{d'd}$ will be given by (3.8) and (3.7) respectively. For $E(s_d) \geq E(s_{d'})$, we have:

$$\begin{aligned}&\frac{2\pi_d^T}{3\left(L|\overline{V}_{0d}| + (N - |\overline{V}_{0d}|)\right)} e^{-(E(s_d)-E(s_{d'}))/T} \\ =&\frac{2\pi_{d'}^T}{3\left(L(|\overline{V}_{0d'}|+1) + (N - (|\overline{V}_{0d'}|+1))\right)}.\end{aligned} \tag{3.11}$$

  Since $s_{d'}$ is obtained from s_d using an add action, $|\overline{V}_{0d}| = |\overline{V}_{0d'}| + 1$ which means that $L|\overline{V}_{0d}| + (N - |\overline{V}_{0d}|) = L(|\overline{V}_{0d'}|+1) + (N - (|\overline{V}_{0d'}|+1))$ in (3.11) above. So, (3.11) becomes:

$$\begin{aligned}&\pi_d^T e^{-(E(s_d)-E(s_{d'}))/T} = \pi_{d'}^T, \\ \implies &\frac{e^{E(s_d)/T}}{\sum_d e^{E(s_d)/T}} e^{-(E(s_d)-E(s_{d'}))/T} = \frac{e^{E(s_{d'})/T}}{\sum_d e^{E(s_d)/T}},\end{aligned}$$

  which is true. Therefore, the given π_d^T satisfies $\pi_d^T p_{dd'} = \pi_{d'}^T p_{d'd}$ for the add action. This can be similarly shown for $E(s_d) < E(s_{d'})$ as well.

- **Drop:** If the transition from s_d to $s_{d'}$ occurs due to a drop action, $p_{dd'}$ and $p_{d'd}$ will be given by (3.7) and (3.8) respectively. Also, in this case, $|\overline{V}_{0d'}| = |\overline{V}_{0d}| + 1$. Following the same steps as for the add action, it can be shown that the given π_d^T satisfies $\pi_d^T p_{dd'} = \pi_{d'}^T p_{d'd}$ for the drop action as well.

Therefore, we conclude that the steady state distribution of the DTMC $\{X_n^T\}_{n\geq 1}$ is given by $\pi_d^T = \frac{e^{E(s_d)/T}}{\sum_{s_d} e^{E(s_d)/T}} \forall\ s_d \in \chi$. □

For a fixed value of T, the DTMC is time homogeneous with steady state distribution π_d^T as given in Theorem 2. However, when T varies as a function of time n, the DTMC no longer remains time homogeneous and the steady state distribution cannot be determined in the same manner. We require this non time homogeneous DTMC $\{X_n\}_{n\geq 1}$ to converge to a reward maximizing state. In the following theorem, we show that this does indeed happen.

Theorem 3. *For the non time homogeneous DTMC* $\{X_n\}_{n\geq 1}$, $\lim_{n\to\infty} P(X_n = s_d)$ *exists, call it* π_d. *Moreover,* $\pi_d = \lim_{T\to 0} \pi_d^T$. *Specifically,*

$$\pi_d = \begin{cases} 1/|\arg\max_d E(s_d)|, & \forall\ d \in \arg\max_d E(s_d), \\ 0, & \textit{otherwise.} \end{cases} \tag{3.12}$$

Thus, π_d *is a uniform distribution over the optimal reward maximizing resource allocation states.*

Proof. The result follows directly from Theorem 1 of [102]. □

We have mentioned the parameter T above, while discussing the TPM. Now, we elaborate its significance in greater detail. SA involves an exploration versus exploitation trade-off. Exploration involves transitioning to new states even if their rewards are lower than the current state of the DTMC whereas exploitation refers to only transitioning to a new state if it provides a reward greater than the current one. SA achieves a balance between exploration and exploitation through this parameter T. T is kept very high in the beginning so that the algorithm can explore a large number of states quickly. As the time index increases, T goes on decreasing and so does the likelihood of transitioning to lower reward states. $T = 1/\log(n)$, n being the time index is known to be the optimal cooling schedule [102]. This form of T ensures that the algorithm escapes local optima faster and converges to the global optimum as T goes to 0. Specifically, by varying T, we can achieve the required $\lim_{n\to\infty} P(E(X_n) = E^\star) = 1$. In the next section, we compare the results of the RS with the optimal solution obtained by solving the BLP $\mathbf{B}_{\mathbf{\Delta}}^\star$ for small input sizes.

Table 3.1: Performance comparison of RS and BLP

No. of groups	**RS**	**BLP**	**% Error**
2	93.53	96	2.57
3	90.05	94	4.2
4	86.54	91	4.9

3.3.2 Performance comparison of the RS and the BLP

The optimal resource allocation can be obtained by solving BLP $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ from Section 3.1. BLPs, as mentioned before, are inherently hard to solve. They can however be solved for small input sizes. Using the computing power at our disposal (Intel i7, 2.90 GHz quad-core processor with 16 GB RAM), we were able to obtain a solution of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$ for an input size of up to 4 groups. Note that the search space scales as $(L+1)^N$ where L is the number of groups and N is the number of PRBs in a sub-frame. So, even for 4 groups and 100 PRBs, the search space consists of 5^{100} states which is why the BLP fails to give a solution for more than 4 groups. The outputs of the BLP and the RS for up to 4 groups, averaged over 100 different channel conditions are tabulated in Table 3.1. As we can see, the output of the RS is close (difference in number of PRBs saved $< 5\%$) to the optimal obtained by solving the BLP.

The randomized scheme works iteratively to obtain an optimal solution. It can take several thousand iterations to converge and so, it is not guaranteed to converge within the sub-frame duration of 1 ms. We require resource allocation schemes that can output a near optimal solution (if not optimal) every sub-frame. We now present two heuristic schemes that give us a reasonably good performance in a time efficient manner. We compare the output of one of the proposed schemes with the output of the RS and show that it gives a solution very close to the optimum and takes significantly less time to run than the RS.

3.4 Heuristic Schemes for Resource Allocation

In this section, we propose two heuristic schemes for resource allocation in multicast. The first scheme allocates PRBs greedily and the second one makes use of Linear Programming (LP) relaxation of the BLP. Allocation of resources in LTE is done in every sub-frame. So, for brevity, we fix a sub-frame t and omit it from notations in this section. Grouping strategy Δ impacts resource allocation via r_{ij}^{Δ}, which is the rate achievable by group i in PRB j under policy Δ. Our aim is to propose resource allocation schemes for any given Δ. So, we fix Δ and omit it from the notations as well.

3.4.1 Greedy Allocation

Algorithm 4: Greedy Resource Allocation Scheme

Input: Rates r_{ij} for all $i \in [L]$ and $j \in [N]$

Initialize: $\mathcal{N} = [N]$, $\mathcal{L} = [L]$ and $x_{ij} = 0$ for every i, j
while $\mathcal{N} \cap \mathcal{L} \neq \phi$ **do**
    Assign $(i^\star, j^\star) = \arg\max_{(i,j)\in\mathcal{N}\times\mathcal{L}} r_{ij}$
    $x_{i^\star j^\star} \leftarrow 1$, $\mathcal{N} \leftarrow \mathcal{N} \setminus \{j^\star\}$
    if $\sum_{j\in[N]} x_{i^\star j} r_{i^\star j} \geq R$ **then**
        $\mathcal{L} \leftarrow \mathcal{L} \setminus \{i^\star\}$
    end
end

The pseudo code for this scheme is given in Algorithm 4. $\mathcal{N}$ and $\mathcal{L}$ denote the unallocated PRBs and the groups whose rate requirements are not yet satisfied, respectively. These quantities are updated every iteration and are monotone non-increasing. The algorithm terminates when either of the two sets becomes empty. In each iteration, the algorithm determines indices $i^\star$ and $j^\star$ from $\mathcal{L}$ and $\mathcal{N}$, respectively, that correspond to the maximum r_{ij}. PRB $j^\star$ is allotted to group $i^\star$ and is removed from $\mathcal{N}$. Also, if the total sum rate on all the PRBs allotted to $i^\star$ is greater than or equal to the requirement R, then $i^\star$ is also removed from $\mathcal{L}$. Next iteration starts with the new values of $\mathcal{N}$ and $\mathcal{L}$. Note that $\mathcal{N}$ is monotone decreasing, thus, the algorithm terminates in at most N iterations.

At the termination, if only $\mathcal{N} = \phi$ and $\mathcal{L}$ is non-empty, then the greedy resource allocation scheme fails to output a feasible resource allocation, else variables x_{ij}'s yield the required resource allocation. The algorithm has a complexity of $\mathcal{O}(LN^2)$. The resource allocation thus obtained is inherently fair as the algorithm provides the minimum required rate R to all the UEs.

3.4.2 LP-relaxation Based Allocation

Recall that the optimal resource allocation can be obtained as a solution to the BLP $\mathbf{B}^{\star}_{\mathbf{\Delta}}$. BLPs are inherently hard to solve and cannot be solved in reasonable time except for very small input sizes. A standard approach used for obtaining an approximate solution of BLPs is to do LP-relaxation of the BLP i.e., relax the binary variables (in our case, x_{ij}s) to take values in the interval $[0, 1]$. The resulting LP can be solved in polynomial-time. Let $\tilde{x}_{ij}$ for all i, j denote the optimal solution of the relaxed LP. Now, $\tilde{x}_{ij}$s are real numbers and we need to convert them to binary values without violating the constraints of $\mathbf{B}^{\star}_{\mathbf{\Delta}}$. To do so, we use the greedy algorithm given in Algorithm 5. In each iteration, PRB j is assigned to an unsatisfied group i if it has the largest value of $\tilde{x}_{ij}$ for that PRB. This is intuitive, as a higher value of $\tilde{x}_{ij}$ indicates that group i was assigned a larger share of PRB j by the LP. The algorithm has a complexity of $\mathcal{O}(LN^2)$. Note that the resource allocation obtained using this scheme is inherently fair as the algorithm ensures that the required rate R is provided to all the UEs. We shall refer to this scheme as the LPr scheme from this point onwards.

3.4.2.1 Performance Comparison of RS and LPr

In order to compare the performance of the LPr scheme to that of the RS, we simulate an LTE cell with all the multicast UEs requiring the same content from the eNB. PRBs are allocated to the UEs using the RS as well as the LPr scheme. We gradually increase the number of UEs in the cell starting from 10 UEs and go up to 100, adding 10 UEs at a time. For each of the resulting 10 scenarios, the PRB allocation is done for 100 different fading variations using both the schemes. The average number of PRBs saved is used as a measure for performance comparison. The results of the simulations are plotted in Figure 3.1. Each point in the curves has been obtained by averaging over 100 different

Algorithm 5: Rounding off algorithm for LP-relaxation

Input: $\tilde{x}_{ij}$ for all $i \in [L]$ and $j \in [N]$

Initialize: $\mathcal{N} = [N]$, $\mathcal{L} = [L]$ and $x_{ij} = 0$ for every i, j

while $\mathcal{N} \cap \mathcal{L} \neq \phi$ **do**

  Assign $(i^\star, j^\star) = \arg\max_{(i,j)\in\mathcal{N}\times\mathcal{L}} \tilde{x}_{ij}$

  $x_{i^\star j^\star} \leftarrow 1$, $\mathcal{N} \leftarrow \mathcal{N} \setminus \{j^\star\}$

  if $\sum_{j\in[N]} x_{i^\star j} r_{i^\star j} \geq R$ **then**

    $\mathcal{L} \leftarrow \mathcal{L} \setminus \{i^\star\}$

  end

end

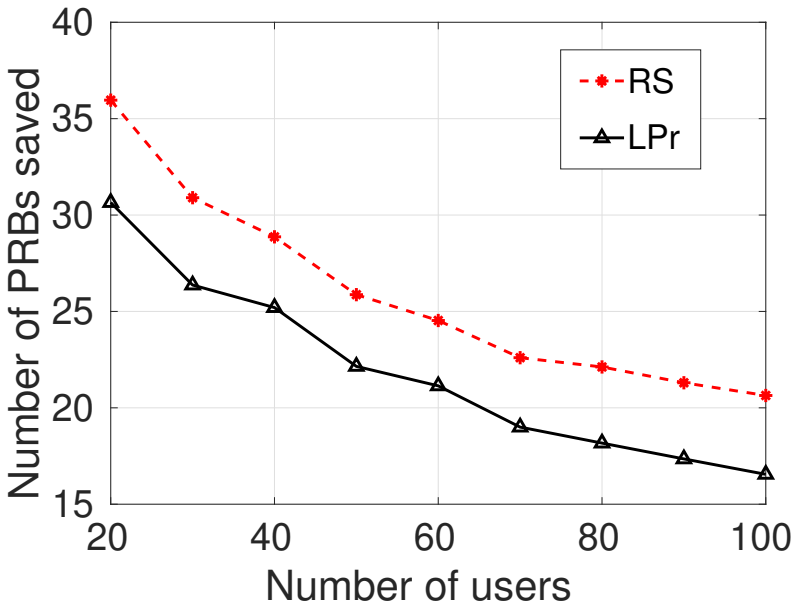


Figure 3.1: Number of PRBs saved under LPr and RS.

channel gain variations. Note that all the groups achieved the required rates at all points in the two curves. Both the algorithms show a similar trend as the number of UEs in the cell increases. Even though the RS saves more PRBs throughout, the ratio of the number of PRBs saved by the RS to the number of PRBs saved by the LPr scheme is no more than 1.25.

Table 3.2: Time taken in seconds to run RS and LPr

No. of UEs	**RS**	**LPr**	**Ratio**
20	0.082	0.015	5.47
40	0.086	0.019	4.53
60	0.089	0.021	4.24
80	0.097	0.017	5.71
100	0.096	0.018	5.33

3.4.2.2 Time Comparison of RS and LPr

Recall that, in LTE, the allocation of PRBs is done every sub-frame. Since a sub-frame spans only 1 ms in time, it is important for whatever resource allocation scheme we employ, to be time efficient. We now do a time comparison of RS and LPr schemes.

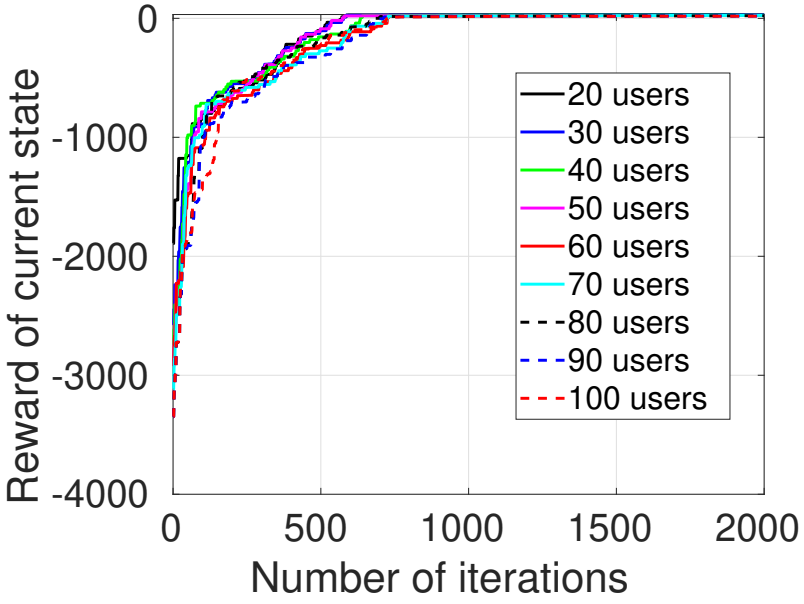


Figure 3.2: Variation of the reward of the state of RS with the increasing number of iterations.

The RS is an iterative algorithm and cannot be guaranteed to converge within the span of a sub-frame. While simulating the RS in this chapter, we run 10^5 iterations. However, for the time comparison here, we first see how the reward of the current state of the RS changes as a function of the number of iterations. Figure 3.2 illustrates the

change in the reward of the current state of the RS as a function of the number of iterations for different number of UEs in the cell. We can observe from the figure that the output saturates well before 2000 iterations in each curve. So, for the sake of time comparison with the LPr scheme, we consider the time taken by just 2000 iterations of the RS. Table 3.2 illustrates the time taken by the RS and the LPr scheme for different number of UEs in the cell. The time taken is averaged over 200 different channel gains. We observe that the RS takes about 5 times more time to run than the LPr scheme even with just 2000 iterations while providing only marginal performance gains as seen in the previous section. Note that in practice, depending upon the system, we might need to run the algorithm for a much larger number of iterations.

From the performance and time comparisons of the LPr and the RS, we conclude that LPr performs nearly as well as the RS in 5 times lesser duration than the RS. Thus, the LPr scheme is a suitable resource allocation scheme for practical implementation. In the next section, we present a heuristic scheme for the grouping of UEs for multicast transmission.

3.5 Heuristic Scheme for Grouping

We proved in Section 3.1 that obtaining the optimal grouping strategy $\Delta^\star$ that maximizes the performance measure $\overline{S}^{\Delta^\star}$ is an NP-hard problem. Indeed, even quantifying $\overline{S}^{\Delta}$ for a given grouping strategy Δ is a very difficult task as the channel gains and hence the rates vary over time. This is because obtaining the optimal resource allocation in a given sub-frame itself is an NP-hard problem (Lemma 1). However, even if some genie provides us with the value $\overline{S}^{\Delta}$ for any given Δ, determining the optimal $\Delta^\star$ is still NP-hard (Lemma 2). Hence, in this section, we present the following heuristic algorithm for grouping.

3.5.1 Hybrid Grouping Policy

Under the hybrid grouping policy, the eNB fixes the SNR thresholds for groups and then UEs are assigned to various groups based on their average SNR values. 3GPP standards for LTE [7] define 15 CQI values with 15 indicating the best and 1 indicating the worst

channel. In keeping with the number of CQI values, we fix the number of grouping intervals to be 15. In LTE, a range of SNR values get mapped to a CQI value [7] (many to one map). Let the minimum SNR that can be mapped to a CQI value c be denoted by $\text{SNR}_{\min}(c)$. We define a threshold corresponding to CQI c such that with a large probability (0.9), the instantaneous SNR of every UE in that group will stay above or at $\text{SNR}_{\min}(c)$. Specifically, a threshold $D(c)$ is defined such that,

$$P\{\text{SNR} \geq \text{SNR}_{\min}(c) | \text{SNR}_{\text{avg}} = D(c)\} = 0.9. \tag{3.13}$$

To compute $D(c)$, we need the distribution of SNR which depends on the distribution of $h_{iu}[t]$. $H_{iu}[t]$ (the fast-fading component of $h_{iu}[t]$ as defined in Section 3.1) are i.i.d exponential with mean 1. Given that the average SNR is equal to $D(c)$, the distribution of the instantaneous SNR is exponential with parameter $D(c)$. Therefore, (3.13) can be written as:

$$e^{-\frac{\text{SNR}_{min}(c)}{D(c)}} = 0.9, \tag{3.14}$$

$$\implies D(c) = \frac{\text{SNR}_{\min}(c)}{\log(10/9)}. \tag{3.15}$$

Now that the thresholds have been defined, UEs are classified into groups on the basis of their average SNR values. UEs with average SNR values greater than or equal to $D(15)$ are classified into Group 1 and those with SNR below $D(2)$ are grouped into Group 15. UEs with average SNR between $D(14)$ and $D(15)$ are put into Group 2 and so on. Thus, Group 1 (Group 15) corresponds to the UEs with the best (worst) channel.

As the size of a group grows, the probability that one or more of the UEs in that group will experience a poor channel increases. Therefore, the performance of the grouping scheme may start worsening with increasing group sizes. To prevent this, we propose a second layer of grouping. If the number of UEs in a group exceeds a certain maximum value, it is divided into smaller groups. We fix the maximum group size such that all the UEs in a group experience a good channel in at least 10% of the PRBs in a sub-frame. Since the thresholds have been set so that the instantaneous SNR of a UE remains above $\text{SNR}_{min}(c)$ with probability 0.9 and the channels are independent across UEs and across PRBs, this probability is given by:

$$p = \sum_{j=\lfloor 0.1N \rfloor}^{N} \binom{N}{j} 0.9^{kj}(1 - 0.9^k)^{(N-j)}, \tag{3.16}$$

where k denotes the group size. We need this probability to be large. For example, in the 20 MHz LTE system with $N = 100$, $p = 0.9452$ for $k = 18$. Therefore, we would fix the maximum group size at 18 for this system and whenever a group grows beyond 18 UEs, the group would be split into smaller groups of size 18 or less. Note that p is a monotonically decreasing function of k.

After the UEs are classified into groups, the rate for a particular group is set at the value corresponding to the weakest UE in the group. Once the achievable rate for each group is determined using the 3GPP mappings [7], the PRB allocation is done according to the resource allocation schemes discussed in the previous section.

3.6 Simulations

Our simulation setup is comprised of an LTE cell of radius 375 m in accordance with the simulation parameters for macro cell propagation model in [1]. We have used the MATLAB [103] LTE simulator designed in [104] to conduct our simulations. LTE specific physical layer conditions have been created using channel models recommended by 3GPP [1]. The SNR to CQI and CQI to rate mapping has also been done according to 3GPP specifications [1].

An eNB located at the center of the cell multicasts the eMBMS content to all the multicast groups. Rate requirement for each UE (R) is taken to be 1 Mbps. The UEs are distributed uniformly at random within the cell and are grouped using the hybrid grouping policy proposed in Section 3.5. For dividing the UEs into groups, we need to determine the average SNR received at the UEs. For calculating the average SNR, we use shadowing and path loss models provided in 3GPP specifications [1]. The channel gain of each UE may be different in different PRBs. The channel gains are determined by: 1): Path Loss, 2): Shadowing and 3): Multipath due to reflections from the surrounding environment. After the grouping is done, PRBs are allocated using the policies proposed in Section 3.4. Resource allocation policies make use of the instantaneous SNR values for taking the allocation decisions. To determine the instantaneous SNR of users, we take Rayleigh fading into account in addition to path loss, shadowing and multipath. We compare the performance of the proposed schemes with unicast transmission. The

Table 3.3: System Simulation parameters [1]

Parameters	**Values**
System bandwidth	20 MHz
Center frequency	2 GHz
eNB cell radius	375 m
Path loss model	$\mathrm{L} = 128.1 + 37.6 \log 10(d)$, d in kilometers
Shadowing	Log Normal Fading with 10 dB standard deviation
White noise power density	-174 dBm/Hz
eNB noise figure	5 dB
eNB transmit power	46 dBm
PRB width	180 kHz
Number of PRBs	100 per sub-frame
ITU path loss model	ITU-R M.2135-1 [105]

performance of the resource allocation policies is also compared to that of the widely used Proportional Fair (PF) policy [14, 30, 36]. Other parameters relevant to our simulations are given in Table 3.3.

For a given grouping and resource allocation, the system performance is affected by two sources of randomness, (1) channel variations around mean on account of fast fading and (2) average channel gain variations on account of user positions. We evaluate the performance of our schemes by averaging over these two sources of randomness. Towards this end, we consider 100 random UE placements and the performance of each placement is evaluated and averaged over 1000 sub-frames with different channel gains. In addition to unicast and the proposed grouping policy, we also simulate a random grouping where each UE is placed in one of 10 groups uniformly at random. Under the proposed hybrid grouping policy, a maximum of 15 groups can be formed. However, the actual number of groups formed will depend upon the average SNR of the users. The average number of groups formed during the simulations is given in Table 3.4.

3.6.1 Results

Figure 3.3 illustrates the plot of the number of PRBs saved against the number of UEs in the cell for greedy and LPr schemes under various grouping policies. The following observations can be made from these plots:

- Unicast performs the worst and is unable to support the rate requirements of more than 20 UEs successfully.
- Random grouping is able to support up to 30 UEs successfully. The number of PRBs saved rapidly decreases to 0 beyond 30 UEs.
- With hybrid grouping, greedy allocation saves greater than 10 PRBs for up to 70 UEs. Using LPr saves around 20 PRBs even for 100 UEs.

Figure 3.4 illustrates the number of sub-frames (out of 1000) for which the allocations are rendered infeasible for greedy and LPr allocation under various grouping policies. We observe that:

- Unicast and random grouping quickly become completely infeasible beyond 40 UEs under greedy scheme and beyond 50 under the LPr scheme.
- Using greedy allocation, the number of infeasible cases for hybrid grouping is zero for up to 40 UEs. Using LPr, the allocation is always feasible.

We have also conducted simulations for a rate requirement of $R = 2$ Mbps. The corresponding results are plotted in Figure 3.5 and Figure 3.6. We observe a similar relative performance of the policies as that for $R = 1$ Mbps.

Figure 3.7 and Figure 3.8 illustrate the number of PRBs saved at the eNB for different UE placements. For every M number of UEs in the cell, 100 different placements have been considered. Out of these, 90% closest to the mean have been plotted as a scatter plot. The means of the observations have also been indicated in the figures. The following conclusions can be drawn from these plots:

- For varying UE placements, the number of PRBs saved at the eNB is between ± 5 PRBs around the mean number saved for all the schemes.

Table 3.4: Average number of groups formed

No. of UEs	**No. of groups**	**No. of UEs**	**No. of groups**
10	5.39	20	6.94
30	7.75	40	7.96
50	8.45	60	8.39
70	8.66	80	8.77
90	8.77	100	9

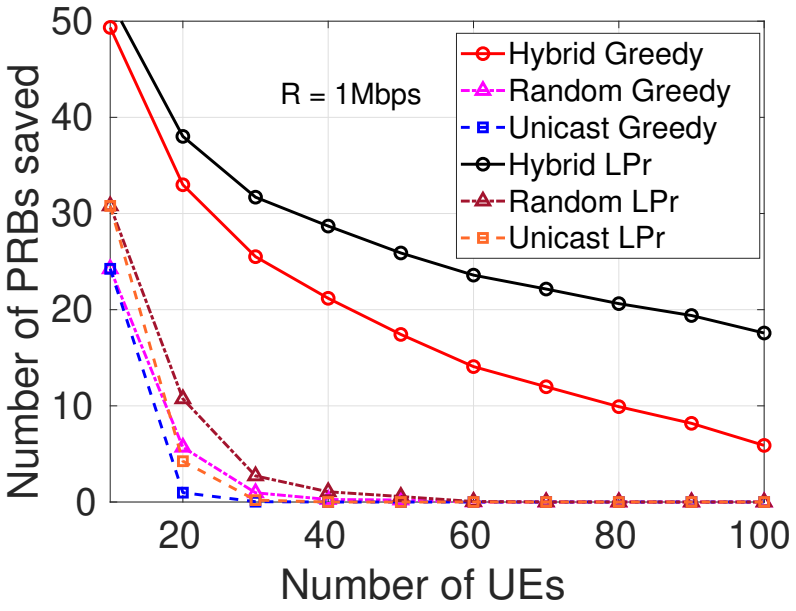


Figure 3.3: Number of PRBs saved under various policies

- Overall trend of the number of PRBs saved as the UE count increases is the same as observed in Figure 3.3.

In addition to QoS, it is also important to guarantee a good QoE in video streaming. QoE is known to be a function of various QoS parameters of the network [106]. The QoE of a video stream primarily depends on the delay, delay jitter and the packet loss rate in the network [107, 108]. To study the impact of our policies on the QoE of users, we evaluate their performance using data from an actual video stream. For this purpose, we use an H.264/AVC encoded video of Star Wars IV (obtained from `(http://trace.eas.asu.edu)`) [16]. For transmitting this video stream, the required rate R is changed every

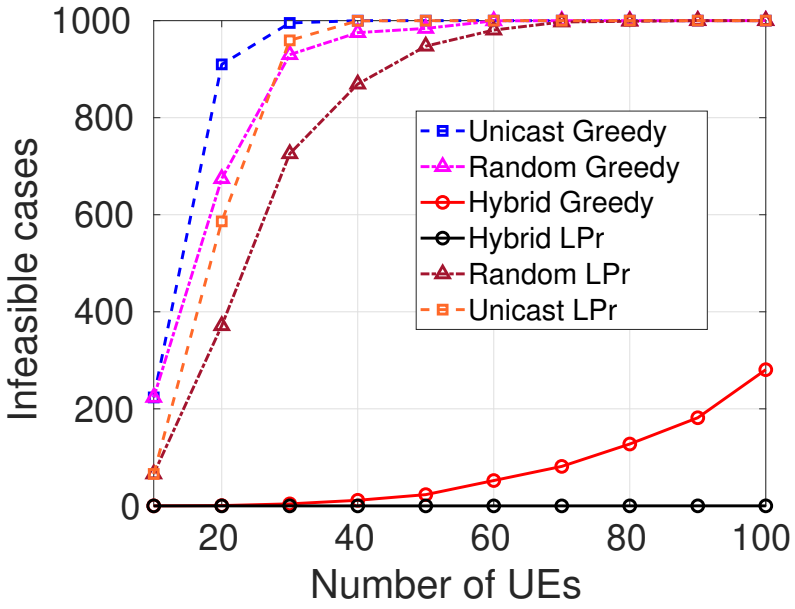


Figure 3.4: Number of infeasible cases under various

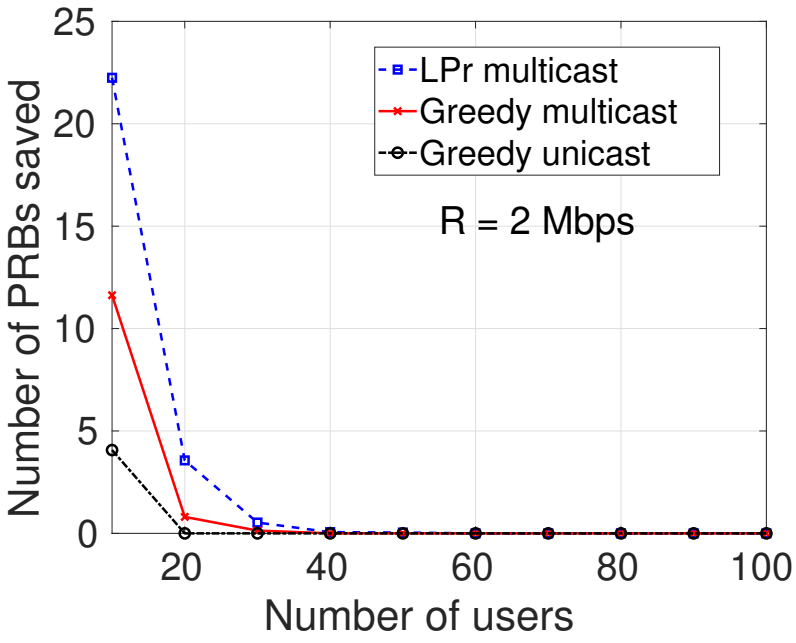


Figure 3.5: Number of PRBs saved under various resource allocation policies for $R = 2$ Mbps

sub-frame according to the requirement of the video frame being transmitted. Under our policies, packets of the video are transmitted as soon as they arrive. As a result, the access network does not induce any additional delay and jitter in the video stream. The users

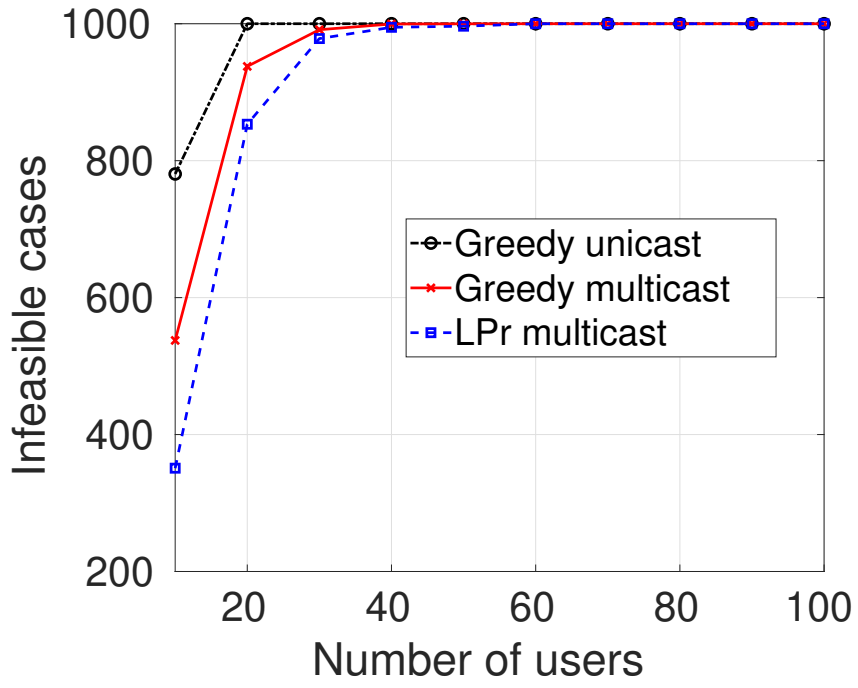


Figure 3.6: Number of infeasible cases under various resource allocation policies for $R = 2$ Mbps

only experience the delay incurred due to the core network. Therefore, our allocation policies do not introduce any QoE degradation. Figure 3.9 and Figure 3.10 show the histograms of the number of PRBs saved while transmitting the frames of the Star Wars IV video over the LTE multicast environment. The proposed resource allocation policies are able to meet the requirements of the video stream in far lesser number of resources than unicast transmission. Under greedy allocation, multicast always saves more than 70 PRBs in a sub-frame whereas unicast is never able to save more than 50 PRBs. Under LPr, multicast transmission nearly always saves 80 or more PRBs in a sub-frame and unicast is unable to save more than 50 PRBs.

These simulation results clearly establish the feasibility of the proposed grouping and resource allocation algorithms for use in multicast systems. The hybrid grouping policy provides a significant advantage over unicast and random grouping. The poor performance of random grouping reinforces the importance of proper grouping algorithms. It clearly shows that if users are randomly thrown together without taking their channel conditions into account, multicast transmission may not provide any advantage over unicast. Among the proposed resource allocation policies, LPr scheme does better than the greedy policy.

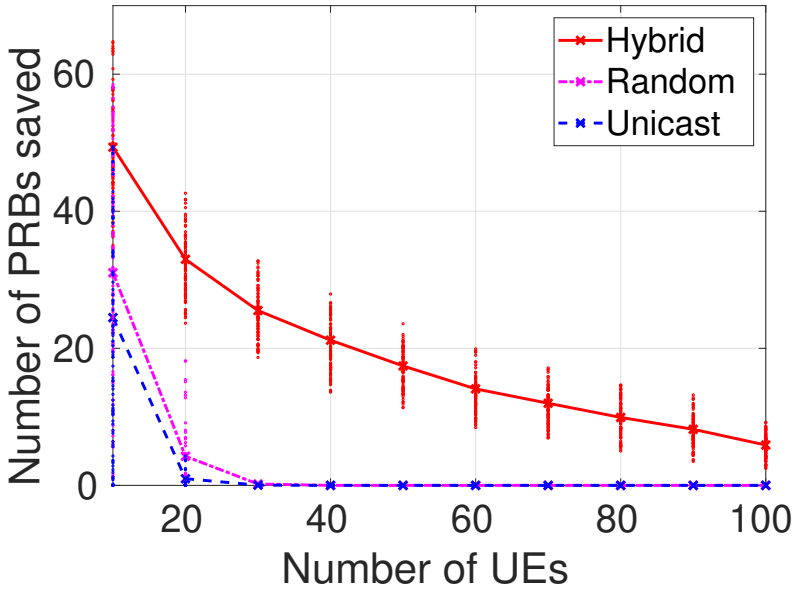


Figure 3.7: Number of PRBs saved for different UE placements under the Greedy scheme

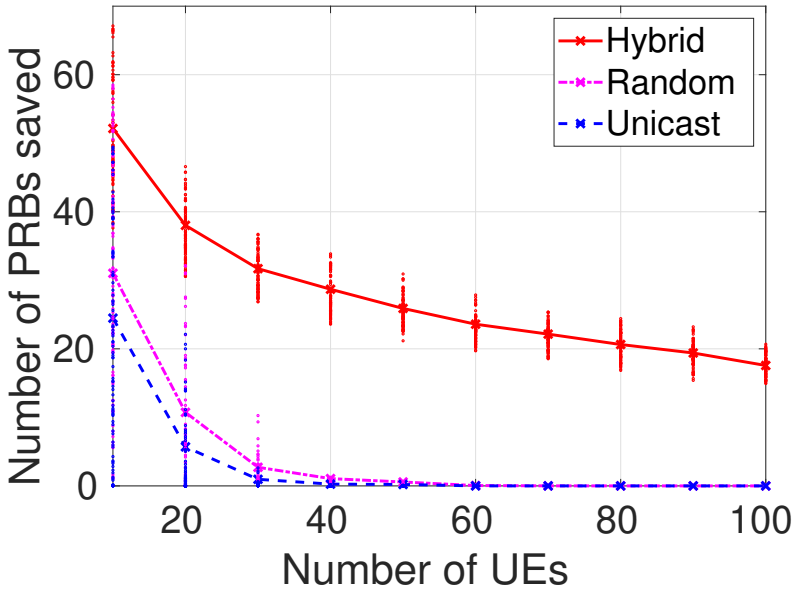


Figure 3.8: Number of PRBs saved for different UE placements under the LPr scheme

It satisfies the multicast users in a lesser number of PRBs and successfully meets their rate requirements in every sub-frame.

Next, we compare the performance of our resource allocation policies with the widely used Proportional Fair (PF) policy from the existing literature. Since we have already established that the LPr policy performs better of the two proposed policies, we only use

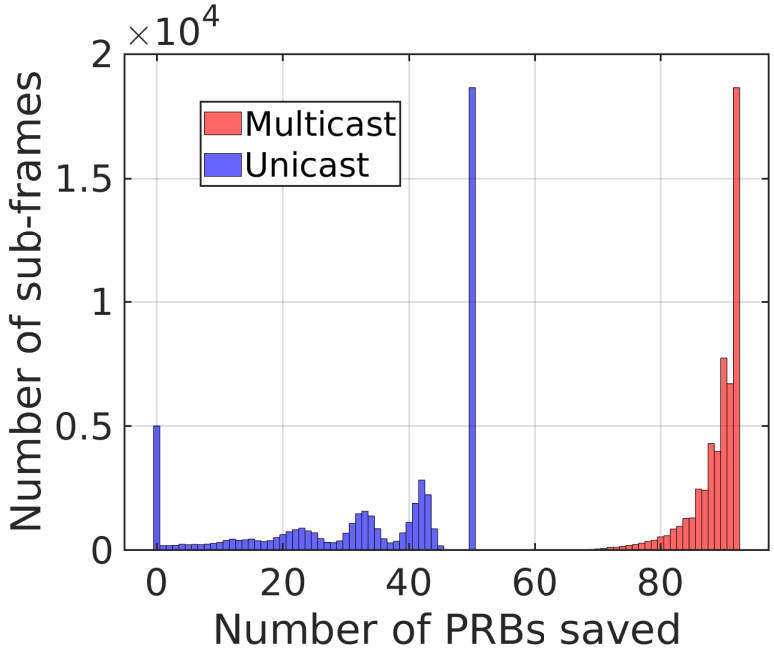


Figure 3.9: Histogram of number of PRBs saved for a real-time video stream under the Greedy scheme

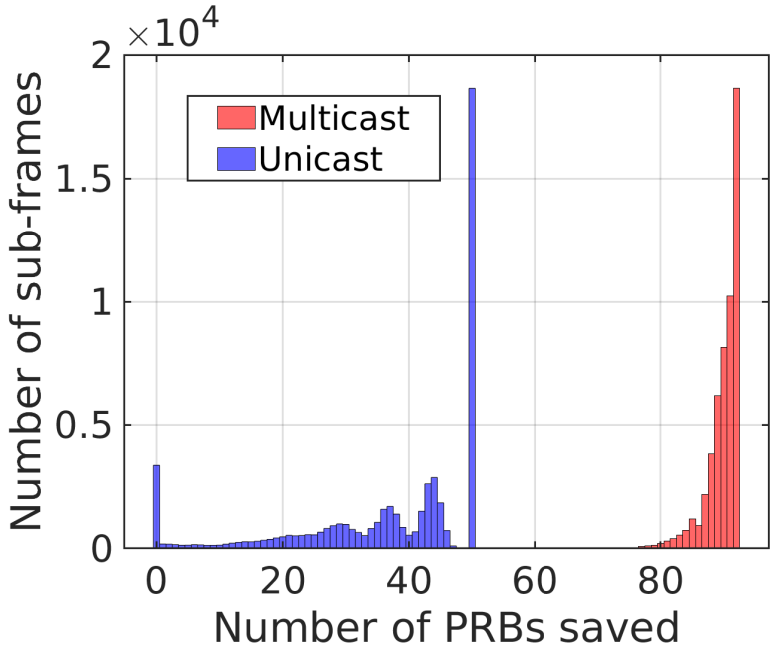


Figure 3.10: Histogram of number of PRBs saved for a real-time video stream under the LPr scheme

LPr for comparison in the next set of simulations.

For comparing our LPr policy to PF, we consider a scenario with both multicast

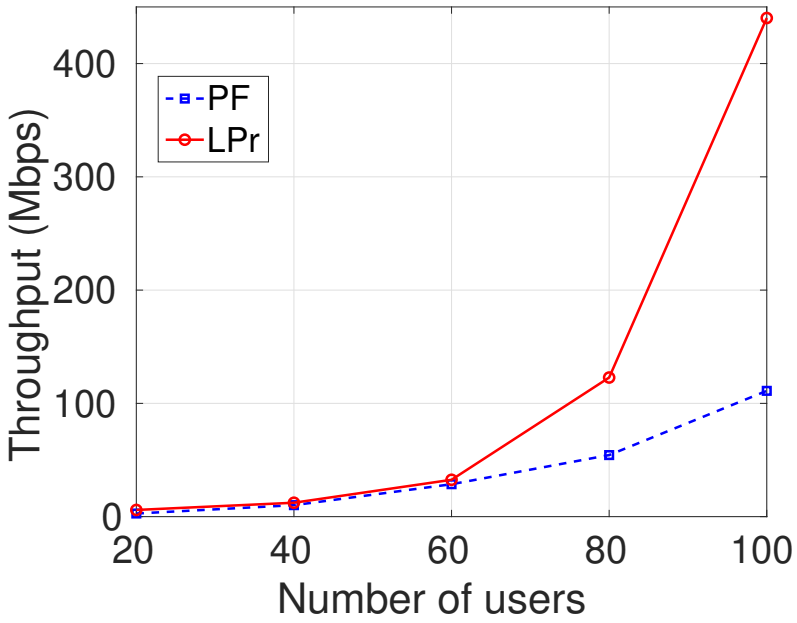


Figure 3.11: Comparison of the average system throughput under LPr and PF

and unicast UEs present in the system. We use the system throughput and user rate satisfaction as the metrics for comparison. In order to compare our scheme with PF, we first allocate the required number of PRBs to multicast groups using the proposed LP relaxation. Allocation to the unicast UEs is then done using the PF policy since the rate requirements of the unicast UEs are not fixed.

In Figure 3.11, we see the average sum throughout provided by LPr and PF allocation. As can be observed from the figure, LPr results in a significantly better system throughput. Even though it first uses a chunk of PRBs to satisfy the rate requirement of the multicast UEs, LPr policy still provides a better overall throughput. In Figure 3.12, we plot the percent of unsatisfied multicast groups as a function of the number of users in the system. The PF policy nearly always fails to meet the rate requirements of the multicast users. Even with no constraint on the amount of resources it can use, it fails to meet the rate requirements of the multicast groups and does not even do better in terms of the overall system throughput. On the other hand, LPr meets the requirements of all the multicast groups in minimum number of PRBs and also results in a significantly better system throughput.

These simulation results clearly indicate that suitability of the proposed policies for

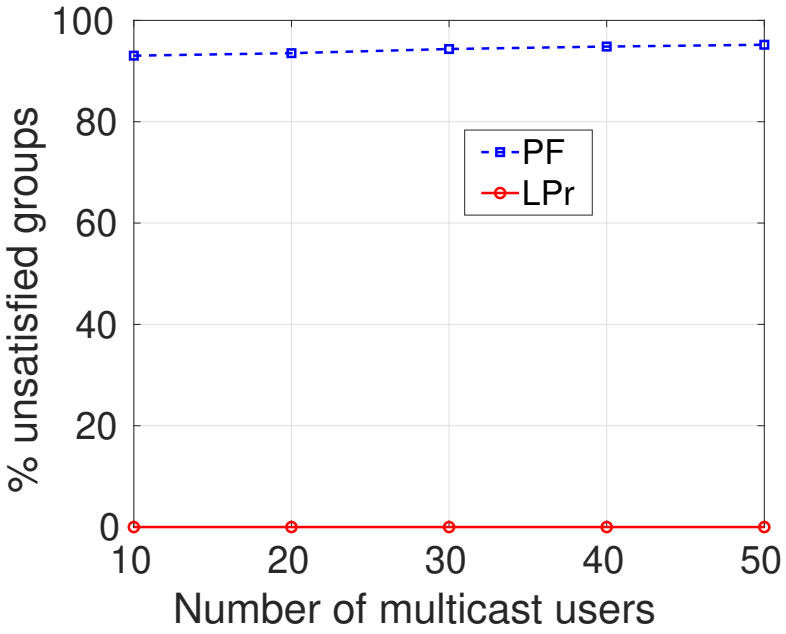


Figure 3.12: Comparison of the percent unsatisfied groups under LPr and PF

use in multicast transmissions. While PF schemes work well in a unicast only scenario, they are not suitable for rate constrained streaming systems that require a certain rate to be provided to the subscribers in every sub-frame. Grouping users according to the hybrid grouping policy and allocating resources using the LPr policy provides the best performance in terms of resource utilization and user satisfaction. In the next section, we discuss some generalizations of the proposed policies.

3.7 Generalizations

In this chapter, we have primarily focused on grouping and resource allocation for multicast streaming for non-layered video coding such as H.264/AVC. Each eMBMS service has a certain rate associated with it at which all the users subscribed to it are served. There are, however, several ways in which a streaming service may differ from the one considered in this chapter. For instance, the network might need to handle layered video coding, introduce rate adaptation or switch to the next generation 5G system. The proposed policies can be adapted for use in these and many more scenarios with little or no changes. We now discuss some of these generalizations in greater detail.

- *Heterogeneous quality demands*: The users subscribed to the same eMBMS service

may want to see different qualities of the same video stream. Some users may want ultra HD quality while others may prefer a lower quality video for a lesser price. This heterogeneity in user demands can be handled by treating the users who require the same quality as a separate group with a specific rate requirement. The proposed allocation policies can be used for allocating resources in such a system.

- *Rate adaptation*: The proposed policies can also be used in streaming systems with rate adaptation as long as the rate adaptation takes place on a slower time scale than a sub-frame (1 ms). When the rate requirements change, the grouping of users can be changed accordingly. As discussed in Section 3.1, we can allow for the groups to change every K sub-frames, where K is large. Even if rate adaptation occurs on the order of a few seconds, K would be of the order of a few 1000 sub-frames. Thus, the proposed policies can also be used in systems with rate adaptation.

- *SVC*: When SVC is used for encoding the streaming content, different sets of users may require a different number of enhancement layers of the video. The base layer, however, needs to be transmitted to all the users. The algorithms proposed in this chapter can then be used for transmitting the base layer to all the subscribed users and separate algorithms can then be used on top of these policies for opportunistically transmitting the enhancement layers to the groups with good channel conditions [54, 55].

- *5G*: Even though the policies proposed in this chapter have been discussed in the context of an LTE system, the policies can be used in next generation 5G systems as well. The proposed hybrid grouping policy makes use of the SNR to CQI mappings to define the grouping thresholds. Similar mappings are also defined in 5G [109] which can be similarly used to define the SNR thresholds. The proposed resource allocation policies are technology agnostic. The bandwidth in 5G is also divided into PRBs [110] and the proposed policies can be used to determine resource allocation for 5G systems without any changes.

Thus, the proposed grouping and resource allocation policies are suitable for use in a versatile range of systems and can also be easily adapted for use in the next generation 5G systems.

3.8 Conclusions

In this chapter, we have formulated the problems of grouping and resource allocation for multicast systems that require a certain rate to be provided to all the multicast users. These problems are aimed at satisfying the rate requirements of the users in minimum possible number of PRBs. We have proved that both, the optimal grouping problem and the optimal resource allocation problem, are NP-hard and therefore, no polynomial-time algorithms exist for determining their optimal solutions. We have designed a Randomized Scheme (RS) that works iteratively for estimating the optimal resource allocation. The output of the RS provides a benchmark for performance evaluation of heuristic resource allocation schemes. We have proposed two efficient heuristics for resource allocation, a greedy and an LP-relaxation based scheme. The LP-relaxation scheme results in feasible resource allocations that save nearly as many PRBs as that saved by using the RS in about one-fifth the time taken by the RS. We have proposed a heuristic scheme for multicast group formation as well. We call this grouping scheme as the hybrid grouping policy. It divides the users subscribed to a multicast service into groups based on their average SNR values such that the users in a group experience similar channel gains. Using extensive simulations, we have shown that using the proposed policies for grouping and resource allocation results in significant resource conservation. Therefore, using these policies for multicast streaming can help alleviate the burden on our network resources. In addition to the multicast system discussed here, the proposed resource allocation policies can be easily adapted for use in more general systems such as those with rate adaption and scalable video content streaming. The proposed grouping and resource allocation schemes can act as an enhancement to eMBMS. These enhancements will not only improve the performance of eMBMS but will also make its multicast operations more flexible and versatile.

The resource allocation policies proposed in this chapter are constrained to serve all the users in every sub-frame. This makes the performance of multicast groups dependent on the weakest user in the group. However, if the users can tolerate a certain amount of packet loss, the system will no longer be constrained to serve every user in every sub-frame allowing for more flexibility in resource allocation. In the next chapter, we address the problem of resource allocation in such a loss tolerant multicast system.

Chapter 4

Resource Allocation for Loss Tolerant Multicast Video Streaming

For lossless multicast streaming discussed in Chapter 3, resource allocation policies need to serve all users in each sub-frame. Hence, the streaming content for a multicast group cannot be transmitted at rates higher than what can be decoded by the weakest user in the group. This makes the system performance dependent upon the users experiencing the worst channel states. It also results in dissatisfaction of users with good channel states who can achieve much higher data rates with unicast transmissions. In this chapter, we address these issues using loss tolerant video streaming. Video streams can tolerate a certain amount of packet loss without any significant degradation in the quality of video perceived by the users. In fact, video streams are tolerant of packet losses as high as 40% [13]. For an H.264 encoded video, decoders like FFmpeg and JM can conceal as much as 39% packet loss with no observable deterioration of video quality [13]. This property can be leveraged to selectively allow some packet loss in video streams as long as the users receive the desired video quality. We model such a loss tolerant multicast system for video streaming and show that it significantly reduces the bandwidth consumption of video streams. The loss tolerant nature of video streaming has *not* been exploited for performance improvement of multicast in the existing literature.

In this chapter, we design efficient resource allocation algorithms for multicast video streaming that allow for some controlled packet losses, depending upon the video quality requirements of users. Allowing for some losses in a multicast stream gives us the flexibility

of not having to serve all users in a multicast group in each sub-frame. This means that the weakest user may no longer be a bottleneck. The transmission rates in some sub-frames can be higher than what can be decoded by the weakest user. This leads to increased system throughput and greater user satisfaction.

We model a loss tolerant MBMS system in which each user may have a different loss tolerance. We convert the problem of resource allocation in this system to the problem of stabilizing a virtual queueing system. We prove that stabilizing the token queues in this virtual queueing system is equivalent to satisfying the loss requirements of the users. For allocating resources in loss tolerant MBMS systems, we propose two online loss optimal policies that do not require any statistical information of the channel states of users for making allocation decisions. Channel states can vary arbitrarily and can also be correlated across users. The proposed policies are throughput optimal in the sense that they can stabilize the queueing system whenever any other policy, including offline policies with complete information of the channel states, can do so. A mechanism for a polynomial-time implementation of the proposed policies using Maximum Weight Bipartite Matching (MWBM) is also presented.

We evaluate the performance of the proposed policies using extensive simulations. Since the proposed policies are primarily designed for video streaming services, we make use of video traces from actual videos to simulate realistic video traffic patterns for the simulations. We use traces from five different videos obtained from the video trace library of Arizona State University (`http://trace.eas.asu.edu/`) [16, 17]. We compare the performance of our policies to that of the throughput optimal Exponential (Queue length) (EXP-Q) rule [18]. The EXP-Q rule proposed in [18], is designed for use in a single channel system where multiple flows are contending for obtaining the channel. We modify the EXP-Q rule for use in a more complicated multi-channel multicast system. Under the modified EXP-Q rule, several users encounter losses greater than their thresholds. On the other hand, our proposed policies successfully meet the loss requirements of all users. By allowing for some loss, we are able to satisfy the video quality requirements of a larger number of users.

The rest of this chapter is organized as follows. We discuss the system model and the problem formulation in Section 4.1. The construction of the queueing system and related

results are presented in Section 4.2. In Section 4.3, we present the proposed resource allocation algorithms and in Section 4.4, we discuss the polynomial-time implementation of these algorithms. The details of the simulations are given in Section 4.5 and we conclude in Section 4.6.

4.1 System Model and Problem Formulation

In this section, we first explain the system model in detail and then we define the resource allocation problem for a loss tolerant MBMS system.

4.1.1 System Model

Our system consists of an LTE cell with L different MBMS sessions. There are M UEs in the cell that can subscribe to any of these sessions. $[M]$ and $[L]$ denote the set of UEs and the set of multicast groups, respectively. UEs subscribed to the i^{th} video stream form multicast group G_i and we use $i(k)$ to denote the index of the group to which UE k belongs. The number of UEs in G_i is denoted by K_i. Each MBMS group is allocated one PRB in each sub-frame. A resource allocation policy Γ decides which PRB will be allocated to which group in each sub-frame. We define an allocation vector $\mathbf{B^{\Gamma}[t]}$ for policy Γ in sub-frame t. $\mathbf{B^{\Gamma}[t]}$ is a vector of length L that specifies which PRB, if any, has been assigned to each group. Note that Γ is completely defined by the value of $\mathbf{B^{\Gamma}[t]}$ in each sub-frame. We use $B_i^{\Gamma}[t]$ to denote the i^{th} entry of vector $\mathbf{B^{\Gamma}[t]}$. If G_i is not scheduled for reception in sub-frame t, then $B_i^{\Gamma}[t] = 0$, otherwise $B_i^{\Gamma}[t]$ takes the value of the PRB number allocated to G_i. The i^{th} MBMS service requires data to be transmitted to its subscribers at rate R_i. Whenever a PRB is allocated to multicast group G_i, data is transmitted in that PRB at the corresponding rate R_i. For each MBMS stream, a data packet arrives at the beginning of a sub-frame and is transmitted in the same sub-frame.

The channel states of UEs vary across time and frequency. As a result, the channel experienced by a UE varies from one sub-frame to another and also across PRBs within a sub-frame. Depending on the CQI of UE k in PRB j of sub-frame t, there is a certain maximum MCS that can be supported in that PRB for this UE [7] and a corresponding maximum rate, $r_{kj}[t]$ that it can successfully decode. As a result, a UE may not receive

the transmitted content successfully even after a PRB has been assigned to its multicast group. When a UE successfully receives data in a sub-frame, we say that the UE has been *'served'* in that sub-frame. Note that a UE being *scheduled* and being *served* is not the same. We distinguish between these two terms below:

- We say that a UE has been scheduled in a sub-frame if a PRB is allocated to its group in that sub-frame. For instance, UE $k \in G_i$ is said to have been scheduled for reception in sub-frame t under policy Γ if $B_i^{\Gamma}[t] \neq 0$.

- We say that a UE has been served in a sub-frame if it has been scheduled in that sub-frame and is able to successfully decode the received content. For instance, UE $k \in G_i$ is said to have been served in sub-frame t under Γ if $B_i^{\Gamma}[t] = j \neq 0$ and $R_i \leq r_{kj}[t]$.

We denote the loss encountered by UE k under policy Γ in sub-frame t by $\ell_k^{\Gamma}[t]$. For $k \in G_i$ and $B_i^{\Gamma}[t] = j \neq 0$, we have:

$$\ell_k^{\Gamma}[t] = \begin{cases} 0, & \text{if } R_{i(k)} \leq r_{kj}[t], \\ 1, & \text{otherwise.} \end{cases} \tag{4.1}$$

For $B_{i(k)}^{\Gamma}[t] = 0$, the UE is not scheduled for reception and so, $\ell_k^{\Gamma}[t] = 1$.

The video streams can tolerate a certain amount of packet loss without significant quality degradation. Each UE in the system has some loss tolerance depending upon its channel state and the video resolution chosen by it. A higher resolution would typically mean a lower loss tolerance and vice versa. We use $\tilde{\ell}_k$ to denote the fractional loss tolerable by UE k. $\tilde{\boldsymbol{\ell}} = [\tilde{\ell}_1, \ldots, \tilde{\ell}_M]$ denotes the loss tolerance vector for the system.

A compressed video stream is made up of Groups of Pictures (GoPs). A GoP comprises a series of Intra-coded (I), Predicted (P) and Bidirectional predicted (B) frames. I frames are self-contained and do not require other frames to be decoded. P frames are dependent on their preceding I frames for being correctly decoded, and B frames are dependent on both preceding and following I and/or P frames for being correctly decoded. Due to this, it is difficult to estimate the impact of loss of I and P frames on the video quality [16]. Therefore, we assume that all I and P frames of the videos are transmitted without any loss and we use loss tolerant streaming only for transmitting B frames.

For I and P frames, the eNB can allocate sufficient resources so that these frames are transmitted without any loss. We now formally define the resource allocation problem.

4.1.2 Problem Definition

We begin by stating some important definitions that will be used in formulating the problem.

Definition 3. *Feasible resource allocation: Resource allocation in a sub-frame is said to be feasible if it assigns at most one PRB to each multicast group such that no two groups are assigned the same PRB. In other words, a feasible resource allocation in sub-frame t corresponds to an allocation vector $\mathbf{B^{\Gamma}[t]}$ such that no two non-zero elements of the vector are equal i.e., if $B_i^{\Gamma}[t] \neq 0$, then $B_i^{\Gamma}[t] \neq B_{i'}^{\Gamma}[t]$ for every $i' \neq i$.*

Definition 4. *Feasible resource allocation policy: A feasible resource allocation policy Γ is a policy that chooses a feasible allocation vector in each sub-frame.*

A resource allocation policy can make use of the knowledge of current channel states of UEs, allocation information of the previous sub-frames, loss tolerance of UEs and losses encountered by UEs in the past to make allocation decisions in a sub-frame. It could even be an off-line policy that could make allocation decisions in advance with the knowledge of the channel states of users in all sub-frames including the ones in future.

Definition 5. *Average packet loss: We denote the average packet loss encountered by a UE k under resource allocation Γ by $\bar{\ell}_k^{\Gamma}$. It is the total packet loss per unit time and can be mathematically expressed as follows:*

$$\bar{\ell}_k^{\Gamma} = \lim \sup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \ell_k^{\Gamma}[t].$$

$\bar{\boldsymbol{\ell}}^{\mathbf{\Gamma}} = [\bar{\ell}_1^{\Gamma}, \ldots, \bar{\ell}_M^{\Gamma}]$ *denotes the system loss vector for policy* Γ.

Definition 6. *Feasible region of a policy: The feasible region of a resource allocation policy Γ, $\mathcal{L}^{\Gamma}$, is the set of all loss tolerance vectors, $\tilde{\boldsymbol{\ell}}$s that can be satisfied by Γ i.e., $\tilde{\boldsymbol{\ell}} > \bar{\boldsymbol{\ell}}^{\mathbf{\Gamma}}$ with probability (w.p.) 1.*

Definition 7. *Feasible region of the system: The feasible region of the system is the set of loss vectors $\mathcal{L} = \bigcup_{\Gamma} \mathcal{L}^{\Gamma}$ where the union is over all feasible Γ.*

Definition 8. *Optimal policy: The optimal resource allocation policy* $\Gamma^\star$ *is a policy whose feasible region is the set of loss vectors* $\mathcal{L}^{\Gamma^\star} = \bigcup_\Gamma \mathcal{L}^\Gamma$.

Our objective here is to determine the optimal resource allocation policy $\Gamma^\star$. We design this optimal policy using results from queueing theory. Towards this end, we convert the resource allocation problem in a loss tolerant MBMS system to the problem of stabilizing a virtual queueing system and prove that stabilizing the resulting system is equivalent to meeting the loss requirements of UEs.

4.2 Queueing System for Resource Allocation in Loss Tolerant MBMS Systems

We convert the problem of resource allocation in a loss tolerant MBMS system to the problem of obtaining a throughput optimal policy for a queueing system. Towards this end, we first discuss the construction of the queueing system.

4.2.1 Construction

We construct a virtual queueing system (Figure 4.1) with fictitious queues corresponding to each user. The length of the queue of a user is an indicator of how much loss a user has encountered. The state of this queueing system is completely described by the lengths of these queues. The arrival and departure processes of these queues is defined below. Note that, since these queues are virtual, no physical entities arrive or depart from the queues. Arrivals and departures merely represent an increase or a reduction of the queue lengths. We use the term *'token'* to refer to the virtual entities that make up these queues and refer to the queues as *'token queues'*.

The arrival process for the token queue of k^{th} UE is denoted by $\{\lambda_k[t]\}_{t\geq 1}$. $\lambda_k[t]$ is a binary random variable indicating arrival of a virtual token to k^{th} queue in sub-frame t, and has the expected value $\lambda_k = 1 - \tilde{\ell}_k$. This value of the expected arrival rate will mean that stabilizing this virtual queue will ensure that, in the actual system, UE k is served in more than $1 - \tilde{\ell}_k$ of the sub-frames. Arrivals across sub-frames are assumed to be independent and identically distributed. Across users, the arrival processes are assumed

to be independent. Let $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_M]$ denote the system arrival rate vector.

We define another indicator random variable $\mu_k^{\Gamma}[t]$ that indicates whether or not UE k has been served in sub-frame t under Γ. $\mu_k^{\Gamma}[t] = 1$ if and only if (iff) k is served in sub-frame t. Thus, a departure from the token queue corresponds to the successful delivery of a packet in the actual physical system. If $B_{i(k)}^{\Gamma}[t] = j$, then, $\mu_k^{\Gamma}[t] = 1$ iff $j \neq 0$ and $R_{i(k)} \leq r_{kj}[t]$. Otherwise, $\mu_k^{\Gamma}[t] = 0$. Let $Q_k^{\Gamma}[t]$ denote the length of queue k at the beginning of sub-frame t under Γ. Note that,

$$Q_k^{\Gamma}[t+1] = \max\{Q_k^{\Gamma}[t] + \lambda_k[t] - \mu_k^{\Gamma}[t], 0\}.$$

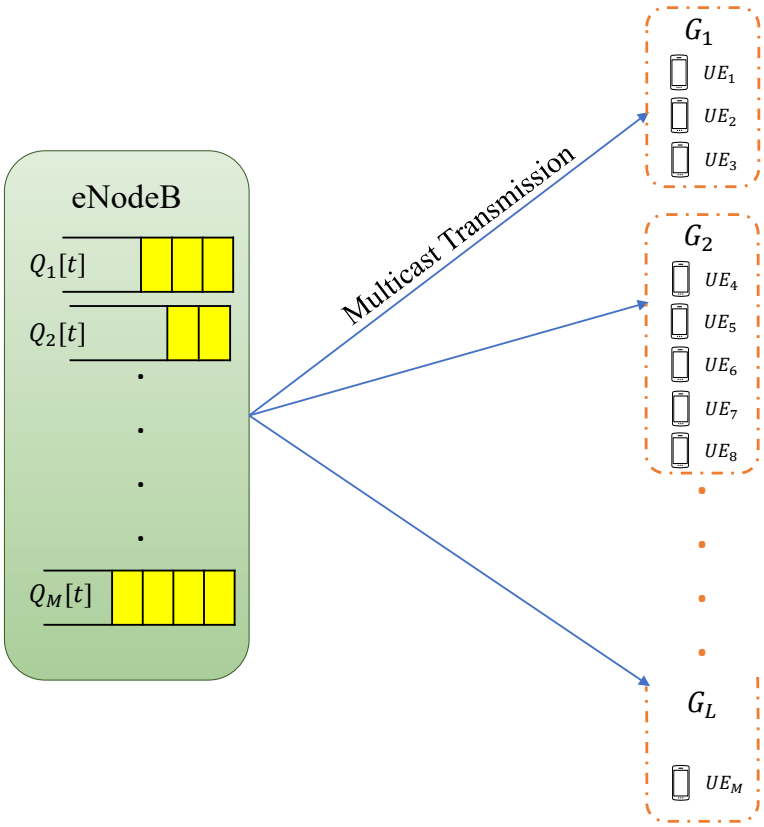


Figure 4.1: Virtual queueing system model

Now, the stability region of the queueing system thus constructed can be defined as follows:

Definition 9. *Stability region of the queueing system: The queueing system is said to be stable if the expected queue lengths stay finite for every queue i.e.,* $\sup_t \mathbb{E}[Q_k[t]] < \infty$ *for*

every k. Note that, for our corresponding physical system, this will mean that each UE k is served in at least $1-\tilde{\ell}_k$ of the sub-frames and hence, its loss requirements are met. A resource allocation policy that stabilizes the system is called a stable resource allocation policy. The stability region of a resource allocation policy Γ is the set of arrival rate vectors for which the system is stable under Γ. The stability region of the queueing system is the union of the stability regions of all feasible Γ's. We denote it as $\mathcal{S}$.

Definition 10. *Throughput optimality: A resource allocation policy Γ is said to be throughput optimal if Γ can stabilize the queueing system if some policy can do so. This means that if the queueing system is at all stabilizable, Γ will stabilize it.*

This virtual queueing system can be maintained at the eNB as illustrated in Figure 4.1. Since the eNB knows the loss requirements of UEs as well as their channel states, it has all the information needed to maintain the queueing system. In the next section, we examine the stability region of the constructed queueing system and relate it to the feasible region of the optimal resource allocation policy.

4.2.2 Feasible Region of the Optimal Resource Allocation Policy and Stability Region of the Queueing System

In this section, we prove that stabilizing the constructed queueing system is equivalent to meeting the loss requirements of the multicast UEs in the loss tolerant MBMS system. This is stated in the following result.

Theorem 4. *The loss requirement of a UE is met iff its token queue in the queueing system is stable. Therefore, the feasible region of the optimal allocation policy $\Gamma^\star$, $\mathcal{L}^{\Gamma^\star}$ is equivalent to the stability region of the queueing system $\mathcal{S}$ i.e., $\tilde{\boldsymbol{\ell}} \in \mathcal{L}^{\Gamma^\star}$ iff $(\mathbf{1}-\tilde{\boldsymbol{\ell}}) \in \mathcal{S}$. Here, $\mathbf{1}$ is a vector of ones of same size as $\tilde{\boldsymbol{\ell}}$.*

This will establish the equivalence of the stability region of the constructed queueing system and the feasible region of the optimal resource allocation policy. We now present a few results that are needed for proving Theorem 4. We begin by defining a few terms.

Define a set $\mathcal{B} = \{B_1, \ldots, B_{|\mathcal{B}|}\}$ containing all possible PRB allocation vectors to L groups. The cardinality of this set $|\mathcal{B}| = \binom{N}{L} \times L!$. In LTE, channel states are quantified

in terms of CQI values. According to 3GPP standards [7], a total of 15 CQI values are defined in LTE. Since the number of CQI values is finite, the possible channel states of UEs can take finitely many values. We define a set $\mathcal{C}$ that contains all possible channel state combinations of all UEs in the system. $\mathcal{C}$ will, therefore, be a set of 15^M CQI vectors, each of size M. Let g be the probability distribution over $\mathcal{C}$. That is, the channel state of the system in a sub-frame t, $C(t) = C$ w.p. $g(C)$. We denote by $\boldsymbol{\mu_{B_iC}}$, the vector of service rates of UEs corresponding to allocation B_i in CQI state $C \in \mathcal{C}$. Note that $\boldsymbol{\mu_{B_iC}}$'s are binary vectors of size M. We use $\boldsymbol{\mu_C} = \{\boldsymbol{\mu_{B_iC}}\}_{B_i \in \mathcal{B}}$ to denote the set of possible service rate vectors in channel state C. For a given $C \in \mathcal{C}$, define a distribution $\mathbf{w_C} = \{w_{B_iC}\}$ over the set of $\boldsymbol{\mu_{B_iC}}$'s where w_{B_iC} denotes the probability of choosing allocation B_i in channel state $C \in \mathcal{C}$. Using these definitions, we define the following Linear Program (LP):

$$
\begin{aligned}
LP(\delta): \quad & \sum_{C \in \mathcal{C}} \sum_{B_i \in \mathcal{B}} g(C) w_{B_iC} \boldsymbol{\mu_{B_iC}} = \boldsymbol{\lambda} + \delta, \\
& w_{B_iC} \geq 0 \ \forall \ B_i \in \mathcal{B}, \ C \in \mathcal{C}, \\
& \sum_{B_i \in \mathcal{B}} w_{B_iC} = 1, \ \forall \ C \in \mathcal{C},
\end{aligned}
$$

where δ is a non-negative real number. Note that $\{\mathbf{w_C}\}_{C \in \mathcal{C}}$ are the variables in this LP. Denote by $\Lambda(\delta)$ the set of arrival rate vectors $\boldsymbol{\lambda}$ such that the feasible region of $LP(\delta)$ is non-empty. Define two sets, $\Lambda^\circ = \bigcup_{\delta > 0} \Lambda(\delta)$ and $\overline{\Lambda} = \bigcup_{\delta \geq 0} \Lambda(\delta)$. In the next result, we establish the relation between Λ°, $\overline{\Lambda}$ and stability region of the queueing system $\mathcal{S}$. This result is essential for relating the feasible region of the optimal resource allocation policy to the stability region of the queueing system.

Theorem 5. $\Lambda^\circ \subseteq \mathcal{S} \subseteq \overline{\Lambda}$.

Proof. We begin by constructing the following randomized resource allocation policy Γ_δ based on $LP(\delta)$ defined in Section 4.2.2:

Definition 11. *Randomized allocation policy Γ_δ: Γ_δ chooses an allocation vector in a sub-frame according to a feasible solution $\mathbf{w_C}$ of $LP(\delta)$. If the system is in channel state C, Γ_δ chooses allocation vector B_i w.p. w_{B_iC} i.e., $P(\mathbf{B}^{\mathbf{\Gamma}_\delta}[\mathbf{t}] = B_i | C(t) = C) = w_{B_iC}$ $\forall$ t and decisions across sub-frames are independent. δ is an input parameter for Γ_δ.*

The definition of Γ_δ will be used for proving various results in this and the following sections. Consider $\boldsymbol{\lambda} \in \Lambda^\circ$. By the definition of Λ°, this means that, there exists $\delta > 0$ such that $LP(\delta)$ is feasible for arrival rate vector $\boldsymbol{\lambda}$. Let $\mathbf{w_C} = \{w_{B_iC}\}$ denote a feasible solution of $LP(\delta)$. Therefore, we can use policy Γ_δ to make scheduling decisions in each sub-frame according to $\mathbf{w_C}$. Let $A_k[t]$ denote the arrival process of queue k. $A_k[t] = 1$ if there is an arrival to queue k in sub-frame t and 0 otherwise. $D_k^{\Gamma_\delta}[t]$ denotes the departure process of k under Γ_δ. $D_k^{\Gamma_\delta}[t] = 1$ if a token departs from k under Γ_δ in sub-frame t and 0 otherwise. We have:

$$Q_k^{\Gamma_\delta}[t+1] = \max\{(Q_k^{\Gamma_\delta}[t] + A_k[t] - D_k^{\Gamma_\delta}[t]), 0\},$$

where $Q_k^{\Gamma_\delta}[t]$ is the length of the token queue of UE k at time t under policy Γ_δ. For simplicity of notation, we omit the Γ_δ superscript from $Q_k^{\Gamma_\delta}[t]$ and $D_k^{\Gamma_\delta}[t]$ through the rest of this section. Since a departure from queue k means that UE k has been successfully served, the corresponding service rate $\mu_k^{\Gamma_\delta}[t] = 1$ and we can write the above equation as:

$$Q_k[t+1] = \max\{(Q_k[t] + A_k[t] - \mu_k^{\Gamma_\delta}[t]), 0\}.$$

The state of the queueing system in a sub-frame can be completely defined by the queue lengths of all the token queues in that sub-frame. We denote the state of the system in sub-frame t by the vector $\mathbf{Q[t]} = [Q_1[t], \ldots, Q_M[t]]$. Since scheduling decisions made under Γ_δ only consider the current state of the system, the evolution of states of the system $\{\mathbf{Q[t]}\}_{\mathbf{t}\geq\mathbf{0}}$ under Γ_δ forms a Discrete Time Markov Chain (DTMC). This DTMC is countable, irreducible and aperiodic. We prove this in the following result.

Lemma 6. *The DTMC $\{\mathbf{Q[t]}\}_{\mathbf{t}\geq\mathbf{0}}$ is countable, irreducible and aperiodic.*

Proof. We prove these properties as follows:

- **Countable**: The state space of the DTMC is the set of all M-tuples $(Q_1[t], \ldots, Q_M[t])$ where $Q_k[t] \in \mathbb{N}$ ($\mathbb{N}$ denotes the set of natural numbers). It forms an M dimensional Cartesian product of $\mathbb{N}$ which is a countable set. Therefore, the state space of the DTMC and hence the DTMC itself is countable (by Theorem 2.13 in [111]).

- **Irreducible**: Let $\mathbf{Q}$ and $\mathbf{Q'}$ denote any two states of the DTMC. The DTMC can transition from any state $\mathbf{Q}$ to a state $\mathbf{Q'}$ in the following steps:

– Step 1: Schedule all UEs for service until all queues are empty. This is accomplished in $\max_k Q_k$ sub-frames.

– Step 2: For the next $\max_k Q'_k$ sub-frames, the token queue of UE k has an arrival and no departure for the first Q'_k sub-frames. In the remaining $(\max_k Q'_k - Q'_k)$ sub-frames, there is no new arrival and no departure in queue k. At the end of this step, the DTMC is in state $\mathbf{Q}'$.

These steps define at least one path of length $(\max_k Q_k + \max_k Q'_k)$ from any state $\mathbf{Q}$ to any other state $\mathbf{Q}'$. Therefore, the DTMC is irreducible.

- **Aperiodic**: If the DTMC is in state $\mathbf{Q[t]}$ and no new token arrives in any queue and no queue is scheduled for service in sub-frame t, the state of the DTMC remains unchanged. Therefore, self loops exist and the DTMC is aperiodic.

□

We now begin the proof of Theorem 5.

Proof. We prove Theorem 5 in two steps. We first establish that $\Lambda^\circ \subseteq \mathcal{S}$ in Lemma 7 and then show that $\mathcal{S} \subseteq \overline{\Lambda}$ in Lemma 8.

Lemma 7. *Every $\boldsymbol{\lambda} \in \Lambda^\circ$ is a stabilizable arrival rate vector. Hence, $\Lambda^\circ \subseteq \mathcal{S}$.*

Proof. To prove this, we first show using Foster's theorem [112] that DTMC $\{\mathbf{Q[t]}\}_{\mathbf{t}\geq\mathbf{0}}$ is positive recurrent and hence the queue lengths do not grow infinitely under Γ_δ. Using the Lyapunov function $f(\mathbf{Q[t]}) = \sum_{k=1}^{M} Q_k^2[t]$, we have:

$$\begin{aligned} &f(\mathbf{Q[t+1]}) - f(\mathbf{Q[t]}) \\ &\leq \sum_{k=1}^{M} [(A_k(t) - \mu_k^{\Gamma_\delta}[t])^2 + 2Q_k[t](A_k[t] - \mu_k^{\Gamma_\delta}[t])]. \end{aligned}$$

Hence,

$$\begin{aligned} &\mathbb{E}[(f(\mathbf{Q[t+1]}) - f(\mathbf{Q[t]}))|\mathbf{Q[t]}] \\ &\leq \mathbb{E}\left[(\textstyle\sum_{k=1}^{M}[(A_k(t) - \mu_k^{\Gamma_\delta}[t])^2 + 2Q_k[t](A_k[t] - \mu_k^{\Gamma_\delta}[t])])|\mathbf{Q[t]}\right], \\ &\leq M + 2\mathbb{E}\left[(\textstyle\sum_{k=1}^{M} Q_k[t]A_k[t] - \sum_{k=1}^{M} Q_k[t]\mu_k^{\Gamma_\delta}[t])|\mathbf{Q[t]}\right], \\ &\leq M + 2\textstyle\sum_{k=1}^{M} Q_k[t]\lambda_k - 2\sum_{k=1}^{M} Q_k[t]\mathbb{E}\left[\mu_k^{\Gamma_\delta}[t]|\mathbf{Q[t]}\right]. \end{aligned}$$

From $LP(\delta)$, we have $\mathbb{E}\left[\mu_k^{\Gamma_\delta}[t]|\mathbf{Q[t]}\right] = \lambda_k + \delta$. Therefore,

$$\begin{aligned}
&\mathbb{E}[(f(\mathbf{Q[t+1]}) - f(\mathbf{Q[t]}))|\mathbf{Q[t]}] \\
&\leq M + 2\sum_{k=1}^{M} Q_k[t]\lambda_k - 2\sum_{k=1}^{M} Q_k[t](\lambda_k + \delta), \\
&\leq M - 2\sum_{k=1}^{M} Q_k[t]\delta.
\end{aligned}$$

Defining set $\mathcal{A} = \{\mathbf{Q} : \sum_{k=1}^{M} Q_k \leq \frac{M+1}{2\delta}\}$, we have:

$$\mathbb{E}[(f(\mathbf{Q[t+1]}) - f(\mathbf{Q[t]}))|\mathbf{Q[t]}] < \begin{cases} -1, & \forall\ \mathbf{Q[t]} \notin \mathcal{A}, \\ \infty, & \text{otherwise.} \end{cases}$$

Thus, by Foster's theorem [112], the DTMC is positive recurrent so the expected queue lengths in the queueing system are finite. Therefore, Γ_δ stabilizes the system for arrival rate vector $\boldsymbol{\lambda} \in \Lambda^\circ$. Thus, $\boldsymbol{\lambda} \in \mathcal{S}$ which implies that $\Lambda^\circ \subseteq \mathcal{S}$. □

This proves the first part of our result. We now need to show that $\mathcal{S} \subseteq \overline{\Lambda}$. In the interest of simplicity of notation, we assume that under a policy Γ that stabilizes the system, the following limit exists w.p. 1.

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{B_i C}^{\Gamma}[t], \tag{4.2}$$

where $\mathbb{1}_{B_i C}^{\Gamma}[t]$ is an indicator random variable that is 1 if allocation vector B_i is chosen by Γ under channel state C in sub-frame t and zero otherwise. Now consider the following sets of sample paths:

- A_1 : the set of sample paths on which Strong Law of Large Numbers (SLLN) holds for the arrival rates i.e., $\frac{\sum_{i=1}^{t} \lambda_k[t]}{t} \to \lambda_k$ as $t \to \infty$, $\forall\ k$. This is a probability 1 set i.e., $P(A_1) = 1$.

- A_2 : set of sample paths on which $\frac{\sum_{i=1}^{t} \mathbb{1}_{\{C(t)=C\}}}{t} \to g(C)$ as $t \to \infty$, $\forall\ C$ (SLLN holds) where $\mathbb{1}_{\{C(t)=C\}}$ is an indicator random variable that is 1 if the channel state in sub-frame t is C and 0 otherwise. Since g is a probability distribution over the set of channel states $\mathcal{C}$, we have, $P(A_2) = 1$.

- A_3 : the set of sample paths on which service rate under Γ is $\geq \boldsymbol{\lambda}$. Since Γ stabilizes the system, we have $P(A_3) = 1$.

- A_4 : the set of sample paths over which the limit in (4.2) exists. Since we assume that this limit exists w.p. 1, $P(A_4) = 1$.

Since A_1, A_2, A_3, A_4 are probability 1 sets, their intersection,

$$A = \bigcap_{i=1}^{4} A_i \tag{4.3}$$

is also a probability 1 set. We refer to the sample paths belonging to this set A as *non-trivial sample paths*.

We now prove the second part of our result.

Lemma 8. *If $\boldsymbol{\lambda} \notin \overline{\Lambda}$, then $\boldsymbol{\lambda} \notin \mathcal{S}$. Thus, $\mathcal{S} \subseteq \overline{\Lambda}$.*

Proof. We prove this result using a contradiction. Let $\boldsymbol{\lambda} \notin \overline{\Lambda}$ be a stabilizable arrival rate vector i.e., $\boldsymbol{\lambda} \in \mathcal{S}$. Since $\boldsymbol{\lambda}$ is a stabilizable arrival rate vector, there exists some allocation policy Γ that can stabilize the system for arrival rate $\boldsymbol{\lambda}$.

We observe the scheduling decisions made by this Γ along a non-trivial sample path from the set A defined in (4.3). Let v_{B_iC} denote the fraction of time for which Γ chooses the allocation vector B_i in channel state C along such a sample path. Since Γ stabilizes the system, the rate of departures must equal the arrival rate in the system. Therefore:

$$\sum_{C \in \mathcal{C}} \sum_{B_i \in \mathcal{B}} g(C) v_{B_iC} \boldsymbol{\mu_{B_iC}} = \boldsymbol{\lambda},$$

$$\text{where } v_{B_iC} \geq 0 \ \forall \ B_i \in \mathcal{B}, \ C \in \mathcal{C},$$

$$\sum_{B_i \in \mathcal{B}} v_{B_iC} = 1 \ \forall \ C \in \mathcal{C}.$$

This implies that $\boldsymbol{v} = \{v_{B_iC}\}$ is a feasible solution of $LP(\delta)$ and that,

$$\boldsymbol{\lambda} \in \Lambda(0) \implies \boldsymbol{\lambda} \in \overline{\Lambda}, \tag{4.4}$$

which is a contradiction. Therefore, $\boldsymbol{\lambda} \notin \overline{\Lambda}$ is not stabilizable i.e., any stabilizable $\boldsymbol{\lambda}$ must be contained in $\overline{\Lambda}$. Hence, $\mathcal{S} \subseteq \overline{\Lambda}$. □

From Lemmas 7 and 8, we have, $\Lambda^\circ \subseteq \mathcal{S} \subseteq \overline{\Lambda}$, which is the required result. This concludes the proof. □

□

From here onwards, we consider Λ° to be the stability region of the queueing system. We now prove the main result of this section, stated in Theorem 4 above. We state the theorem again for ease of the reader.

Theorem 4. *The loss requirement of a UE is met iff its token queue in the queueing system is stable. Therefore, the feasible region of the optimal allocation policy* $\Gamma^\star$*,* $\mathcal{L}^{\Gamma^\star}$ *is equivalent to the stability region of the queueing system* $\mathcal{S}$ *i.e.,* $\tilde{\boldsymbol{\ell}} \in \mathcal{L}^{\Gamma^\star}$ *iff* $(\mathbf{1} - \tilde{\boldsymbol{\ell}}) \in \mathcal{S}$*. Here,* $\mathbf{1}$ *is a vector of ones of same size as* $\tilde{\boldsymbol{\ell}}$*.*

Proof. We need to show that the loss requirement of a UE is met iff its token queue is stable. We first argue that the stability of the queueing system implies that the loss requirements of UEs are met. If the queue corresponding to UE k is stable, it means that there exists a policy Γ that stabilizes the queue for $\boldsymbol{\lambda} \in \Lambda^\circ$. We can, therefore, construct a randomized policy Γ_δ as defined in Definition 11. Under Γ_δ, the rate of service for queue k is greater than λ_k which means that UE k is served in greater than $(1 - \tilde{\ell}_k)$ of the sub-frames. Therefore, the loss encountered by UE k is less than $\tilde{\ell}_k$ and its loss requirement is met.

Now, let us assume that the loss requirement of UE k is met. We show that this ensures the stability of its token queue. Since the loss requirement $\tilde{\ell}_k$ is achievable, there exists a policy Γ that satisfies the loss requirement. This means that, under Γ, the UE is served in greater than $(1 - \tilde{\ell}_k)$ fraction of sub-frames. Since the arrival rate $\lambda_k = (1 - \tilde{\ell}_k)$, the queue is served at a rate greater than the arrival rate. Hence, Γ stabilizes the token queue. From these arguments, we conclude that the loss requirement of a UE is met iff its corresponding token queue is stable. Therefore, the feasible region of the optimal allocation policy $\Gamma^\star$, $\mathcal{L}^{\Gamma^\star}$ is equivalent to the stability region of the queueing system, $\mathcal{S}$. □

We have now established that the stability region of the constructed queueing system is same as the feasible region of the optimal resource allocation policy $\Gamma^\star$. Therefore, here onwards, we do not explicitly consider meeting the loss requirements of UEs. Instead, we focus our attention on stabilizing the token queues corresponding to each UE knowing that stabilizing the token queues of UEs will ensure that their respective loss requirements are met.

4.3 Proposed Resource Allocation Algorithms

In this section, we propose online loss optimal policies for resource allocation in loss tolerant MBMS systems. We also present their efficient polynomial-time implementations in a later section.

4.3.1 Loss Optimal Resource Allocation

Loss Optimal Resource Allocation (LORA) makes scheduling decisions in a sub-frame t based on the token queue lengths of users $Q_k[t]$'s. Note that the queues being scheduled here are the fictitious queues of the constructed virtual queueing system. Scheduling of a token queue here is equivalent to the corresponding UE being served in the actual system. For ease of notation, we use Γ_0 to denote LORA in notations and equations[1]. In each sub-frame t, Γ_0 chooses service vector $\boldsymbol{\mu^{\Gamma_0}[t]}$ according to the following optimization problem.

$$\boldsymbol{\mu^{\Gamma_0}[t]} = \underset{\boldsymbol{\mu^{\Gamma_0}[t]} \in \boldsymbol{\mu_C}}{\arg\max} \sum_{k=1}^{M} Q_k[t] \mu_k^{\Gamma_0}[t], \tag{4.5}$$

where $\mu_k^{\Gamma_0}[t]$ is the service rate of UE k in sub-frame t under policy Γ_0. Γ_0 maximizes the sum of the queue lengths of UEs served in sub-frame t. This corresponds to those users being served in the actual system who have experienced the highest packet losses. It has a brute force computational complexity of $\mathcal{O}(M\binom{N}{L}L!)$. We have already established in Section 4.2 that stabilizing the token queues ensures that the loss requirements of UEs are met. Therefore, to prove that Γ_0 can successfully meet the loss requirements of the multicast UEs, it is sufficient to show that Γ_0 stabilizes the constructed queueing system. We prove this in the following result.

Theorem 6. *For any stabilizable arrival rate vector $\boldsymbol{\lambda}$, Γ_0 stabilizes the queueing system.*

This theorem implies that as long as the system is stabilizable, i.e., there exists some policy Γ that can stabilize the queueing system, so can Γ_0. Note that Γ is not restricted to using the same information that is available to Γ_0. Γ could be using information of the

[1]The names of policies (e.g., LORA) and their symbols (e.g., Γ_0) are used interchangeably throughout this chapter.

past and future allocations and channel conditions to make allocation decisions. Despite that, we claim that Γ_0 will successfully stabilize the system using only the knowledge of the current state of the queueing system to make the scheduling decisions.

Proof. Let $D_k^{\Gamma_0}[t]$ denote the departure process of queue k under Γ_0. We have:

$$Q_k^{\Gamma_0}[t+1] = \max\{(Q_k^{\Gamma_0}[t] + A_k[t] - D_k^{\Gamma_0}[t]), 0\},$$

where $Q_k^{\Gamma_0}[t]$ denotes the queue length of the token queue of k at time t under Γ_0. For the sake of simplicity of notation, we omit the Γ_0 superscript from $Q_k^{\Gamma_0}[t]$ and $D_k^{\Gamma_0}[t]$ through the rest of this section. Since a departure from queue k means that UE k was successfully served, the service rate $\mu_k^{\Gamma_0}[t] = 1$ and we can write the above equation as:

$$Q_k[t+1] = \max\{(Q_k[t] + A_k[t] - \mu_k^{\Gamma_0}[t]), 0\}.$$

The state of the queueing system is completely defined by the vector $\mathbf{Q[t]} = [Q_1[t], \ldots, Q_M[t]]$. The evolution of $\mathbf{Q[t]}$ forms a DTMC since the scheduling decisions made by Γ_0 in t are based solely on the state of the system in sub-frame t. The DTMC is countable, irreducible and aperiodic. The proof that the DTMC has these properties follows the same arguments as in Lemma 6 in Section 4.2.2. We now show using Foster's theorem [112] that this DTMC is positive recurrent and hence the token queues do not grow infinitely.

Using the Lyapunov function $f(\mathbf{Q[t]}) = \sum_{k=1}^{M} Q_k^2[t]$, we have:

$$\begin{aligned} &f(\mathbf{Q[t+1]}) - f(\mathbf{Q[t]}) \\ &= \sum_{k=1}^{M} [(A_k(t) - \mu_k^{\Gamma_0}[t])^2 + 2Q_k[t](A_k[t] - \mu_k^{\Gamma_0}[t])]. \end{aligned}$$

Hence,

$$\begin{aligned} &\mathbb{E}[(f(\mathbf{Q[t+1]}) - f(\mathbf{Q[t]}))|\mathbf{Q[t]}] \\ &= \mathbb{E}\left[(\sum_{k=1}^{M} [(A_k(t) - \mu_k^{\Gamma_0}[t])^2 + 2Q_k[t](A_k[t] - \mu_k^{\Gamma_0}[t])])|\mathbf{Q[t]}\right], \\ &\leq M + 2\sum_{k=1}^{M} Q_k[t]\lambda_k - 2\mathbb{E}\left[(\sum_{k=1}^{M} Q_k[t]\mu_k^{\Gamma_0}[t])|\mathbf{Q[t]}\right]. \end{aligned} \tag{4.6}$$

Let $\mu_k^{\Gamma_\delta}[t]$ denote the service rate for UE k in sub-frame t under the randomized policy Γ_δ (Definition 11). Then, from (4.5), we have:

$$\sum_{k=1}^{M} Q_k[t]\mu_k^{\Gamma_0}[t] \geq \sum_{k=1}^{M} Q_k[t]\mu_k^{\Gamma_\delta}[t]. \tag{4.7}$$

Therefore, from (4.6) and (4.7):

$$\begin{aligned}
&\mathbb{E}[(f(\mathbf{Q[t+1]}) - f(\mathbf{Q[t]}))|\mathbf{Q[t]}] \\
&\leq M + 2\sum_{k=1}^{M} Q_k[t]\lambda_k - 2\mathbb{E}\left[(\sum_{k=1}^{M} Q_k[t]\mu_k^{\Gamma_\delta}[t])|\mathbf{Q[t]}\right], \\
&\leq M + 2\sum_{k=1}^{M} Q_k[t]\lambda_k - 2\sum_{k=1}^{M} Q_k[t](\lambda_k + \delta), \\
&\leq M - 2\sum_{k=1}^{M} Q_k[t]\delta.
\end{aligned}$$

Now for set $\mathcal{A} = \{\mathbf{Q} : \sum_{k=1}^{M} Q_k \leq \frac{M+1}{2\delta}\}$, we have:

$$\mathbb{E}[(f(\mathbf{Q[t+1]}) - f(\mathbf{Q[t]}))|\mathbf{Q[t]}] < \begin{cases} -1, & \forall\ \mathbf{Q[t]} \notin \mathcal{A}, \\ \infty, & \text{otherwise.} \end{cases}$$

Thus, by Foster's theorem [112], the DTMC is positive recurrent which means that the expected queue lengths in the queueing system will be finite. Therefore, Γ_0 stabilizes the system and hence meets the loss requirements of UEs. □

We now have a loss optimal policy that meets the loss requirements of users by making allocation decisions based on the UE token queue lengths. However, in addition to the amount of packet loss in a video stream, we would also like to control the pattern in which these losses occur. Even if a user has a high tolerance for loss, we would like to avoid large number of consecutive packet losses. Starving users for a large number of consecutive sub-frames may lead to user dissatisfaction and result in users leaving the multicast session. Therefore, a loss tolerant resource allocation policy should also restrict the amount of consecutive packet losses encountered by a UE in addition to the long term average packet loss. We propose such a policy in the next section. This policy improves upon LORA by increasing the scheduling probability of a UE every time it is left unserved. This ensures that users do not remain unserved for long periods at a stretch which leads to better loss performance, reduced burstiness of packet losses, and improved user satisfaction.

4.3.2 Priority Loss Optimal Resource Allocation

Priority Loss Optimal Resource Allocation (p-LORA) also makes scheduling decisions in a sub-frame based on the queue lengths $Q_k[t]$'s in that sub-frame. However, in p-LORA, we use an additional priority vector to increase the probability of serving a previously unserved queue. Here also, the queues being scheduled are the fictitious queues of the constructed virtual queueing system. Scheduling of a token queue here is equivalent to the corresponding UE being served in the actual system. We use Γ_P to denote p-LORA in notations and equations. In every sub-frame t, Γ_P chooses service vector $\boldsymbol{\mu^{\Gamma_P}[t]}$ according the following optimization problem:

$$\boldsymbol{\mu^{\Gamma_P}[t]} = \underset{\boldsymbol{\mu^{\Gamma_P}[t]} \in \boldsymbol{\mu_C}}{\arg\max} \sum_{k=1}^{M} \left(Q_k[t] + (c_k[t]+1) \times s\right) \mu_k^{\Gamma_P}[t], \tag{4.8}$$

where $\mu_k^{\Gamma_P}[t]$ is the service rate of UE k in sub-frame t under Γ_P, $c_k[t]$ is the priority weight ascribed to the token queue of UE k, s is a positive constant. $c_k[t]$ is defined as:

$$c_k[t] = \begin{cases} 0, & \text{if } \mu_k[t-1] = 1, \\ \min(c_k[t-1]+1, \kappa), & \text{otherwise.} \end{cases}$$

κ is the maximum value that the priority weights can take. Also, $c_k[0] = 0, \forall\ k$. We use $\bar{c}[t] = [c_1[t], \ldots, c_M[t]]$ to denote the vector of priority weights of all the queues in sub-frame t. Since increasing $c_k[t]$ increases the contribution of UE k in (4.8), it is more likely to be served by the resource allocation policy. Γ_P has a brute force computational complexity of $\mathcal{O}(M\binom{N}{L}L!)$.

When using policy Γ_P for resource allocation, the state of the queueing system can be completely defined by the queue lengths of all the token queues and the value of the priority counter of each queue. We denote the state in sub-frame t under policy Γ_P by the vector $\mathbf{Q^{\Gamma_P}[t]} = [Q_1^{\Gamma_P}[t], \ldots, Q_M^{\Gamma_P}[t], \bar{c}[t]]$. Since scheduling decisions under Γ_P in a sub-frame are based only on the state of the system in that sub-frame, the evolution of states of the system form a DTMC. In the next result we prove that this DTMC is countable, irreducible and aperiodic.

Lemma 9. *The DTMC formed by the evolution of the states under* Γ_P $\mathbf{Q^{\Gamma_P}[t]} = [Q_1^{\Gamma_P}[t], \ldots, Q_M^{\Gamma_P}[t], \bar{c}[t]]$ *is countable, irreducible and aperiodic.*

Proof. For the sake of simplicity of notation, we omit the Γ_P superscript from the notations in this proof.

- **Countable**: The state of the DTMC $\mathbf{Q}[\mathbf{t}]$ comprises the queue lengths of M UEs and their priority weights. We have already shown in Lemma 6 that the state space of queue lengths $(Q_1[t], \ldots, Q_M[t])$ is a countable set. The state space of the priority weights of UEs is an M dimensional Cartesian product over the finite set $\{1, 2, \ldots, \kappa\}$ and is therefore a finite countable set (Theorem 2.13 in [111]). Therefore, the states of the DTMC $\mathbf{Q}[\mathbf{t}]$ form a $2M$ dimensional Cartesian product over two countable sets, the state space of queue lengths and the state space of priority weights. Therefore, the state space of the DTMC and hence the DTMC itself is countable (Theorem 2.13 of [111]).

- **Irreducible**: Consider that the DTMC is in state $\mathbf{Q} = \{Q_1, \ldots, Q_M, \bar{c}_k\}$. We will show that a finite length path exists from $\mathbf{Q}$ to any state $\mathbf{Q}' = \{Q'_1, \ldots, Q'_M, \bar{c'}_k\}$. The DTMC can transition from $\mathbf{Q}$ to $\mathbf{Q}'$ in the following steps:

  - *Step 1*: Schedule all UEs for service until all queues are empty. This is accomplished in $\max_k Q_k$ sub-frames.
  - *Step 2*: A new token arrives in every queue and no queue is scheduled for service for the next $\min_k Q'_k$ sub-frames. At the end of this step, all queue lengths are equal to $\min_k Q'_k$ and all priority weights are equal to $\min(\min_k Q'_k, \kappa)$.
  - *Step 3*: For the next $\max_k Q'_k - \min_k Q'_k$ sub-frames, UEs in $\arg\max_k Q'_k$ see one arrival and no departure. Every other UE k' see an arrival and no departure for the first $(Q'_{k'} - \min_k Q'_k)$ sub-frames and one arrival and one departure for the remaining $\max_k Q'_k - Q'_{k'}$ sub-frames. At the end of this step, the queue length of UE k is equal to Q'_k.
  - *Step 4*: In the next sub-frame, there is one arrival and one departure in every queue. This makes the priority weights all equal to 0 while the queue lengths remain unchanged.
  - *Step 5*: In the next $\max_k c'_k$ sub-frames, there is no arrival and no departure for UEs in $\arg\max_k c'_k$. For every other UE k', there is an arrival and a departure

in the first $(\max_k c'_k - c'_{k'})$ of these sub-frames and no arrival and no departure in the remaining $c'_{k'}$ sub-frames. At the end of this step, the DTMC is in the desired state $\mathbf{Q}'$.

This defines one finite length path from any state $\mathbf{Q}$ to any other state $\mathbf{Q}'$ of length $(\max_k Q_k + \max_k Q'_k + 1 + \max_k c'_k)$. Hence, the DTMC is irreducible.

- **Aperiodic**: Consider state $\mathbf{Q}[\mathbf{t}]$ where all queues are empty and all priority weights are 0. If there is one arrival in each queue in slot $(t+1)$ and every queue is served, the queues remain empty and the priority weights remain 0. Therefore, this state has a self loop and hence has period 1. Since we have already shown that the DTMC is irreducible, all states have period 1 because periodicity is a class property. Hence, the DTMC is aperiodic.

□

We now prove that Γ_P is throughput optimal, i.e., Γ_P will stabilize the queueing system if any other policy can do so.

Theorem 7. *For any stabilizable arrival rate vector $\boldsymbol{\lambda}$, Γ_P stabilizes the queueing system.*

Theorem 7 implies that if the queueing system under consideration is at all stabilizable, Γ_P will stabilize it.

Proof. Let $D_k^{\Gamma_P}[t]$ denote the departure process of queue k under Γ_P. We have:

$$Q_k^{\Gamma_P}[t+1] = \max\{(Q_k^{\Gamma_P}[t] + A_k[t] - D_k^{\Gamma_P}[t]), 0\},$$

where $Q_k^{\Gamma_P}[t]$ is the queue length of queue k at time t under Γ_P. For simplicity of notation, we omit the Γ_P superscript from $Q_k^{\Gamma_P}[t]$ and $D_k^{\Gamma_P}[t]$ in the rest of this section. Since a departure from queue k in sub-frame t means that $\mu_k^{\Gamma_P}[t] = 1$, we can write the above equation as:

$$Q_k[t+1] = \max\{(Q_k[t] + A_k[t] - \mu_k^{\Gamma_P}[t]), 0\}.$$

Under this policy, the evolution of the state of the queueing system $\mathbf{Q}[\mathbf{t}] = [Q_1[t], \ldots, Q_M[t], \bar{c}[t]]$ forms a DTMC. We have proved in Lemma 9 that this DTMC is countable, irreducible and aperiodic. We now show using Foster's theorem [112] that this DTMC is positive recurrent and hence the queues do not grow infinitely.

Using the following Lyapunov function $f(\mathbf{Q}[\mathbf{t}]) = \sum_{k=1}^{M} Q_k^2[t]$, we have:

$$\begin{aligned} &f(\mathbf{Q}[\mathbf{t}+\mathbf{1}]) - f(\mathbf{Q}[\mathbf{t}]) \\ &= \sum_{k=1}^{M} [(A_k(t) - \mu_k^{\Gamma_P}[t])^2 + 2Q_k[t](A_k[t] - \mu_k^{\Gamma_P}[t])]. \end{aligned}$$

Hence, as in (4.6), we have:

$$\begin{aligned} &\mathbb{E}[(f(\mathbf{Q}[\mathbf{t}+\mathbf{1}]) - f(\mathbf{Q}[\mathbf{t}]))|\mathbf{Q}[\mathbf{t}]] \\ &\leq M + 2\sum_{k=1}^{M} Q_k[t]\lambda_k - 2\mathbb{E}\left[(\sum_{k=1}^{M} Q_k[t]\mu_k^{\Gamma_P}[t])|\mathbf{Q}[\mathbf{t}]\right]. \end{aligned} \tag{4.9}$$

Let $\mu_k^{\Gamma_\delta}[t]$ denote the service rate for UE k in sub-frame t under the randomized policy Γ_δ. Then, from (4.8), we have:

$$\begin{aligned} &\sum_{k=1}^{M} \left(Q_k[t]\mu_k^{\Gamma_P}[t] + (c_k[t]+1)s\mu_k^{\Gamma_P}[t]\right) \\ &\geq \sum_{k=1}^{M} \left(Q_k[t]\mu_k^{\Gamma_\delta}[t] + (c_k[t]+1)s\mu_k^{\Gamma_\delta}[t]\right). \end{aligned} \tag{4.10}$$

Therefore, from (4.9) and (4.10):

$$\begin{aligned} &\mathbb{E}[(f(\mathbf{Q}[\mathbf{t}+\mathbf{1}]) - f(\mathbf{Q}[\mathbf{t}]))|\mathbf{Q}[\mathbf{t}]] \leq M + 2\textstyle\sum_{k=1}^{M} Q_k[t]\lambda_k \\ &-2\mathbb{E}\left[(\textstyle\sum_{k=1}^{M} Q_k[t]\mu_k^{\Gamma_\delta}[t] + (c_k[t]+1)(\mu_k^{\Gamma_\delta}[t] - \mu_k^{\Gamma_P}[t])s)|\mathbf{Q}[\mathbf{t}]\right], \\ &\leq M + 2\textstyle\sum_{k=1}^{M} Q_k[t]\lambda_k - 2\sum_{k=1}^{M} Q_k[t](\lambda_k + \delta) \\ &-2\mathbb{E}\left[(\textstyle\sum_{k=1}^{M} -(\kappa+1)s\right], \\ &\leq M - 2\textstyle\sum_{k=1}^{M} Q_k[t]\delta + 4Ms. \text{ (for } \kappa = 1) \end{aligned}$$

Defining set $\mathcal{A} = \{\mathbf{Q} : \sum_{k=1}^{M} Q_k \leq \frac{4Ms+M+1}{2\delta}\}$, we have:

$$\mathbb{E}[(f(\mathbf{Q}[\mathbf{t}+\mathbf{1}]) - f(\mathbf{Q}[\mathbf{t}]))|\mathbf{Q}[\mathbf{t}]] < \begin{cases} -1, & \forall\ \mathbf{Q}[\mathbf{t}] \notin \mathcal{A}, \\ \infty, & \text{otherwise.} \end{cases}$$

Thus, by Foster's theorem [112], the DTMC is positive recurrent meaning that the expected queue lengths in the queueing system will be finite. So, Γ_P stabilizes the system and hence meets the loss requirements of all UEs. □

In the next section, we discuss a generalization of the Exponential (Queue length) rule (EXP-Q) which was proposed in [18]. We use the EXP-Q rule as a benchmark for performance evaluation of our policies since it is a well known throughput optimal policy for scheduling multiple flows over a time varying wireless channel. The EXP-Q rule also minimizes the maximum delay encountered in the system [113]. The rule, however, considers that there is a single channel that can be used by one flow at a time. We generalize the EXP-Q rule for use with multicast transmission and with multiple time varying channels so that it can be used for scheduling in the system under consideration.

4.3.3 Modified Exponential (Queue length) Rule (Γ_E)

The EXP-Q rule is a throughput optimal policy [18] that schedules a single queue k in a time slot t such that:

$$k \in \arg\max_k \gamma_k \mu_k[t] \exp\left(\frac{a_k Q_k[t]}{\beta + [\bar{Q}[t]]^\eta}\right), \tag{4.11}$$

where $\mu_k[t]$ is the rate of service of queue k in sub-frame t, a_k, γ_k and η are constants and $\bar{Q}[t] = (1/N)\sum_k a_k Q_k[t]$. The EXP-Q rule is designed for use in a system where a single time varying channel is shared by multiple flows. We generalize the EXP-Q rule to include multicast transmission and multiple channels (in the form of PRBs) available for scheduling multiple multicast and unicast flows. In the existing form, the EXP-Q rule cannot be used for resource allocation in such a system. Therefore, we modify it as follows.

We use Γ_E to denote the modified EXP-Q rule. Since we have multiple channels available and multiple groups can be scheduled for service in a sub-frame, the policy has to determine an allocation vector $\mathbf{B}^{\mathbf{\Gamma_E}}[\mathbf{t}]$ instead of choosing a single entity to be scheduled in a sub-frame. As defined in Section 4.1, $\mathbf{B}^{\mathbf{\Gamma_E}}[\mathbf{t}]$ is a vector that specifies which PRB is allocated to which multicast group. We define Γ_E as the policy that chooses service vector $\boldsymbol{\mu}^{\boldsymbol{\Gamma_E}}[\boldsymbol{t}]$ according to the following optimization problem:

$$\boldsymbol{\mu}^{\boldsymbol{\Gamma_E}}[\boldsymbol{t}] = \arg\max_{\boldsymbol{\mu}^{\boldsymbol{\Gamma_E}}[\boldsymbol{t}] \in \boldsymbol{\mu_C}} \sum_{k=1}^{M} \gamma_k \mu_k^{\Gamma_E}[t] \exp\left(\frac{a_k Q_k[t]}{\beta + [\bar{Q}[t]]^\eta}\right), \tag{4.12}$$

where $\mu_k^{\Gamma_E}[t]$ is the service rate of UE k in sub-frame t under Γ_E. In (4.12), we sum the quantity defined in (4.11) over all users in the system. Γ_E chooses the service vector that

maximizes this quantity. It then determines the allocation vector $\mathbf{B}^{\mathbf{\Gamma_E}}[\mathbf{t}]$ corresponding to the service vector $\boldsymbol{\mu}^{\boldsymbol{\Gamma_E}}[\boldsymbol{t}]$. Note that the queue lengths in (4.12) correspond to the fictitious queues of the constructed virtual queueing system. Scheduling of a token queue here is equivalent to the corresponding UE being served in the actual system. The allocation vector $\mathbf{B}^{\mathbf{\Gamma_E}}[\mathbf{t}]$ gives the allocation for the actual system. Γ_E can also be used for joint allocation of resources to unicast and multicast transmissions. It has a brute force computational complexity of $\mathcal{O}(M\binom{N}{L}L!)$.

4.3.4 Computational Complexity

The resource allocation policies discussed in this section have a brute force computational complexity of $\mathcal{O}(M\binom{N}{L}L!)$. This makes them unsuitable for use in practical systems unless we design efficient mechanisms for their implementation. The policies discussed in this section can be implemented in polynomial-time using Maximum Weight Bipartite Matching (MWBM) [97] based algorithms. We discuss the details of this implementation in the next section. We first present the algorithm for implementing Γ_0 in detail. The same algorithm can be used for implementing Γ_P and Γ_E by replacing the edge weights of Γ_0 with those of Γ_P and Γ_E respectively.

4.4 Polynomial-time Implementation of LORA, p-LORA and Modified EXP-Q

We make use of MWBM for efficient polynomial-time implementations of the resource allocation policies proposed in Section 4.3. MWBM brings down the computational complexity of their implementation to $\mathcal{O}(NL^2)$. The policies can thus be implemented in polynomial-time. We begin with the construction of the underlying bipartite graph which is the same for all the policies except for the edge weights which change according to the policy under consideration. We discuss the implementation of Γ_0 in detail. The procedure and proof involved can be directly used for Γ_P and Γ_E with modified edge weights. The modifications for Γ_P and Γ_E are given at the end of this section.

We construct a bipartite graph $\mathcal{G} = (U, V, E)$ where vertex set U is the set of L multicast groups, vertex set V is the set of N PRBs, E is the set of edges connecting

nodes in U to nodes in V. We define the service rate of UE $k \in G_i$ in PRB j in sub-frame t as follows:

$$\nu_k^j[t] = \begin{cases} 0, & \text{if } R_i > r_{kj}[t] \\ 1, & \text{otherwise.} \end{cases}$$

The weight of an edge connecting vertex $i \in U$ to vertex $j \in V$, $w_i^j[t]$ is the sum of the products of the queue lengths of UEs in group G_i and their achievable service rates in PRB j in sub-frame t i.e., $w_i^j[t] = \sum_{k \in G_i} Q_k[t]\nu_k^j[t]$. The resulting bipartite graph is illustrated in Figure 4.2. A MWBM of $\mathcal{G}$ that matches every node in U to a unique node from V results in an allocation equivalent to Γ_0. We prove this in the following result.

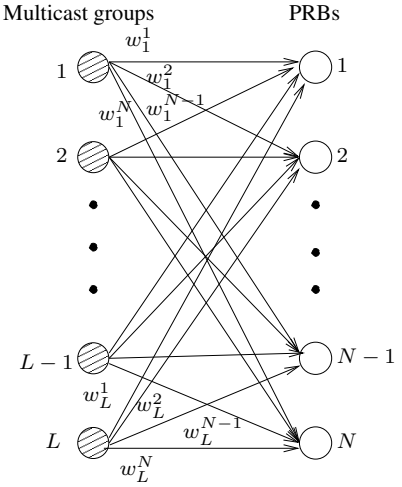


Figure 4.2: Bipartite graph between multicast groups and PRBs

Lemma 10. *Maximum weight bipartite matching for graph $\mathcal{G}$ results in resource allocation according to policy Γ_0.*

Proof. A matching for graph $\mathcal{G}$ selects edges that share no common vertices. This means that each group from U will be matched to exactly one PRB from V and each PRB from V will be matched to at most one group from U. Therefore, the requirement of assigning no more than 1 PRB to each group is satisfied. Since PRBs in V are matched to no more than one group from U, we have $B_i^\Gamma[t] \neq B_{i'}^\Gamma[t]\ \forall\ \{i, i' \in [L] : B_i^\Gamma[t], B_{i'}^\Gamma[t] \neq 0\}$ as required by Definition 12. Thus, the solution of the MWBM gives us a feasible resource allocation. Next, we show that the resulting allocation is consistent with the allocation decisions made by policy Γ_0.

MWBM picks edges such that the sum of the weights of the edges chosen is maximized. Therefore, it maximizes the quantity $\sum_{i\in U}\sum_{k\in G_i} Q_k[t]\nu_k^j[t] = \sum_{k=1}^{M} Q_k[t]\mu_k^{\Gamma_0}[t]$ which is same as in (4.5). Hence, resource allocation done using MWBM on $\mathcal{G}$ is consistent with policy Γ_0. □

The same algorithm can be used for implementing Γ_P and Γ_E by changing the edge weights. For Γ_P we have:

$$w_i^j[t] = \sum_{k\in G_i} \left(Q_k[t] + (c_k[t]+1)\times s\right)\nu_k^j[t]. \tag{4.13}$$

For Γ_E the edge weights are:

$$w_i^j[t] = \sum_{k\in G_i} \gamma_k \nu_k^j[t] \exp\left(\frac{a_k Q_k[t]}{\beta + [\bar{Q}[t]]^\eta}\right). \tag{4.14}$$

Proof similar to that of Lemma 10 follows to show that the MWBM for graph $\mathcal{G}$ with the edge weights defined in (4.13) and (4.14) results in the implementation of policies Γ_P and Γ_E respectively.

In the next section, we present the results of the simulations performed for evaluating the performance of the resource allocation algorithms proposed in this work.

4.5 Simulation Results

We study the performance of the proposed allocation algorithms in an LTE MBMS system. We simulate an LTE cell with L different MBMS video streams and UEs distributed uniformly at random throughout the cell. All UEs are subscribed to one of the L streams. UEs subscribed to the same MBMS service receive the relevant content on common PRBs. We use the MATLAB [103] based LTE simulator designed in [104] for these simulations. The channels and 3GPP mappings used are the same as specified in Section 3.6. Other relevant simulation parameters are given in Table 4.1.

Each multicast service has a certain rate requirement and each UE can tolerate some amount of packet loss. The loss tolerable by a UE depends on the quality of video required by it and its channel conditions. PRBs are allocated to the multicast groups according to the resource allocation policies discussed in Section 4.3. We observe the packet loss

Table 4.1: System Simulation parameters [1]

Parameters	**Values**
System bandwidth	20 MHz
eNB cell radius	150 m
Path loss model	$\mathrm{L} = 128.1 + 37.6 \log 10(d)$, d in kilometers
Lognormal shadowing	Log Normal Fading with 10 dB standard deviation
White noise power density	-174 dBm/Hz
eNB noise figure	5 dB
eNB transmit power	46 dBm
PRB width	180 kHz
Number of PRBs	100 per sub-frame

encountered by UEs and compare the performance of the proposed schemes with that of the modified EXP-Q rule [18]. In rest of this section, we refer to the modified EXP-Q rule as simply the EXP-Q rule for brevity. It should be noted that it is the modified EXP-Q rule defined in Section 4.3.3 that is used in all the simulations.

Since the proposed policies are primarily meant for use with video streaming services, we use traces from actual videos to generate traffic for these simulations. We have used video traces of five different videos in these simulations. The video traces have been obtained from the video trace library available at `http://trace.eas.asu.edu/` [16, 17]. The videos used are those of Silence of the Lambs, Star Wars IV, Tokyo Olympics, NBC News and Sony Demo. The videos are all H.264/AVC encoded with a GoP size of 16 with 15 B frames in each group.

As discussed in Section 4.1.1, I and P frames are needed for decoding other frames in a GoP. Therefore, we ensure that all I and P frames of the videos are transmitted without any loss and we use lossy allocation policies only for sending the B frames. This is a recommended practice in network simulations with video traces [16] since it is difficult to estimate the impact of loss of I and P frames on the video quality [16]. For sending

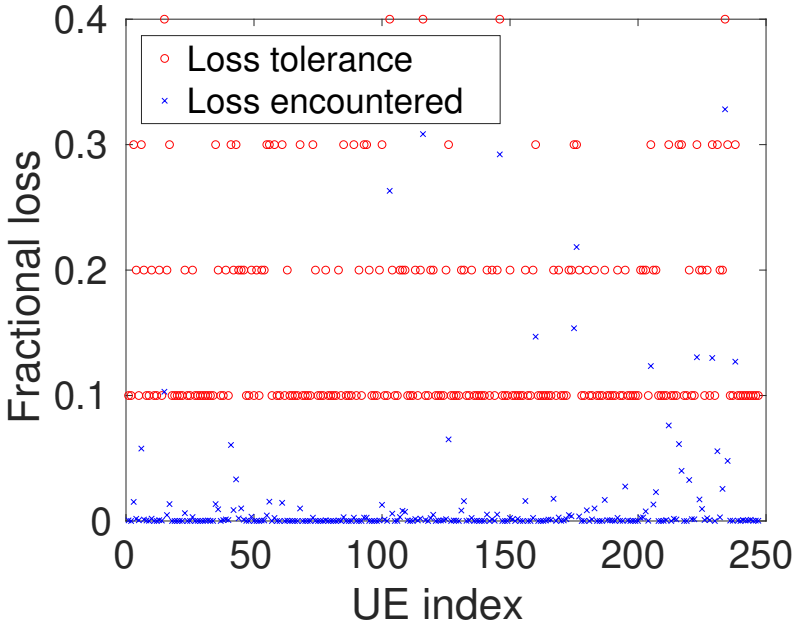


Figure 4.3: Tolerable loss versus loss encountered using LORA

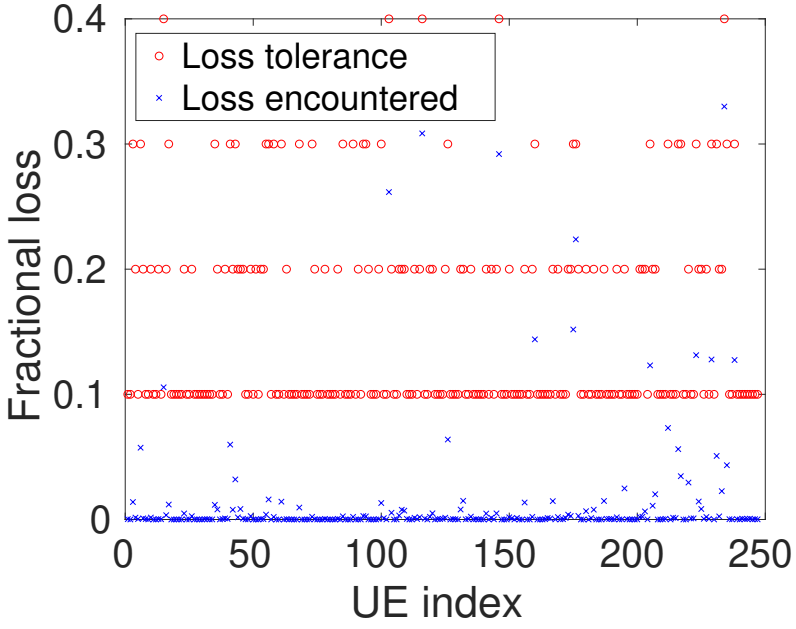


Figure 4.4: Tolerable loss versus loss encountered using p-LORA

I and P frames, we allocate sufficient resources to the groups and transmit at the rate corresponding to the weakest UE to ensure that those frames are successfully received by all users.

Figures 4.3, 4.4 and 4.5 compare the losses encountered by UEs to their respective

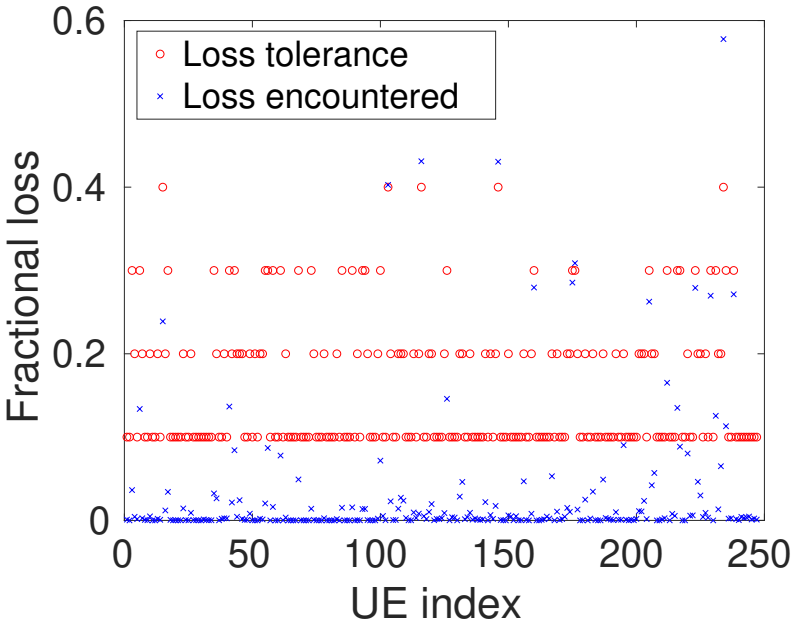


Figure 4.5: Tolerable loss versus loss encountered using EXP-Q

loss tolerances. For these plots, we run the simulations for the entire duration of all 5 videos and then average the results. Figures 4.3 and 4.4 illustrate this comparison for LORA and p-LORA respectively. Both these policies succeed in meeting the loss requirements of all UEs in the system. The virtual queueing system corresponding to the actual loss tolerant MBMS system is, thus, stable under both the proposed loss optimal policies. Figure 4.5 plots the losses encountered under the modified EXP-Q rule. We observe that several UEs experience losses significantly greater than their tolerable limits and the queueing system is rendered unstable.

Figure 4.6 compares the plots of the average losses encountered by UEs under the three policies. For this, the losses encountered per second have been exponentially averaged for a UE. Every point in the plot is then obtained by averaging over all UEs. We observe that the EXP-Q rule results in a better loss performance than LORA. Even though the EXP-Q rule results in better average loss performance than LORA, it fails to meet the loss requirements of several UEs. On the other hand, despite a greater average system packet loss, LORA is able to meet the loss requirements of all UEs. p-LORA leads to the least average packet loss among the three policies.

Figure 4.7 illustrates the Peak Signal-to-Noise Ratio (PSNR) degradation encountered by each video under the three policies. PSNR degradation is calculated as the

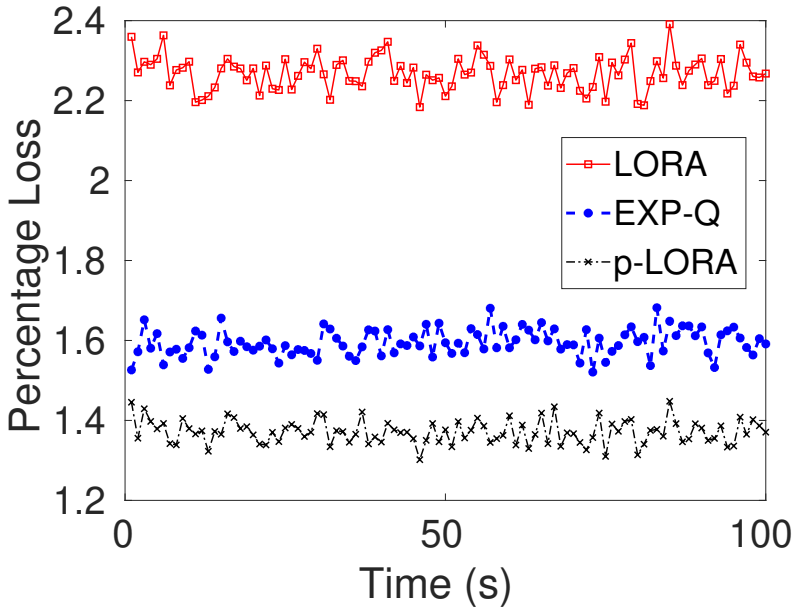


Figure 4.6: Comparison of average losses in LORA, EXP-Q and p-LORA schemes

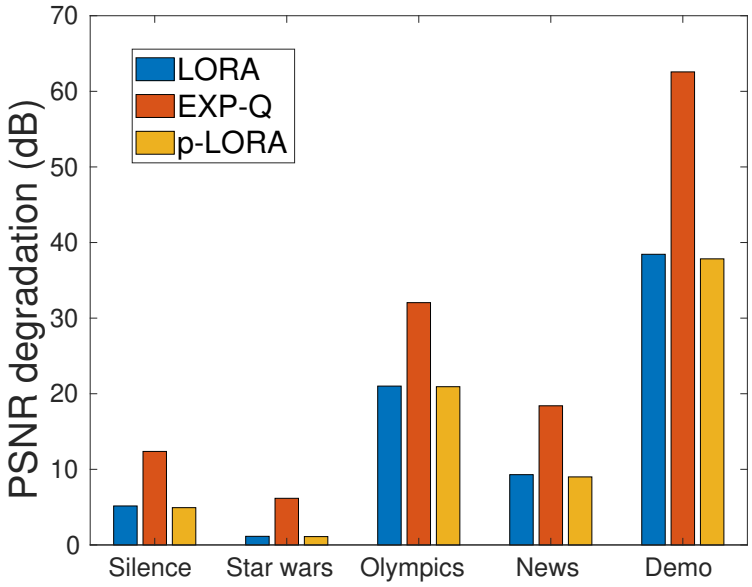


Figure 4.7: PSNR degradation of different videos

difference between the PSNR of the actual video and that of the received video. We observe that the EXP-Q rule results in the highest degradation in PSNR. LORA and p-LORA result in a significantly less loss in PSNR of the received video streams.

As discussed in Section 4.3, in addition to the amount of packet loss, the patterns in which the losses occur also have a major bearing on the users' experiences. While

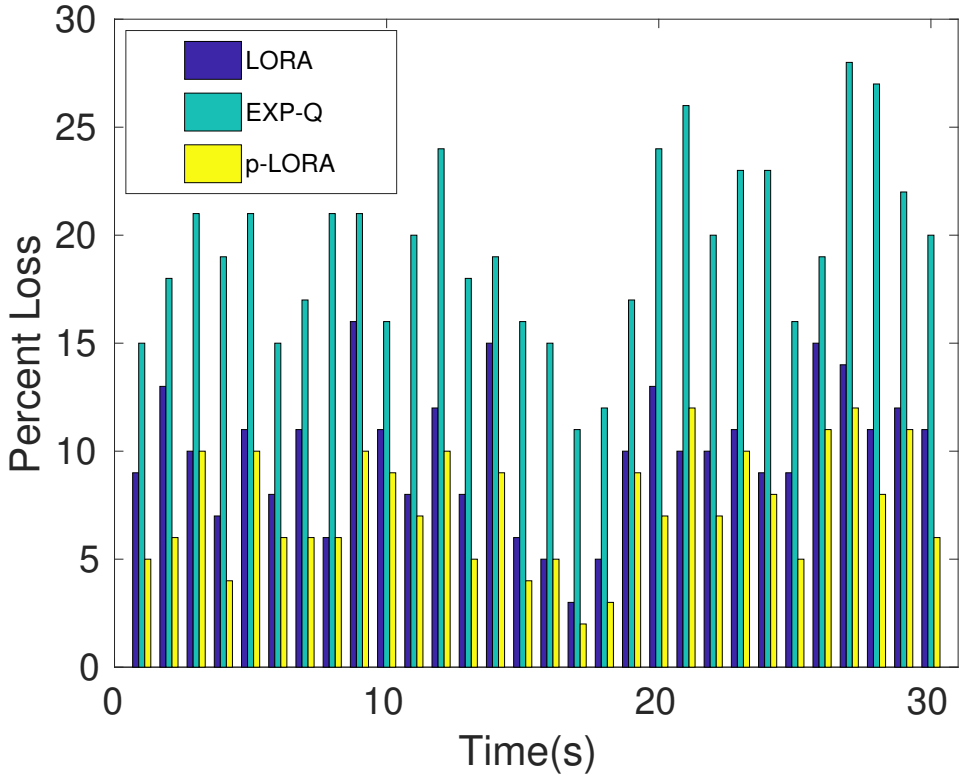


Figure 4.8: Loss pattern of a UE (losses per sub-frames)

some amount of packet loss spread more or less uniformly through a session may lead to no degradation in quality at all, a concentrated packet loss can be extremely annoying in a video stream and may even lead to UEs quitting the MBMS session. To observe the temporal pattern of packet loss encountered under the three policies, we plot the percentage packet loss pattern of a user with high loss tolerance, under heavy traffic conditions. This is plotted for all the policies as a function of time in Figure 4.8. The EXP-Q rule results in the most variable loss pattern. The losses per second see jumps as high as 10% from one second to another. LORA does better than the EXP-Q rule. However, p-LORA provides the most uniform loss pattern of the three policies. It, therefore, controls the burstiness of the losses encountered and ensures that no UE is starved for long periods at a stretch.

These simulation results clearly establish the effectiveness and superiority of the proposed loss optimal policies. The use of traces of actual videos further strengthens the

case for using loss tolerant allocation policies for streaming video content.

4.6 Conclusions

Video streams can tolerate a certain amount of packet loss without affecting the quality perceived by the end users. In this chapter, we leverage this property to improve the performance of wireless multicast video streaming. We consider an MBMS system where users can tolerate a certain amount of packet loss depending on various factors, such as the type of video stream and the channel quality experienced by them. We address the problem of resource allocation in such a system. We construct a fictitious virtual queueing system to represent the actual loss tolerant MBMS system. We convert the problem of determining the optimal resource allocation policy for the said system to the problem of stabilizing the constructed virtual queueing system. We propose two loss optimal policies, namely, LORA and p-LORA, for resource allocation in loss tolerant multicast video streaming systems. Since the proposed policies are computationally expensive to implement, we propose an MWBM based algorithm that provides an efficient polynomial-time implementation of the proposed policies.

We also modify the EXP-Q rule [18] for use in multicast transmission with multiple channels. The EXP-Q rule does scheduling of virtual queues of the constructed queueing system which is translated to resource allocation for the actual system. The EXP-Q rule is a known throughput optimal policy, and we use its modified version as a benchmark for evaluating the performance of our policies. We perform extensive simulations to study and compare the performance of LORA, p-LORA, and modified EXP-Q policies. To generate realistic video traffic patterns, we use video traces from actual video streams [16] in our simulations. Simulation results indicate that among these policies, p-LORA results in the least packet loss and the best PSNR of the delivered video streams. Using this policy for streaming video content in wireless multicast systems can significantly improve the system performance and reduce resource utilization of video streaming services.

The policies proposed in this chapter are specifically designed to cater to a loss tolerant video streaming system. In practice, a system may need to handle several different kinds of services simultaneously. While a set of streaming services may be loss tolerant, a high priority service running in parallel may have entirely different service requirements.

Therefore, there is a need for allocation mechanisms that can cater to a heterogeneous variety of services. In the next chapter, we design such a generalized resource allocation mechanism.

Chapter 5

Resource Allocation and Pricing for Heterogeneous User Demands

In Chapters 3 and 4, we have addressed resource allocation problems in a lossless and a loss tolerant multicast system, respectively. For a lossless multicast system, we have designed resource allocation policies that meet the rate requirements of users in a minimum possible number of resources. For a loss tolerant multicast system, we have proposed allocation policies which ensure that packet losses encountered by users stay within some acceptable thresholds. Like most of the literature on resource allocation, all these policies are designed around a specific objective. In practice, the objective of resource allocation may be governed by several different criteria because the base station handles a variety of user traffic simultaneously. The base station may be catering to multicast video streams, unicast transmissions, voice calls, video calls, browsing requests, some high priority traffic, and best-effort services all at once. All these applications may have different QoS requirements. Therefore, resource allocation algorithms designed for meeting one specific objective or for a specific type of service cannot cater to such a heterogeneous mix of users and services. In the current literature on resource allocation, there is no generalized resource allocation algorithm that can be used for simultaneously allocating resources to all types of users and services with varying QoS requirements.

In this chapter, we design a generalized auction based resource allocation algorithm that addresses the issues discussed above. The proposed algorithm provides a unified mechanism for allocating resources to different kinds of services simultaneously. It can be

used irrespective of the parameter being optimized by the allocation policy. Additionally, the algorithm also provides a method to determine the prices to be paid by users according to the QoS experienced by them. The proposed algorithm is equipped to handle a combination of multicast and unicast traffic as well as traffic with different priorities and QoS requirements simultaneously.

The proposed algorithm is based on the Vickrey-Clarke-Groves (VCG) [114] mechanism. It makes the allocation decisions using the bids conveyed by the interested users. The bids of users represent their eagerness for being scheduled. We prove that the proposed algorithm is strategy-proof, i.e., it can successfully elicit the true valuations of the resources from the users. This is an essential property for an auction based resource allocation mechanism since it ensures that users only bid according to their actual requirements so that the decisions made by the policy are socially optimal. The proposed algorithm is suited for use in any cellular mobile system and can be used for simultaneous resource allocation to all kinds of traffic. We also propose a computationally efficient implementation of the proposed mechanism that enables its polynomial-time implementation. Through extensive simulations in an LTE environment, we show the effectiveness of the proposed algorithm for handling a heterogeneous mix of users.

The rest of this chapter is organized as follows. We discuss the system model and formulate the problem in Sections 5.1 and 5.2, respectively. The proposed mechanism for resource allocation is presented in Section 5.3. The polynomial-time implementation of the mechanism is discussed in Section 5.4 and the simulation results are discussed in Section 5.5. Section 5.6 concludes the chapter.

Remark 2. *Throughout this chapter, we discuss the resource allocation problem in the context of an LTE system. However, the problem and the proposed algorithm are general and applicable to a wireless mobile communication system irrespective of the technology being used.*

5.1 System Model

Th system model is analogous to that of Chapter 4 except that we also consider unicast users here, in addition to multicast groups. The system comprises an LTE cell with M UEs

and S MBMS services that the UEs can subscribe to. UEs either subscribe to one of the MBMS services or receive unicast service. The UEs subscribed to an MBMS service form a multicast group that is considered as a single entity for resource allocation. We denote by L the total number of entities inclusive of all unicast UEs and multicast groups. Without loss of generality, we will refer to all the entities as MBMS groups/services, keeping in mind that a unicast UE is simply an MBMS group containing just one UE. The i^{th} MBMS group is denoted by G_i, and we use $i(k)$ to denote the index of a group to which UE k belongs. $[M]$ and $[L]$ denote the set of UEs and the set of MBMS groups, respectively. Each group has an associated rate of transmission at which the UEs belonging to it need to be served. R_i denotes the required rate corresponding to G_i. $r_{kj}[t]$ is the maximum rate achievable by UE k in PRB j in sub-frame t. Say UE k belongs to G_i and PRB j is allocated to this group in sub-frame t, then, data will be transmitted in j at rate R_i and UE k can successfully receive this data only if $r_{kj}[t] \geq R_i$. Accordingly, we define the loss encountered by UE k in PRB j in sub-frame t as:

$$\ell_{kj}[t] = \begin{cases} 0, & \text{if } R_{i(k)} \leq r_{kj}[t], \\ 1, & \text{otherwise.} \end{cases} \tag{5.1}$$

Each UE in the system has a certain valuation for being scheduled for service in a sub-frame. We use $v_k[t]$ to denote this valuation for UE k in sub-frame t. The valuation captures the resource requirement of a UE which could be a function of any number of factors like the data plan of a UE, the quality of video it requires, or the amount of packet loss it has encountered in the past. The valuation of a UE is its private information and is unknown to the eNB and other UEs in the system. In this model, we assume no structure and place no restrictions whatsoever on what the valuations can be. Thus, the proposed algorithm and the results presented in this chapter are independent of the nature of the UE valuations. Note that even though the system comprises multicast groups, valuations are individually calculated by all the users. A unique valuation is not defined for a multicast group. In the next section, we discuss the problem formulation.

5.2 Problem Formulation

We seek to determine an auction based resource allocation policy that is capable of satisfying the service requirements of a heterogeneous mix of unicast users and multicast groups based on their valuations. There are two main challenges in designing such a policy

1. Since the valuations of users are unknown to the eNB, it has to rely on the reported valuations for making the allocation decisions. Malicious users can report false valuations to bias the policy and hog resources. This may degrade the system performance and result in starvation of other users. It is therefore essential that the resource allocation policy successfully elicit the true valuations from the UEs.

2. The second challenge arises due to the existence of multicast groups. Valuations are calculated by users individually and not as a single value that represents the multicast group. Therefore, the policy has to make allocation decisions based on the valuations of individual UEs but allocate the same PRB to the entire group. However, because of distinct channel conditions of the UEs in a group, some of them may not receive the transmitted content successfully. Despite this, the policy should be able to meet the QoS requirements of all the users.

Before stating the problem, we define few essential terms and recall some notations. Recall that each group is allocated one PRB in a sub-frame. We denote a resource allocation policy by Γ and define an allocation vector of length L, $\mathbf{A^{\Gamma}[t]}$ that contains the identities of the PRBs allocated to each group by Γ in sub-frame t. For instance, if its first element $A_1^{\Gamma}[t] = 2$, it means that PRB 2 has been assigned to G_1 in sub-frame t. We also define an allocation indicator random variable $x_{ij}^{\Gamma}[t]$ that indicates whether or not PRB j has been assigned to G_i in sub-frame t under Γ. So,

$$x_{ij}^{\Gamma}[t] = \begin{cases} 1, & \text{if } A_i^{\Gamma}[t] = j, \\ 0, & \text{otherwise.} \end{cases}$$

Definition 12. ***Feasible resource allocation:*** *Resource allocation in a sub-frame is said to be feasible if it assigns at most one PRB to each multicast group such that no two groups are assigned the same PRB. In other words, a feasible resource allocation in sub-frame t corresponds to an allocation vector $\mathbf{A^{\Gamma}[t]}$ such that no two elements in it are equal, i.e., $A_i^{\Gamma}[t] \neq A_{i'}^{\Gamma}[t]$ for every $i' \neq i$.*

In this chapter, we aim to design an auction based resource allocation policy. We assume that each UE k communicates a bid value $b_k[t]$ to the eNB at the beginning of sub-frame t. The resource allocation in a sub-frame is performed based on the bids received in that sub-frame. In addition to resource allocation, the eNB also calculates the prices that the users have to pay for the services. We assume that the users are rational and selfish. Thus, they may report a bid value which is not same as their true valuations if doing so benefits them. These concepts are formalized in the following definitions.

Definition 13. ***Auction based resource allocation policy:*** *An auction based resource allocation policy* Γ *takes the bids of UEs (*$b_k[t]$*'s) as input and outputs a feasible allocation vector* $\mathbf{A^{\Gamma}[t]}$ *and prices to be paid by the UEs (*$p_k^{\Gamma}[t]$*'s) in every sub-frame* t*.*

Definition 14. ***Utility of a UE:*** *The utility of UE* k *in sub-frame* t *under policy* Γ*,* $u_k^{\Gamma}[t]$ *is defined as the difference between the valuation of the UE and the price* $p_k^{\Gamma}[t]$ *it pays for being served in that sub-frame i.e.* $u_k^{\Gamma}[t] = v_k[t] - p_k^{\Gamma}[t]$*. If a UE is not scheduled for reception in sub-frame* t*, its utility in that sub-frame is* 0*.*

Definition 15. ***Social utility:*** *The social utility of the system in sub-frame* t *under policy* Γ*,* $V^{\Gamma}[t]$ *is defined as the sum of the valuations of the UEs scheduled for service by* Γ *in that sub-frame. Using the definition of* $\ell_{kj}[t]$ *and* $x_{ij}^{\Gamma}[t]$*, we can write* $V^{\Gamma}[t] = \sum_k v_k[t] \sum_j x_{i(k)j}^{\Gamma}[t](1 - \ell_{kj}[t])$*.*

Equipped with these definitions, we can now define the problem as follows.

Let Λ *denote the set of all possible resource allocation policies. Our aim is to determine the optimal auction based resource allocation policy* $\Gamma^{\star} \in \Lambda$ *that provides a feasible, social utility maximizing resource allocation in every sub-frame.*

In the next section we propose such a resource allocation policy.

5.3 A VCG Based Mechanism for Generalized Resource Allocation and Pricing

The VCG mechanism [114] is a form of sealed bid auction mechanism that maximizes the social utility of the system. It takes the bids of buyers as input and allocates items to the highest bidders. The price paid by the winning bidders is equal to the 'damage' caused

by them to the rest of the bidders. We explain how this 'damage' is calculated later on this section. It is a known result that in VCG mechanism, bidding of the buyers' true valuations is a dominant strategy [115]. This means that rational participants have no incentive to not report their true valuations of the items. These features make the VCG mechanism suitable for resource allocation. However, in most cases, implementing a VCG mechanism is NP-hard. We now discuss the proposed VCG based resource allocation mechanism.

All the allocations and pricing calculations take place on a sub-frame basis. Therefore, we fix a sub-frame t and eliminate it from the notations in the rest of this chapter for the sake of notational simplicity. Consider PRBs in a sub-frame to be commodities that UEs want to acquire. UEs act as bidders who have certain valuations for acquiring these commodities. Since each group is allotted one PRB in a sub-frame, it follows that each UE can acquire at most one PRB. Also, since all the UEs in an MBMS group are to be served on the same PRB, our system is further bound to allocate the same commodity to all the UEs that belong to the same multicast group. The objective of the VCG mechanism here, is to determine a feasible allocation in each sub-frame that maximizes the sum of winning bids subject to these allocation constraints.

Recall that the valuation of UE k for obtaining a PRB is v_k and the bid submitted by it is denoted by b_k. The VCG mechanism chooses an allocation that maximizes the sum of winning bids given by:

$$\sum_{k}\sum_{j} b_k x^{\Gamma}_{i(k)j}(1-\ell_{kj}). \tag{5.2}$$

Let $\{x^{\Gamma_{-k}}_{ij}\}$ be the allocation indicators under policy Γ in the absence of UE k. Then, the price paid by UE k for service (i.e., the damage caused by it to the other bidders) is:

$$p^{\Gamma}_{k} = \sum_{q}\sum_{j} b_q x^{\Gamma_{-k}}_{i(q)j}(1-\ell_{qj}) - \sum_{q\neq k}\sum_{j} b_q x^{\Gamma}_{i(q)j}(1-\ell_{qj}).$$

The utility of the UE under this allocation is $u^{\Gamma}_{k} = v_k - p^{\Gamma}_{k}$.

The proposed allocation mechanism $\Gamma^{\star}$ works as follows:

1. **Step 1:** The UEs report their bids, b_k's to the eNB at the beginning of the sub-frame.

2. **Step 2:** The eNB determines the allocation vector $\mathbf{A}^{\mathbf{\Gamma}}$ and the corresponding x_{ij}^{Γ}s that maximize the quantity $\sum_k \sum_j b_k x_{i(k)j}^{\Gamma}(1-\ell_{kj})$ and allocates PRBs accordingly.

3. **Step 3:** The price to be paid by UE k, p_k^{Γ} is calculated for every k. These are periodically transmitted to and stored at the Policy and Charging Rules Function (PCRF) for charging purposes.

In the VCG mechanism, there is no incentive for a bidder to misrepresent its valuation since the utility gained by reporting a false valuation is never greater than that achieved by reporting the actual valuation. Therefore, the system can allocate items in an optimal manner without malicious users hogging resources by misrepresenting their requirements. This property is referred to as 'strategy-proofness' of the mechanism. It is this property that makes VCG a social utility maximizing mechanism. Since all bidders are forced to bid their true valuations, maximizing the sum of winning bids is equivalent to maximizing the utility of the system. The strategy-proofness of $\Gamma^{\star}$, however, does not obviously follow from the strategy-proofness of the conventional VCG mechanism due to the structure of the resource allocation problem under consideration. Here, a single commodity is allocated to an entire group of bidders, some of whom may still gain no utility whatsoever. So, the strategy-proofness of $\Gamma^{\star}$ needs to be proved. We do this in the following result.

Theorem 8. $\Gamma^{\star}$ *is strategy-proof.*

Proof. Since we will be dealing with policy $\Gamma^{\star}$ throughout this proof, we drop $\Gamma^{\star}$ from the notations for simplicity e.g., the allocation vector will simply be denoted by $\mathbf{A}$ instead of $\mathbf{A}^{\mathbf{\Gamma}^{\star}}$.

Consider a UE k with its true valuation being v_k. Let B denote the sum of winning bids under $\Gamma^{\star}$ when all UEs report their true valuations and let $\mathbf{A}$ be the corresponding allocation vector. We use B_{-k} to denote the sum of winning bids in the absence of UE k. Then, the price paid by k if it is scheduled under $\Gamma^{\star}$ is $p_k = B_{-k} - (B - v_k)$ and its utility is $u_k = B - B_{-k}$. If k is not scheduled, then $u_k = p_k = 0$. If k reports its true valuation, it either gets scheduled by $\Gamma^{\star}$ or it doesn't. We look at both these cases separately.

***Case 1*:** UE k gets scheduled by reporting $b_k = v_k$ truthfully. Now, let us say that it reports bid b'_k instead. Then, one of the following cases arise:

- $\mathbf{b'_k > v_k}$: If UE k bids a value greater than its valuation, it should continue to be scheduled. Suppose that this is not the case and k is not scheduled when it bids b'_k. Let B' be the sum of winning bids in this case, the rest of the bids being same as for B. Since $b'_k > v_k$, $b'_k - b_k = \delta > 0$. If allocation vector $\mathbf{A}$ is used in this scenario, k will be scheduled and the resulting sum of winning bids will be $B'' = B' + b'_k > B'$ which is a contradiction since $\Gamma^\star$ maximizes the sum of winning bids. Therefore, UE k will be scheduled when it bids $b'_k > v_k$ resulting in sum of winning bids $B' = B + \delta$. Let us now look at the utility obtained by it.

  The price paid by k in this case will be $p'_k = B_{-k} - (B' - b'_k) = p_k$ and its utility is $u'_k = v_k - (B_{-k} - (B' - b'_k)) = B' - B_{-k} - \delta = B - B_{-k} = u_k$. Since the utility gained by UE k remains unchanged, it has no incentive in reporting b'_k instead of b_k.

- $\mathbf{b'_k < v_k}$ and k is not scheduled: This is a trivial case since $u'_k = 0 < u_k$. The utility gained by the UE is reduced and so, there is no incentive for it to bid b'_k instead of b_k.

- $\mathbf{b'_k < v_k}$ and k is still scheduled: Since $b'_k < v_k$, $b_k - b'_k = \delta > 0$. Note that $B' = B - \delta$. Here, $p'_k = B_{-k} - (B' - b'_k) = p_k$ and its utility is $u'_k = v_k - (B_{-k} - (B' - b'_k)) = B' - B_{-k} + \delta = B - B_{-k} = u_k$. Since the utility gained by UE k remains unchanged, it has no incentive in bidding b'_k instead of b_k.

***Case 2*:** UE k does not get scheduled by reporting $b_k = v_k$ truthfully. In this case $u_k = 0$. Now, let us say that it bids b'_k instead. One of the following cases arise:

- $\mathbf{b'_k > v_k}$ and k is still not scheduled: This is a trivial case since $u'_k = 0 = u_k$. Here also, the utility gained by UE k remains unchanged and there is no incentive for it in bidding b'_k instead of b_k.

- $\mathbf{b'_k > v_k}$ and k is scheduled: Since $b'_k > v_k$, $b'_k - b_k = \delta > 0$. Note that $B' - B \leq \delta$. The price paid by k, $p'_k = B_{-k} - (B' - b'_k) = B - (B' - b'_k) \geq b'_k - \delta = v_k$. Therefore, its utility is $u'_k = v_k - p'_k \leq 0$. Since UE k does not gain any additional utility, it has no incentive in bidding b'_k instead of b_k even if it does get scheduled.

- $\mathbf{b'_k} < \mathbf{v_k}$: If UE k bids a value lower than its valuation, it should continue not being scheduled. Suppose that this is not the case and k is scheduled when it bids b'_k. Let B' be the sum of winning bids in this case, the rest of the bids being the same as for B. Let $\mathbf{A}'$ be the corresponding allocation vector. Now, if the same allocation vector is used when the UEs bid their true valuations, the resulting sum of winning bids will be $B'' = B + v_k$ which is a contradiction since B is the maximum bid value obtainable with true valuations. Thus, it is not possible for k to get scheduled when it bids b'_k. Therefore, there is no incentive in bidding b'_k instead of b_k.

We have shown for all possible cases that manipulating the actual valuations in any manner does not result in any utility gain for the UEs under allocation policy $\Gamma^\star$. This proves that $\Gamma^\star$ is strategy-proof. □

5.3.1 Computational Complexity of $\Gamma^\star$

The brute force implementation of $\Gamma^\star$ requires going through all possible resource allocations and calculating the sum of winning bids for each allocation. The optimal allocation can then be obtained by choosing the one that maximizes the sum of winning bids. The computational complexity of this algorithm is $\mathcal{O}(L\binom{N}{L}L!)$. This is computationally very expensive and unsuitable for practical implementation. However, in the problem under consideration here, the proposed mechanism can be implemented in polynomial-time using a Maximum Weight Bipartite Matching (MWBM) based algorithm. In the next section, we present this polynomial-time implementation of $\Gamma^\star$.

5.4 MWBM Implementation of $\Gamma^\star$

We propose a MWBM based implementation of $\Gamma^\star$ that has a computational complexity of $\mathcal{O}(L^2N)$. We first construct the bipartite graph for the matching and then we prove that determining a maximum weight matching for it is equivalent to determining the resource allocation according to $\Gamma^\star$. Construct a bipartite graph $\mathcal{G} = (U, V, E)$ where U is the set of all the MBMS groups $[L]$ and V is the set of all PRBs $[N]$ as shown in Figure 5.1. E denotes the set of edges between the two sets of nodes. The weight of the edge between

vertex $i \in U$ and vertex $j \in V$ is defined as:

$$w_i^j = \sum_{k \in G_i} b_k \times (1 - \ell_{kj}).$$

MWBM of $\mathcal{G}$ is the social utility maximizing resource allocation given by $\Gamma^\star$. We prove this in the following result.

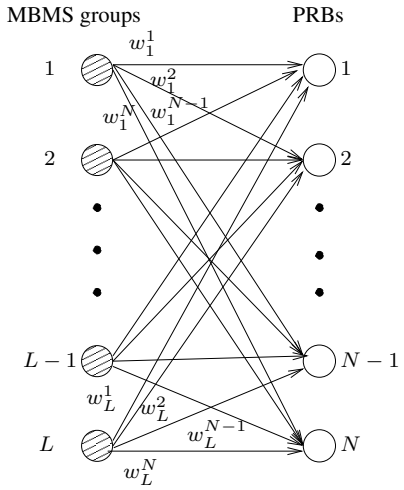


Figure 5.1: Bipartite graph between multicast groups and PRBs

Lemma 11. *MWBM for graph $\mathcal{G} = (U, V, E)$ results in the same resource allocation as that given by $\Gamma^\star$.*

Proof. Let us first establish that MWBM of $\mathcal{G}$ results in a feasible resource allocation. By definition of a matching, MWBM of $\mathcal{G}$ selects edges with no common vertices. Therefore, a vertex from U is matched to at most one vertex from V and vice-versa. This means that each group is given a single PRB in a sub-frame. It also ensures that, in the resulting allocation vector $\mathbf{A}$, $A_i \neq A_{i'}\ \forall\ \{i, i' \in U\}$. Hence, by Definition 12, the resulting resource allocation is feasible. All that is left to show is that this feasible allocation also maximizes the quantity given in (5.2).

Since MWBM searches for a maximum weight matching with no common nodes, it effectively maximizes the quantity,

$$\sum_{i,j} w_i^j x_{ij}^\Gamma = \sum_j \sum_i \sum_{k \in G_i} b_k (1 - \ell_{kj}) x_{i(k)j}^\Gamma = \sum_j \sum_k b_k x_{i(k)j}^\Gamma (1 - \ell_{kj}),$$

which is the same quantity that is maximized by $\Gamma^\star$ as given in (5.2). Thus, MWBM for $\mathcal{G}$ successfully implements the allocation mechanism of $\Gamma^\star$. □

Table 5.1: System Simulation parameters [1]

Parameters	**Values**
System bandwidth	20 MHz
Path loss model	$\mathrm{L} = 128.1 + 37.6 \log 10(d)$, d in kilometers
Shadowing	Log Normal Fading with 10 dB standard deviation
White noise power density	-174 dBm/Hz
eNB noise figure	5 dB
eNB transmit power	46 dBm
Number of PRBs	100 per sub-frame

5.5 Simulations

To study the performance of the proposed allocation mechanism, we implement it in an LTE environment. We first discuss the simulation setup and then present the results. We consider an LTE cell with 100 UEs distributed uniformly at random throughout the cell. The channel models used are in accordance with the 3GPP specifications [1]. There are 5 different MBMS streams available for subscription and all UEs in the cell are either subscribed to one of them or are receiving a unicast service. We run the simulations for a period of 10^5 sub-frames. Other relevant simulation details are given in Table 5.1.

As discussed in Section 5.1, we do not place any restrictions on what the valuations of users can be. For the purposes of these simulations, we assume that each UE has a certain packet loss requirement that needs to be met. The loss tolerable by a UE could be a function of factors like the streaming quality required by it, its channel state, or the kind of video service it has subscribed to. This requirement is known to the UE alone. The valuation of the UEs in a sub-frame is some function of their loss tolerance and the loss they have encountered in the past. The UEs report their respective valuations to the eNB. The eNB then allocates a PRB to each MBMS group using the algorithm detailed in Section 5.4. We compare the performance of our policy with that of a greedy policy Γ^G that maximizes the system throughput.

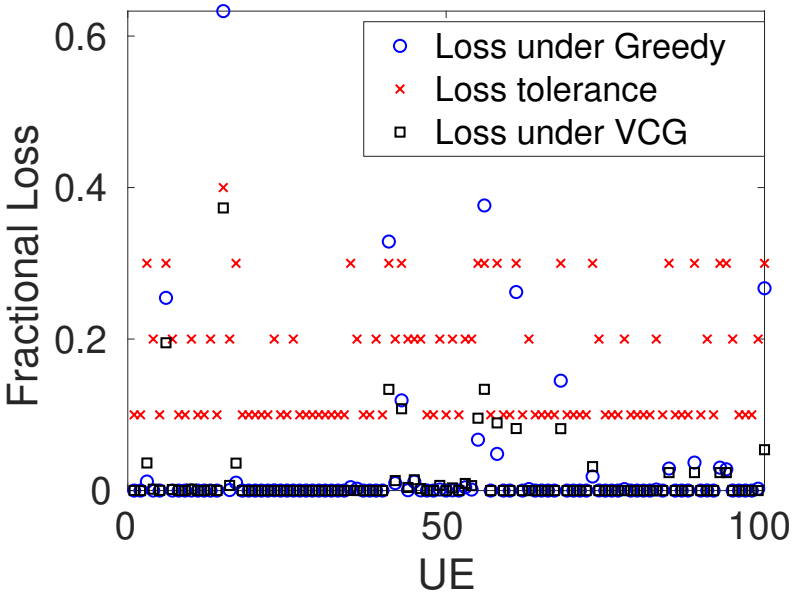


Figure 5.2: Tolerable loss versus loss encountered

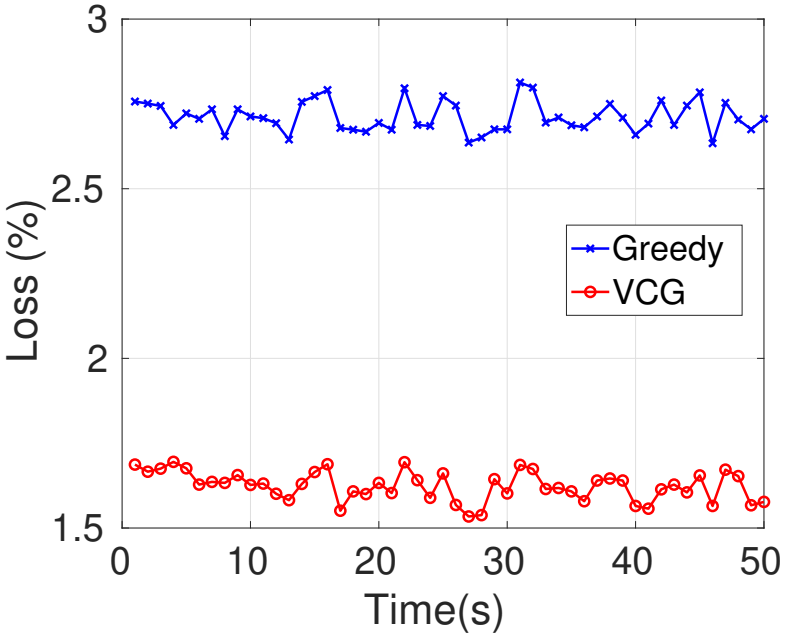


Figure 5.3: Average loss pattern over time

Figure 5.2 shows the plot of the loss tolerance of UEs and the actual loss encountered by them under $\Gamma^{\star}$ and Γ^{G}. We observe that the loss encountered by every UE remains within its tolerable limit under $\Gamma^{\star}$ whereas Γ^{G} fails to meet their loss requirements. $\Gamma^{\star}$ succeeds in meeting the loss requirements of all the UEs even though the eNB has no

knowledge of their loss tolerance or the manner in which the user valuations are calculated. In fact, the eNB is not even aware that the resource allocation algorithm is trying to control the losses encountered by the users. Thus, the proposed algorithm provides a resource allocation mechanism that can be used for optimizing any system parameter. In order to see how the loss patterns evolve over time, we also observe the average percent packet loss under the two policies as a function of time. This is shown in Figure 5.3. We observe that $\Gamma^{\star}$ results in significantly lower losses than the greedy policy.

5.6 Conclusions

Traditionally, resource allocation policies have been designed to optimize a specific parameter, such as maximizing the system throughput or fairness of allocation. A practical cellular mobile system, however, may need to optimize different parameters for different types of traffic that it is serving. There could be services like multicast video streaming, voice calls, or file downloads simultaneously active in a cell. The current literature on resource allocation lacks a generalized resource allocation policy that can be used irrespective of the parameter we may need to optimize. In this chapter, we design such a policy based on auctions. The proposed policy provides a generalized framework for resource allocation that can be used for allocation irrespective of the types of users and services in the system. It can serve a heterogeneous mix of users and services having diverse QoS requirements.

The policy is based on the popular VCG auction mechanism. We have shown that the proposed policy is strategy-proof. Hence, there is no incentive for rational users in the system to misrepresent their valuations for the system resources. The policy takes allocation decisions based on the reported user valuations. The valuations of users are private, unknown to the eNB and other users. Through simulations in an LTE environment, we have shown that the proposed policy succeeds in meeting the QoS requirements of all the users present in the system even though these requirements are not known to the allocating entity. Though VCG mechanisms are generally NP-hard to implement, we have shown that, for the problem under consideration, the proposed mechanism can be implemented in polynomial-time using a maximum weight bipartite matching.

In the next chapter, we propose the idea of using multi-connectivity in multicast transmissions. To the best of our knowledge, as of writing this thesis, multi-connectivity has never been considered for use in multicast transmissions.

Chapter 6

Multi-Connectivity Multicast Streaming

In this chapter, we propose the use of Multi-Connectivity (MC) in multicast transmissions. Multi-connectivity has a potential to significantly improve the performance of multicast video streaming services. It allows users to potentially connect to and receive content from multiple eNBs and over multiple Radio Access Technologies (RATs) simultaneously. Multi-connectivity, specifically Dual Connectivity (DC), is an essential part of the next generation of mobile cellular networks. DC capable devices can connect to at most two Base Stations (BSs) at a time. DC allows users to connect to a primary macro BS and a secondary micro or femto BS [89]. DC is expected to be a key enabler in 5G wireless networks [91]. The high data rate, ultra-reliable low latency, and high mobility requirements of 5G necessitate the reduction of radio link failures due to mobility. The use of DC makes it possible to avoid such failures and ensures seamless connectivity for mobile users [92]. Even though DC has received considerable attention from the research community in the past few years for throughput and handover improvement [89, 93–95], its use with multicast transmissions has not been considered.

Allowing UEs to receive content from several BSs at once is particularly suitable for multicast streaming services. As we shall discuss in Section 6.1, various features of MBMS make it particularly feasible to use multi-connectivity in MBMS transmissions. The only context in which delivery of multicast content from multiple sources is considered in the current 3GPP standards is in the use of Multimedia Broadcast multicast service Single Frequency Networks (MBSFN). In MBSFNs, multiple eNBs in an MBSFN area trans-

mit the same content to the users in synchronization [10]. MBSFN transmissions require strict synchronization between all eNBs in the MBSFN area and an extended Cyclic Prefix (CP) so that users can successfully combine the content received from multiple eNBs. The extended CP reduces the system throughput and the need for tight synchronization between eNBs results in significant control overheads. Multi-connectivity multicast addresses these issues. It provides all the benefits of MBSFNs without the need for strict synchronization and extended CPs. Each eNB can independently optimize its resource allocation and choose the most suitable PRBs for transmitting the multicast content. All users in an MBSFN area can receive the same content from all eNBs and experience improved quality of service.

Since use of multi-connectivity in MBMS has not been considered before, the associated methods have not been standardized in the current 3GPP standards. Therefore, procedures need to be defined for enabling the use of multi-connectivity in MBMS. We define these procedures and the associated control signaling in this chapter. We also formulate the resource allocation problem for a multi-connected multicast system to maximize the number of multicast users served. We prove that this resource allocation problem is NP-hard. Since a multi-connected system involves users receiving content from multiple eNBs, the optimal resource allocation needs to consider a global view of the system to make optimal allocation decisions. This requires the presence of a controller that makes allocation decisions in a centralized manner. We propose a centralized greedy approximation algorithm for solving the resource allocation problem. The proposed algorithm provides an approximation ratio of $(1 - \frac{1}{e})$, which means that the solution provided by it is within $(1 - \frac{1}{e})$ of the optimal solution. This is, in fact, the best possible approximation for the problem. We also propose a distributed greedy allocation policy and compare its performance to that of the centralized policy. Through extensive simulations, we demonstrate the performance improvements provided by multi-connectivity in multicast transmissions. We compare the performance of a multi-connected multicast system with that of single connected and dual connected multicast systems. Our simulation results reveal that multi-connectivity significantly improves the performance of multicast transmissions.

The rest of this chapter is organized as follows. We discuss the proposed concept

of multi-connectivity multicast in Section 6.1. We define the procedures for establishing multi-connectivity in MBMS in Section 6.2. The system model and the resource allocation problem formulation are discussed in Section 6.3. In Section 6.4, we prove NP-hardness of the resource allocation problem. We present the centralized greedy approximation algorithm and prove its approximation ratio in Section 6.5. We then examine the use of distributed resource allocation for MC multicast in Section 6.6. Finally, we present the simulation results in Section 6.7 and conclude in Section 6.8.

6.1 MBMS and Multi-Connectivity

MBMS, as defined in 3GPP standards, is an idle mode procedure [11]. This means that, a UE does not have to be in RRC connected mode to receive MBMS services. Most of the control information relating to MBMS operations is carried on a separate logical channel, the Multicast Control Channel (MCCH) [116]. The only MBMS related information sent over the Broadcast Control Channel (BCCH) is the information needed by UEs to acquire the MCCH(s). This information is carried by the MBMS specific SystemInformationBlock, *SystemInformationBlockType13* (SIB13) [116]. MBMS user data is carried over Multicast Traffic Channels (MTCH). Using the information provided over the MCCH, a UE can read the MTCH corresponding to the MBMS session that it is interested in. MBMS user plane protocol architecture defines an additional Synchronization (SYNC) protocol layer on the transport network layer for content synchronization [117]. It is defined to carry additional information for identifying transmission times and detecting packet loss. The SYNC protocol is terminated in Broadcast Multicast Service Centre (BM-SC) and the eNBs. MBMS session content is forwarded to MBMS GateWay (MBMS-GW) by BM-SC. MBMS-GW then IP multicasts the content to the eNBs.

The streaming content sent to eNBs in a particular region emanates from the same BM-SC. As a result, MBMS contents arriving at these eNBs are in sync. UEs can, therefore, receive and combine multiple copies of the same content received from these eNBs. We propose the use of multi-connectivity in multicast transmissions to take advantage of this inherent synchronization in MBMS systems. MC multicast enables multicast users to obtain multicast content from multiple sources without the need for any additional

synchronization. Moreover, since MBMS is an idle mode procedure, we do not require UEs to establish an RRC connection to any eNB for using MC multicast. A UE may use MC multicast while being in RRC idle mode.

We propose a different dynamic between the primary and secondary eNBs of users than what is traditionally defined for DC in unicast transmissions [90]. Firstly, depending on its capability, a UE can connect to any number of eNBs and receive multicast content from all of them. A UE can also stay in the RRC idle mode if it is not connected to any eNB and receive content from any number of eNBs. For a UE using MC multicast in the RRC idle mode, 'primary' eNB refers to the eNB that it is camped on. For a UE in RRC connected mode, 'primary' eNB refers to the eNB that it is connected to. All other eNBs that the UEs may receive content from are referred to as secondary eNBs. Secondly, in MC multicast, primary and secondary eNBs of a UE do not work in a traditional master-slave configuration. The secondary eNBs are not dictated by the primary eNB in their interaction with the UE. A multicast UE can receive relevant control information and multicast data from multiple eNBs independent of each other. As such, there is no real distinction between the 'primary' and 'secondary' eNBs for a user. Each eNB that serves the UE under MC multicast is equivalent for the UE. Note that, we still use the terms 'primary' and 'secondary' eNB in various places in this chapter for the ease of distinguishing between the various eNBs that a UE is receiving data from.

Multi-connectivity multicast has a potential to provide all the benefits of MBSFN transmissions in a considerably simpler framework. Like in MBSFNs, UEs can receive multicast content from a number of eNBs, resulting in improved SNR, especially for the cell edge users. However, unlike MBSFN operations, eNBs are not required to use the same time and frequency resources (PRBs) for streaming the multicast content. In MC multicast, the same MBMS services are streamed through multiple eNBs and each eNB allocates PRBs to the multicast streams independently. Each eNB can, therefore, optimize the resource allocation to various services in its cell. The resulting frequency diversity significantly improves the probability of reliable reception of the MBMS content. A multicast UE can combine or choose one of the multiple copies of the content received by it. As we shall see in Section 6.7, multi-connectivity results in significant performance improvement of multicast operations.

6.2 Procedures for Establishing Multi-Connectivity in Multicast Transmissions

In this section, we propose the procedures required for establishing multi-connectivity in MBMS. We define multi-connectivity multicast as a user initiated mechanism. As discussed in Section 6.1, a UE needs to acquire the MBMS specific SIB, SIB13 and MCCH from an eNB to begin receiving MBMS session content from it. The procedures for establishing multi-connectivity for UEs in RRC connected and RRC idle modes vary in certain signaling aspects. We explain each of these procedures below.

1. **RRC Idle mode**: A UE in RRC idle mode is informed of the available MBMS sessions by its primary cell that it is camped on. If the UE is interested in an available MBMS session and capable of multi-connectivity, it can choose to receive the content from multiple eNBs in its vicinity. If the UE chooses to receive the session from multiple eNBs, it starts listening to the broadcast channels of its neighboring eNBs. It receives the the MasterInformationBlock (MIB), SIB1, SIB13 of the neighboring cells. The SIB13 obtained from the eNBs contains the MBMS relevant information of these eNBs. The UE can then receive the content from any number of these eNBs where the MBMS session of its interest is available.

   To start receiving the session content, the UE reads MCCH(s) of these eNBs. MCCH contains the information needed by the UE to obtain the relevant MTCH(s). This procedure is illustrated in Figure 6.1. The UE thus receives multiple copies of the same multicast content over MTCHs of multiple eNBs. Depending on the UE capabilities, multiple copies of the content can then be combined to obtain better SNR. It should be noted here that the UE does not need to establish an RRC connection to any of the eNBs in this procedure.

2. **RRC Connected mode**: A UE may have established an RRC connection for some unicast service by the time an MBMS session starts. After a UE establishes an RRC connection to a cell, it stops listening to the broadcast channels of other cells. When such a UE is informed of an MBMS session that it is interested in, it can choose to either receive the content only from the cells that it is connected to or to receive the content from multiple cells using multi-connectivity multicast. The procedure

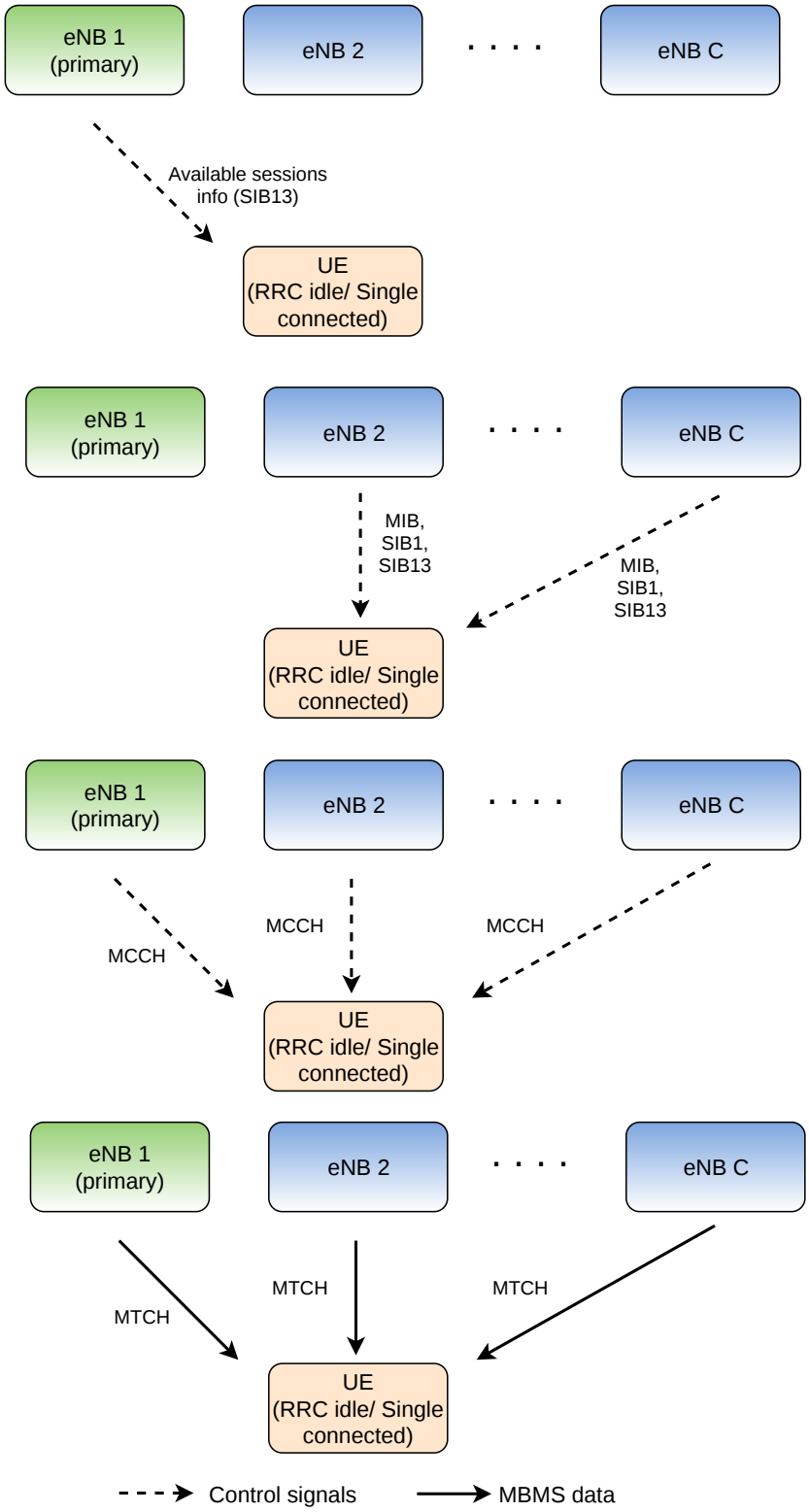


Figure 6.1: Procedure for enabling multi-connectivity multicast for UEs in RRC idle mode and single connected UEs.

for establishing multi-connectivity for such a UE will be different depending upon whether the UE is single connected or dual connected. We propose the procedures for both these cases as follows.

(a) *Single Connected UE*: A single connected UE is notified of the available MBMS sessions by its primary eNB that it is connected to. If the UE is interested in an MBMS session, it can either choose to receive the content from its primary cell alone or use multi-connectivity multicast to receive it from multiple sources. In case the UE decides to receive the MBMS content only from its primary eNB, it reads the corresponding MCCH and receives the content over the relevant MTCH from its primary eNB alone. If the UE chooses to use multi-connectivity multicast instead, it starts listening for broadcast information from its neighboring cells. It acquires the MIB, SIB1 and SIB13 of these cells. It then acquires the MCCH(s) of the additional cells where the session of its interest is available. MCCH provides the UE with the allocation information needed to acquire the MTCH of its interest. This procedure is illustrated in Figure 6.1. Note that, the UE does not establish an RRC connection to any of the secondary eNBs.

(b) *Dual Connected UE*: Consider a UE that is dual connected and receiving some unicast service from two different cells. This UE can choose to receive MBMS content in one of the following ways.

   i. From the primary eNB alone: In this case, the UE acquires MCCH and the relevant MTCH from its primary eNB and no additional procedures are required.

   ii. From its primary eNB and some other eNBs in its neighborhood: This is the same as the case of a single connected UE in 2a above and the same procedures apply.

   iii. From the two eNBs it is dual connected to: If the MBMS session of interest is available in the secondary cell of the UE, it can choose to receive MBMS content from the same two eNBs that it is dual connected to. In the existing 3GPP standards for dual connectivity, the primary eNB acts as the control plane anchor for the UE [90]. All the control information

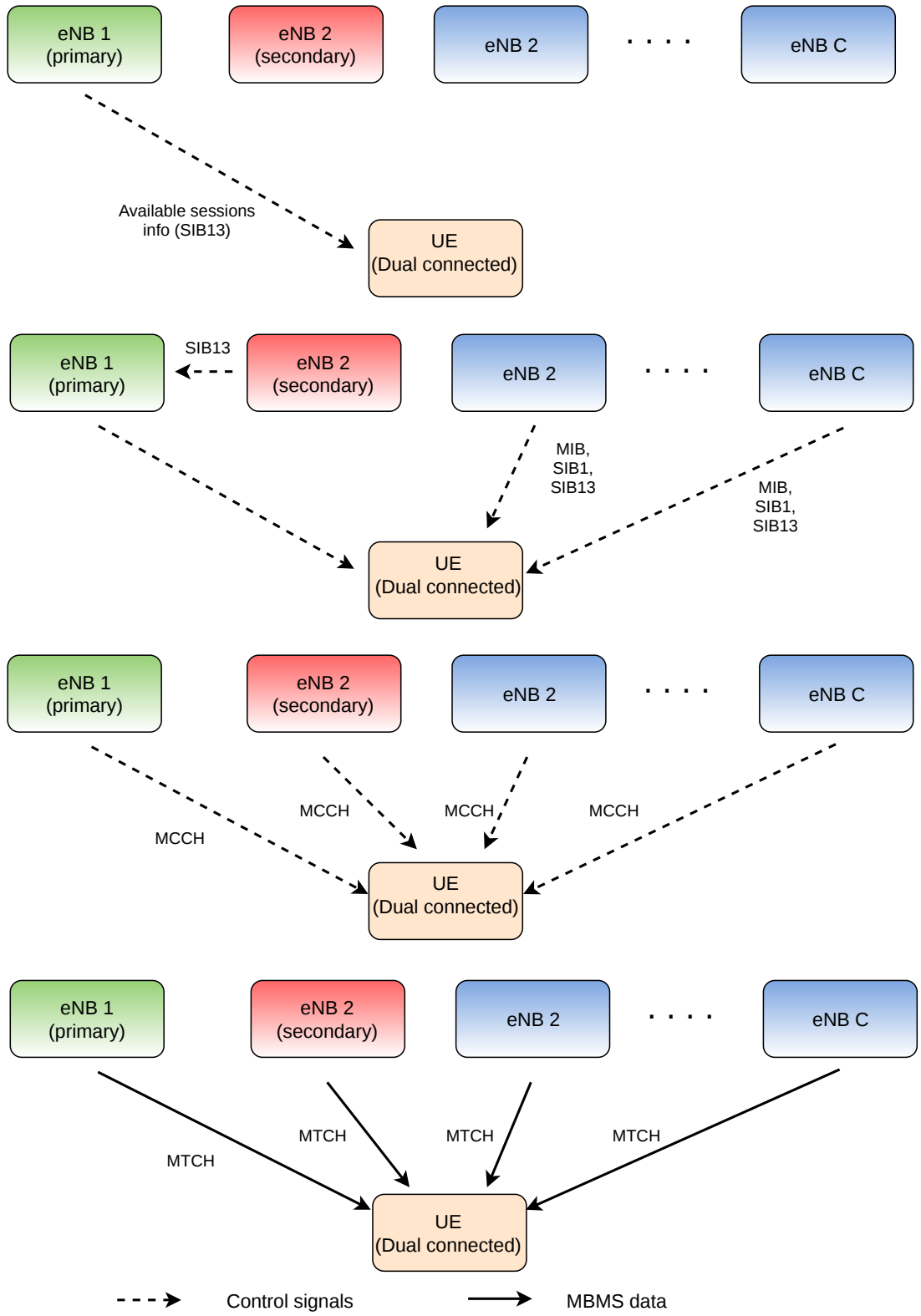


Figure 6.2: Procedure for enabling multi-connectivity multicast for dual connected UEs.

coming from the secondary eNB is transmitted to the UE via the primary eNB. Therefore, we propose that, SIB13 from the secondary eNB is also transmitted over the X2 interface to the primary eNB instead of the UE having to listen for it separately. It can then acquire the MCCH and relevant MTCH from both the cells independently.

iv. From the two eNBs it is connected to as well as other eNBs in its neighborhood: This is a combination of ii. and iii. above and the same procedures are followed. Figure 6.2 illustrates this procedure.

In the next section, we discuss the resource allocation problem in a multi-connectivity multicast system. We formulate the resource allocation problem with the aim of maximizing the number of multicast users successfully served in the system.

6.3 Resource Allocation in MC Multicast

6.3.1 System Model

We consider a C cell LTE system. Each cell has an eNB located at the center. There are M multicast UEs in the system. All UEs are capable of multi-connectivity and can potentially be served by any number of eNBs. There is a multicast session available for streaming in all the cells. The multicast stream has a certain required rate R at which the content needs to be streamed to the subscribed UEs. Multicast content is streamed at this rate R whenever the multicast session is active. All multicast UEs in the system are subscribed to the ongoing multicast session. The UEs subscribed to a multicast session in a cell form a single multicast group and receive the streaming content over the same PRBs. The UEs can potentially receive the multicast streaming content from any number of neighboring eNBs in addition to their primary eNB. A multi-connected UE, therefore, belongs to multiple multicast groups streaming the same content. The multicast stream in a cell is allocated one PRB each sub-frame t. Resource allocation to the various multicast streams can either be done by each eNB independently or by a central controller that manages the eNBs. The multicast data stream in the primary and secondary cells of a UE may or may not be scheduled on the same PRB.

The channel states of UEs vary across time and frequency. As a result, a UE experiences different channels in different sub-frames and across different PRBs in a sub-frame. Depending on the channel state of a UE, there's a certain maximum rate it can successfully decode in a PRB. Since the multicast content is transmitted at rate R, a UE may or may not be successfully served by an eNB it is connected to. For instance, say cell c is streaming the multicast content over PRB j in sub-frame t. Let $r_{jk}^{c}[t]$ be the maximum decodable rate for UE k in PRB j of cell c in sub-frame t. If $R > r_{jk}^{c}[t]$, UE k will not be able to successfully receive the content from c. On the other hand, if $R \leq r_{jk}^{c}[t]$, UE k will be successfully served by c. A multi-connected UE successfully receives data in sub-frame t if it can decode the content from any of the eNBs it is connected to. On the other hand, a UE that is not multi-connected would successfully receive data only if it can decode the content from its primary cell. We now discuss and formally define the resource allocation problem for MC multicast.

6.3.2 Problem Formulation

The problem of resource allocation in an MC multicast system is aimed at serving as many UEs successfully in a sub-frame as possible. Since multi-connected UEs can receive the streaming content from multiple eNBs, the performance of a UE depends on the PRB allocation in multiple cells. The optimal resource allocation for a region must, therefore, optimize over all the cells in that region. Optimal allocation of resources in individual cells is not optimal for a multi-connected system. We now mathematically define the optimal resource allocation problem.

M multicast users distributed in C cells can potentially receive multicast content from all the eNBs in their neighborhood. $[M]$ is the universal set of all users. There are N PRBs available in each cell. Denote by $U_{jc} \subseteq [M]$, the set of users that would be successfully served if PRB j is allocated to the multicast service in cell c. Set $\mathcal{U}_c = \{U_{1c}, U_{2c}, \ldots, U_{Nc}\}$ is the sub-collection of such sets for cell c. Let $\mathcal{U} = \{\mathcal{U}_1, \ldots, \mathcal{U}_C\}$. The resource allocation problem can then be stated as follows:

$\mathbf{K}^{\star}$: Given the universe $[M]$ and a collection of sets $\mathcal{U} = \{\mathcal{U}_1, \ldots, \mathcal{U}_C\}$, we are required to determine $\mathcal{U}' \subseteq \mathcal{U}$ such that $|\bigcup_{U_{jc} \in \mathcal{U}'} U_{jc}|$ is maximized subject to $|\mathcal{U}'| = C$ and $|\mathcal{U}' \cap \mathcal{U}_c| = 1, \ \forall \ c$.

6.4 $\mathbf{K}^\star$ is NP-hard

The optimal resource allocation problem $\mathbf{K}^\star$ is an NP-hard problem. We prove this by reduction from the Maximum Coverage Problem (MCP) [118]. MCP is a well known NP-hard problem and is defined as follows:
MCP takes as input a universe $\mathcal{S}$, a number k and a collection of sets $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$ where each $T_j \subseteq \mathcal{S}$. The objective of MCP is to determine a sub-collection $\mathcal{T}' \subseteq \mathcal{T}$ such that $\mathcal{T}' \in \arg\max_{|\mathcal{T}'| \leq k} |\bigcup_{T_j \in \mathcal{T}'} T_j|$.

Theorem 9. $\mathbf{K}^\star$ *is an NP-hard problem.*

Proof. In order to prove that $\mathbf{K}^\star$ is NP-hard, we first reduce an instance of MCP to an instance of $\mathbf{K}^\star$ in polynomial time. Then, we demonstrate how a solution for $\mathbf{K}^\star$ can be mapped to a solution for MCP. We begin with the reduction.

Algorithm 6: Pseudo-code for reducing MCP to $\mathbf{K}^\star$

Input: MCP with collection of sets $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$ and a number, $k \in \mathbb{N}$

Output: An instance of $\mathbf{K}^\star$ with

1 $C \leftarrow k$

2 $N \leftarrow m$

3 $U_{jc} \leftarrow T_j \; \forall \; j \in \{1, 2, \ldots, m\}, c \in \{1, 2, \ldots, C\}$

The pseudo-code for reducing an instance of MCP to an instance of $\mathbf{K}^\star$ is given in Algorithm 6. We define the total number of cells to be k and the number of PRBs available as m. The set U_{jc} is set to be T_j. This reduction can be accomplished in constant time ($\mathcal{O}(C)$). We now demonstrate how a solution of $\mathbf{K}^\star$ can be mapped to a solution of MCP.

Algorithm 7: Pseudo-code for mapping a solution of $\mathbf{K}^\star$ to a solution of MCP

Input: Solution of $\mathbf{K}^\star$ $\mathcal{U}' \subseteq \mathcal{U}$ such that $|\mathcal{U}'| = C$ and $|\mathcal{U}' \cap \mathcal{U}_c| = 1, \; \forall \; c$

Output: Solution of MCP $\mathcal{T}'$

1 $T_j \in \mathcal{T}'$ iff $U_{jc} \in \mathcal{U}'$ for some c

Let us assume that there exists a polynomial time algorithm for solving $\mathbf{K}^\star$. Say $\mathcal{U}'$ is the solution of $\mathbf{K}^\star$. This means that $|\mathcal{U}'| = k$, $|\mathcal{U}' \cap \mathcal{U}_c| = 1, \; \forall \; c$ and $\mathcal{U}'$ maximizes $|\bigcup_{U_{jc} \in \mathcal{U}'} U_{jc}|$. This solution can be mapped to a solution of MCP as follows:

Construct set $\mathcal{T}'$ such that, $T_j \in \mathcal{T}'$ iff $U_{jc} \in \mathcal{U}'$. Since $|\mathcal{U}'| = k$, we have, $|\mathcal{T}'| \leq k$. Therefore, $\mathcal{T}'$ is a feasible solution of MCP. The pseudo-code for this mapping is given in Algorithm 7. We now need to prove that this is indeed the optimal solution of MCP. We prove this by contradiction as follows.

Say that that there exists $\mathcal{T}'' \subseteq \mathcal{T}$ such that $|\mathcal{T}''| \leq k$ and $|\bigcup_{T_j \in \mathcal{T}''} T_j| > |\bigcup_{T_j \in \mathcal{T}'} T_j|$. We can then construct U'' using $\mathcal{T}''$ as follows. Say $\mathcal{T}'' = \{T_{j_1}, \ldots, T_{j_l}\}, l \leq k$ and say $j_1 < j_2 < \ldots < j_l$. Then, we can construct $\mathcal{U}'' = \{U_{j_1 1}, U_{j_2 2}, \ldots, U_{j_l l}, U_{1(l+1)}, \ldots, U_{1C}\}$. We have, $|\mathcal{U}''| = C$, $|\mathcal{U}'' \cap \mathcal{U}_c| = 1$, $\forall\ c$ and $|\bigcup_{U_{jc} \in \mathcal{U}''} U_{jc}| > |\bigcup_{U_{jc} \in \mathcal{U}'} U_{jc}|$ which is a contradiction to $\mathcal{U}'$ being the solution of $\mathbf{K}^\star$. Therefore, $\mathcal{T}'$ is indeed the optimal solution of MCP.

Algorithm 7 maps a solution of $\mathbf{K}^\star$ to a solution of MCP in constant time ($\mathcal{O}(C)$ assignments). Thus, a polynomial time solution for $\mathbf{K}^\star$ results in a polynomial time solution for MCP as well. This is not possible unless P = NP. Therefore, no polynomial time algorithm exists for solving $\mathbf{K}^\star$ i.e., $\mathbf{K}^\star$ is an NP-hard problem. □

Now that we have proved that the multi-connectivity problem is NP-hard, the best we can do is construct approximation algorithms that provide some performance guarantees. We propose one such greedy algorithm in the next section and prove that the algorithm has an approximation factor of $\left(1 - \frac{1}{e}\right)$.

6.5 Approximation Algorithm for $\mathbf{K}^\star$

We construct a Centralized Greedy Approximation (CGA) algorithm for solving $\mathbf{K}^\star$. The pseudo-code for the algorithm is given in Algorithm 8. The algorithm maximizes the number of additional users served in each step. In the first iteration, CGA chooses U_{jc} from $\mathcal{U}$ that serves the maximum number of users. In the subsequent steps, it picks U_{jc}'s that serve the maximum number of unserved users. In each step, the set picked is from a different sub-collection $\mathcal{U}_c$ i.e., c in the subscript of the chosen sets is different for each set picked. The collection of sets chosen after C such iterations $\mathcal{U}_G$, is the output of the algorithm. In the next result, we prove that the resulting solution has an approximation factor of $\left(1 - \frac{1}{e}\right)$. This means that the solution provided by this greedy approximation algorithm serves at least $\left(1 - \frac{1}{e}\right)$ of the number of users that would be served by the

optimal algorithm.

Theorem 10. *The CGA algorithm (Algorithm 8) is a* $\left(1-\frac{1}{e}\right)$ *approximation for* $\mathbf{K}^\star$*. In fact, no other algorithm can achieve a better approximation unless* $P = NP$*.*

Algorithm 8: Greedy Approximation Algorithm for $\mathbf{K}^\star$

Input: Universe $[M]$, $\mathcal{U} = \{\mathcal{U}_1, \ldots, \mathcal{U}_C\}$, C

Initialize: $\mathcal{U}_G = \phi$

for $n = 1 : C$ **do**

  Pick $U_{j^\star c^\star} \in \mathcal{U}$ that covers the maximum number of elements from $[M] \setminus \bigcup_{U_{jc} \in \mathcal{U}_G} U_{jc}$

  $\mathcal{U}_G \leftarrow \mathcal{U}' \bigcup \{U_{j^\star c^\star}\}$

  $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{U}_{c^\star}$

end

Let OPT denote the number of UEs served by the optimal solution. Let m_n be the total number of UEs served by CGA up to and including the n^{th} iteration. $b_n = OPT - m_n$ is the difference between the number of UEs served by the optimal algorithm and the number of UEs served by CGA up to the n^{th} iteration. Note that $m_0 = 0, b_0 = OPT$ and m_C is the total number of UEs served by CGA.

In order to determine the approximation ratio of CGA, we first prove the following two results that will eventually help us determine the approximation ratio in Theorem 10.

Lemma 12. $m_{n+1} - m_n \geq \frac{b_n}{C}$.

Proof. Let $U_{OPT} = \{U_1^\star, \ldots, U_C^\star\}$ be the optimal solution. Denote by M_n, the set of users served at the end of the n^{th} iteration of CGA and by M_n^C the set of users not yet covered after the end of the n^{th} iteration. We have:

$$\sum_{c=1}^{C} |U_c^\star \bigcap M_n^C| \geq |\bigcup_{c=1}^{C} (U_c^\star \bigcap M_n^C)| \geq OPT - m_n,$$

$$\implies \max_c |U_c^\star \bigcap M_n^C| \geq \frac{(OPT - m_n)}{C} = \frac{b_n}{C}. \tag{6.1}$$

Since CGA picks the set that serves the maximum possible number of yet unserved users in each iteration, we have:

$$m_{n+1} - m_n \geq \max_c |U_c^\star \bigcap M_n^C|. \tag{6.2}$$

From (6.1) and (6.2),

$$m_{n+1} - m_n \geq \frac{b_n}{C}.$$

□

Lemma 13. $b_{n+1} \leq \left(1 - \frac{1}{C}\right)^{n+1} OPT$.

Proof. We prove this result by induction. For $n = 0$, the above equation becomes:

$$\begin{aligned} & b_1 \leq \left(1 - \frac{1}{C}\right) OPT, \\ \implies & OPT - m_1 \leq OPT - \frac{OPT}{C}, \\ \implies & m_1 \geq \frac{OPT}{C} = \frac{b_0}{C}, \end{aligned}$$

which is true (from Lemma 12). Thus, the result holds for $n = 0$. Let us assume that $b_n \leq \left(1 - \frac{1}{C}\right)^n OPT$ and prove that $b_{n+1} \leq \left(1 - \frac{1}{C}\right)^{n+1} OPT$. By the definition of b_n and m_n, we have:

$$b_{n+1} \leq b_n - (m_{n+1} - m_n), \tag{6.3}$$

$$\implies b_{n+1} \leq b_n - \frac{b_n}{C} = b_n \left(1 - \frac{1}{C}\right), \tag{6.4}$$

$$\implies b_{n+1} \leq \left(1 - \frac{1}{C}\right)^{n+1} OPT. \tag{6.5}$$

(6.4) follows from (6.3) by Lemma 12 and (6.5) follows from (6.4) by our assumption that $b_n \leq \left(1 - \frac{1}{C}\right)^n OPT$. Thus, by induction, the result holds for all n. □

We can now prove Theorem 10. We state the theorem again for ease of the reader.

Theorem 10. *The CGA algorithm (Algorithm 8) is a* $\left(1 - \frac{1}{e}\right)$ *approximation for* $\mathbf{K}^\star$. *In fact, no other algorithm can achieve a better approximation unless* $P = NP$.

Proof. From Lemma 13,

$$\begin{aligned} & b_C \leq \left(1 - \frac{1}{C}\right)^C OPT, \\ \implies & OPT - m_C \leq \left(1 - \frac{1}{C}\right)^C OPT \leq \frac{OPT}{e}, \\ \implies & m_C \geq \left(1 - \frac{1}{e}\right) OPT. \end{aligned}$$

Thus, CGA provides a $\left(1 - \frac{1}{e}\right)$ approximation for $\mathbf{K}^\star$.

This is the best possible approximation for $\mathbf{K}^\star$. If there was an algorithm that could provide a better approximation, that algorithm would also provide a better approximation for MCP because a solution for $\mathbf{K}^\star$ can be mapped to a solution of MCP using Algorithm 7. This is a contradiction since the greedy algorithm is known to be the best possible approximation for MCP unless P = NP [119]. Therefore, no other algorithm can provide a better approximation for $\mathbf{K}^\star$. □

6.6 Distributed versus Centralized Allocation

The CGA algorithm discussed in the previous section is a centralized algorithm. It requires a central controller that can make allocation decisions based on a global view of the multicast region. In the absence of such a centralized controller, allocation decisions would be made by each cell individually based only on the knowledge of the UEs connected to it. In such a distributed allocation, each cell allocates resources to the multicast stream independently. In a multi-connected system, this type of allocation does not fully reap the benefits of MC. We illustrate this with the following example. Consider a 2 cell system containing cells c_1 and c_2. There are two PRBs available for allocation in each cell. We denote these as P_1 and P_2. c_1 contains four users, $\{u_1, u_2, u_3, u_4\}$ and c_2 has two users $\{u_5, u_6\}$. All users are subscribed to the same multicast stream. u_1 has a good channel only in P_1 and can successfully receive content only on P_1. Users u_3, u_4, u_5 and u_6 have a good channel only in P_2 and can, therefore, successfully receive content only on P_2. u_2 has a good channel in both the PRBs and would be served on either of them. Users u_1, u_3, u_4 are connected to both the cells and can receive content from either of them.

Let us now look at the allocations that will be done by a distributed policy that is maximizing the number of users served in each cell independently. c_1 considers the users connected to it and allots P_2 to the stream because it serves the maximum number of users (u_2, u_3, u_4). c_2 also optimizes independently and allocates P_2 to the stream to serve (u_3, u_4, u_5, u_6). Under this allocation, u_1 remains unserved even though it was multi-connected, since it could only receive the content over P_1. On the other hand, u_3 and u_4 receive content from both the cells. In contrast, a centralized policy would consider users

of both cells together and allocate P_2 to the stream in c_2 and P_1 in c_1 and successfully serve all users in the system.

Any centralized allocation policy, even if it is sub-optimal, will always do better in terms of the number of users successfully served than a policy which allocates resources in a distributed manner. A centralized policy does not necessarily mean that the policy is optimizing over the entire system. Any form of centralization that looks beyond just the individual cell will reap a better performance than a completely uncoordinated allocation. We now define a distributed greedy allocation policy. We use this policy for the purpose of simulations in Section 6.7.

6.6.1 Distributed Greedy Allocation

In Distributed Greedy (DG) allocation policy, each eNB allocates resources to the multicast streams by optimizing over the users connected to it. This policy solves $\mathbf{K}^\star$ for each cell individually. In each sub-frame, an eNB allocates PRBs to the multicast streams such that maximum number of users associated with it are served. When optimizing over a single cell, the problem can be solved in polynomial time. The pseudo-code for this algorithm is given in Algorithm 9. Recall that the set U_{jc} in Algorithm 9 is the set of all users that would be successfully served if PRB j were allocated to the multicast stream in c. x_{jc} is the indicator random variable that indicates whether or not PRB j has been allocated to the multicast stream in cell c.

Algorithm 9: DG Allocation

Input: Sets $\mathcal{U}_c = \{U_{1c}, \ldots, U_{Nc}\}$ for all $c \in [C]$

Initialize: $x_{jc} = 0$ for every j, c

for $c = 1 : C$ **do**

  Assign $j^\star = \arg\max_j |U_{jc}|$

  $x_{j^\star c} \leftarrow 1$

end

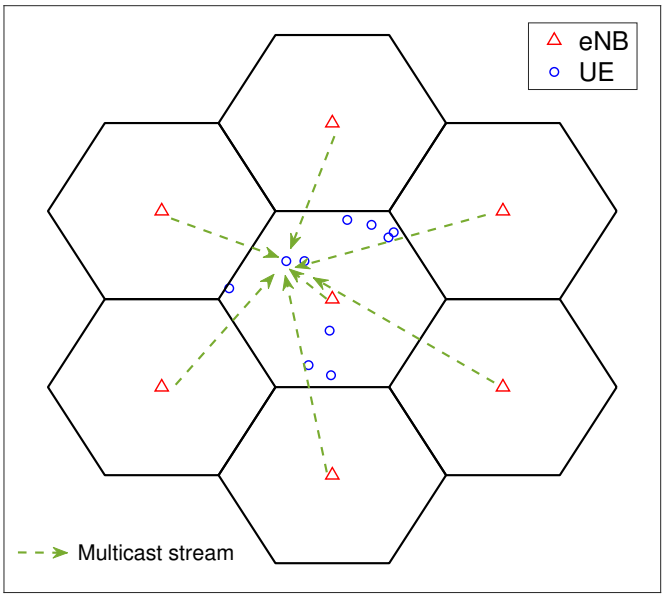


Figure 6.3: A snapshot of the simulation scenario

6.7 Simulations

We simulate a seven cell urban macro scenario [2]. UEs in the system are distributed uniformly at random throughout the cells as shown in Figure 6.3. A single multicast streaming service is available in all the cells. Some other relevant simulation parameters are given in Table 6.1. The cell edge users are multi-connected to all the eNBs in the system. In all the cells, a PRB is allocated to the multicast stream in each sub-frame. Content corresponding to a multicast stream is transmitted at rate R in the PRB allocated to it. The multi-connected users successfully receive a packet in a sub-frame if they can decode the content from any of the eNBs. Other users should be able to decode the content from their primary eNBs for being served.

Resource allocation to the multicast streams is done using the CGA algorithm proposed in Section 6.5 and the DG policy proposed in Section 6.6. We use the number of packets delivered successfully and the number of UEs successfully served in a sub-frame as the performance metrics in these simulations. We compare the performance of the centralized (CGA) and distributed (DG) resource allocation algorithms. We also com-

Table 6.1: System Simulation parameters [1, 2]

Parameters	**Values**
System bandwidth	20 MHz
Cell radius	250 m
Path loss model	$\mathrm{L} = 128.1 + 37.6 \log 10(d)$, d in kilometers
Lognormal shadowing	Log Normal Fading with 10 dB standard deviation
White noise power density	-174 dBm/Hz
eNB noise figure	5 dB
eNB transmit power	46 dBm

pare the performance of MC with Single Connectivity (SC) and MBSFN to establish the performance gains resulting from the use of MC in multicast transmissions. For resource allocation in SC multicast, we use the distributed algorithm from Section 6.6 where each eNB only considers the UEs in its own cell while making the allocation decisions.

In Figure 6.4, we plot the average number of packets successfully received by UEs under the CGA and the DG resource allocation algorithms. We transmit one packet in every sub-frame and plot the average number of packets successfully received by all the UEs in the system over a period of 10 s (10000 sub-frames). We observe that CGA performs much better than the DG policy. It succeeds in successfully serving the UEs in a significantly larger number of sub-frames.

In Figures 6.5 to 6.10, we compare the performance of MC multicast with that of SC multicast. From here onwards, only the CGA algorithm has been used for allocation in MC multicast. For the plots in Figure 6.5 to 6.8, data is transmitted at a fixed rate in each sub-frame. The points in these plots are obtained by averaging over 10000 sub-frames.

Figure 6.5 illustrates the number of packets successfully received under MC and SC as the number of users per cell increases. We observe a decline in the number of packets successfully received as the number of UEs increases. However, the number of packets successfully delivered under MC multicast is much larger than that under SC multicast. Figure 6.6 plots the same metric as a function of cell radius. We observe a decline in the

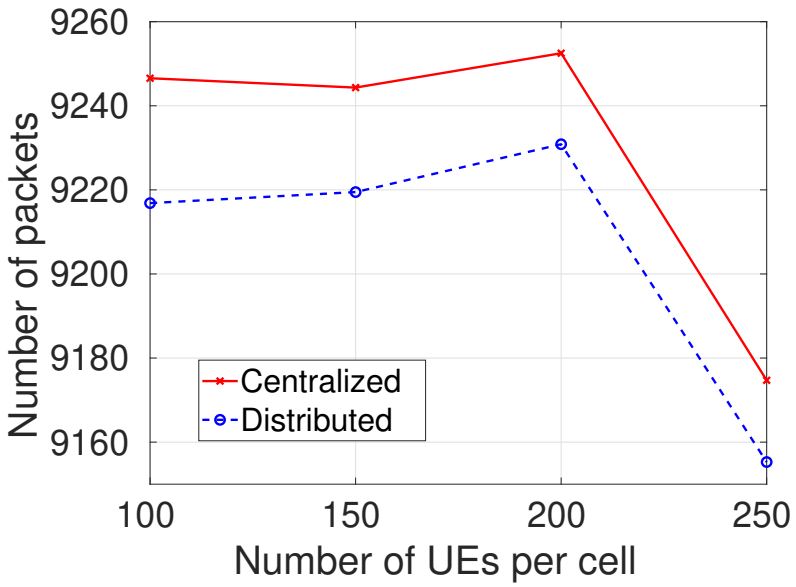


Figure 6.4: Average umber of packets received successfully under MC using centralized and distributed allocation

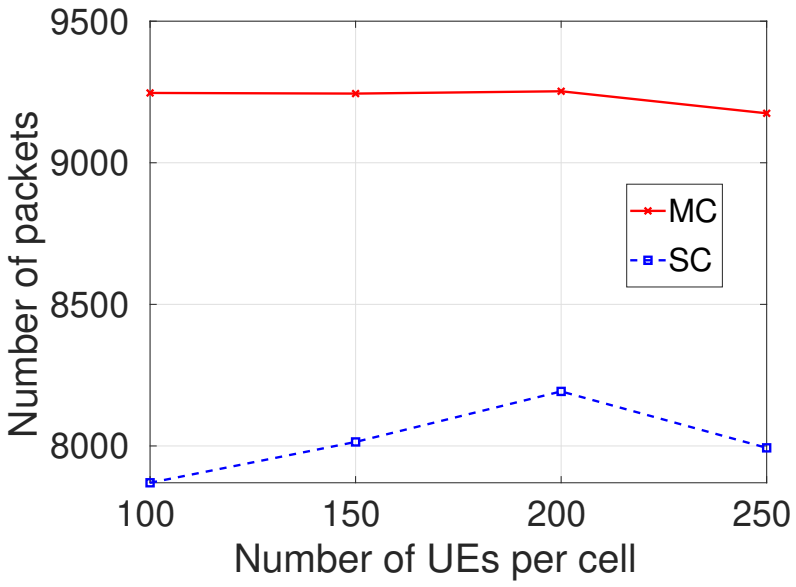


Figure 6.5: Average number of packets received successfully under greedy approximation algorithm as a function of increasing number of users

number of packets successfully received as the cell sizes increase. This is expected since the path loss of the cell edge users increases as the cells become larger. The key thing to note here is that the performance gap between MC and SC follows an increasing trend. The relative performance of MC and SC is similar to what we observe in Figure 6.5.

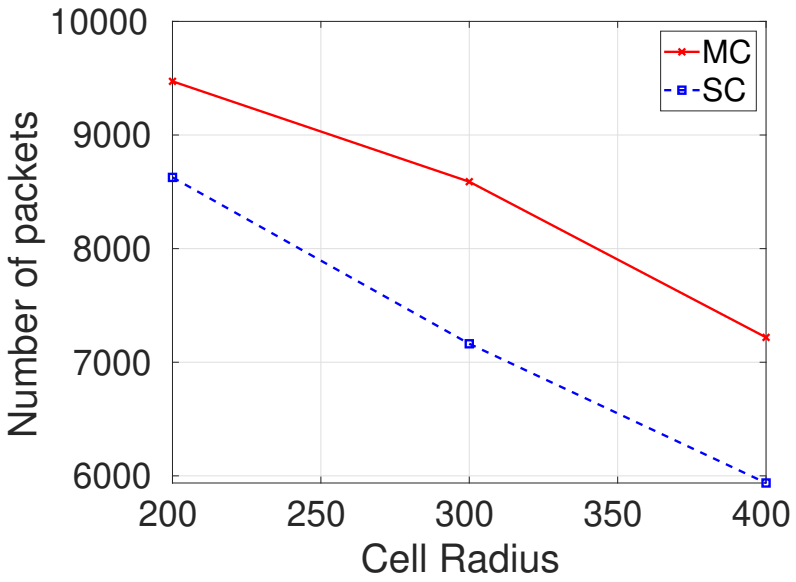


Figure 6.6: Average number of packets received successfully under greedy approximation algorithm as a function of cell radius

Figures 6.7 and 6.8 plot the average number of users left unserved in a cell per sub-frame as a function of increasing number of users and cell radius respectively. The number of unserved users increases as the number of users and cell radius increases. Performance gap between MC and SC increases as the number of users in each cell increases. MC multicast succeeds in serving many more users than SC multicast.

In Figures 6.9 and 6.10, we compare the performance of MC and SC while serving a real-time video stream. To generate realistic video traffic patterns, we use traces of a video of Tokyo Olympics (obtained from `http://trace.eas.asu.edu`) [16]. For these simulations, the rate of transmission varies every sub-frame, according to size of the video frame being transmitted. We run the simulations for the duration of the video stream and then average the results over all the sub-frames. In Figure 6.9, we observe that MC multicast delivers significantly more packets successfully than SC multicast. From Figure 6.10, we observe that many more UEs are left unserved under SC than under MC. The performance gap between the two increases as the number of UEs in the system increases.

In Figures 6.11 and 6.12, we compare the performance of MC multicast with that of MBSFN transmissions. We consider that all the cells in our system belong to the

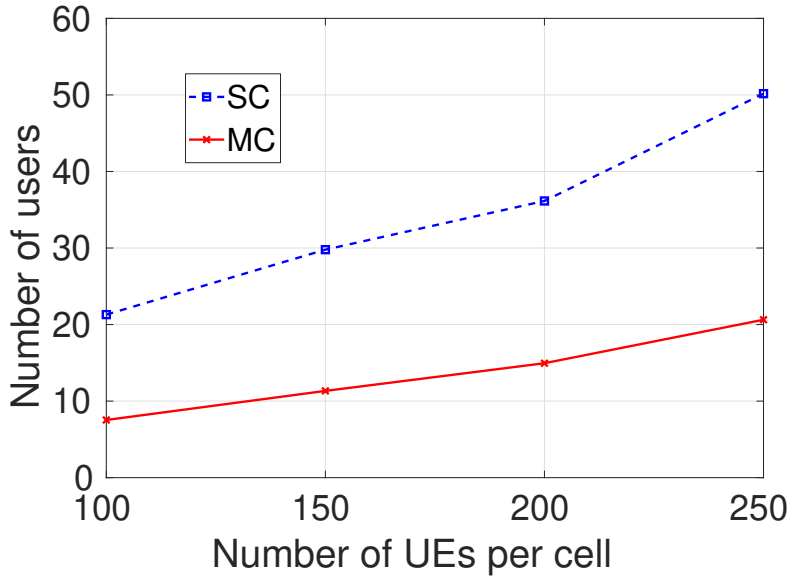


Figure 6.7: Average number unserved users under greedy approximation algorithm as a function of increasing number of users

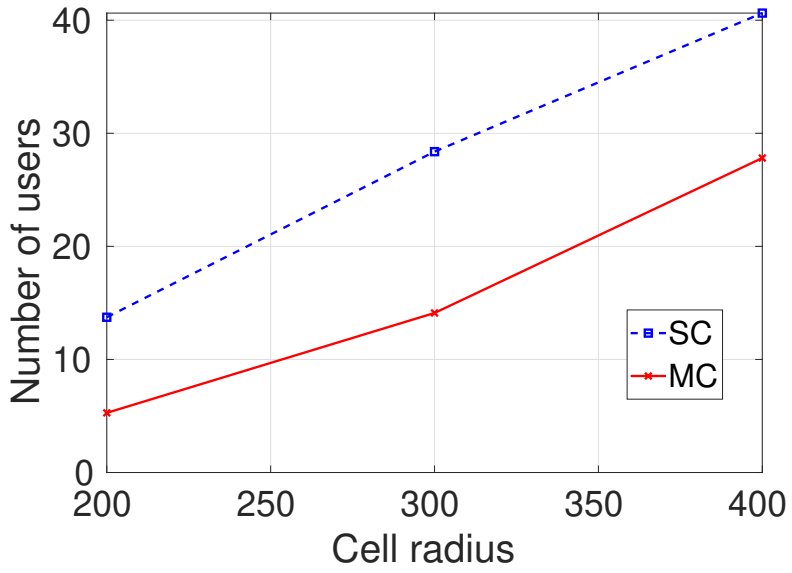


Figure 6.8: Average number unserved users under greedy approximation algorithm as a function of cell radius

same MBSFN. MBSFN requires the multicast content to be transmitted over the same PRB by all eNBs. For resource allocation in MBSFN, we choose a PRB that serves the maximum number of UEs in the entire system. Here too, we use traces of the video of Tokyo Olympics to generate realistic video traffic patterns. We observe that MC

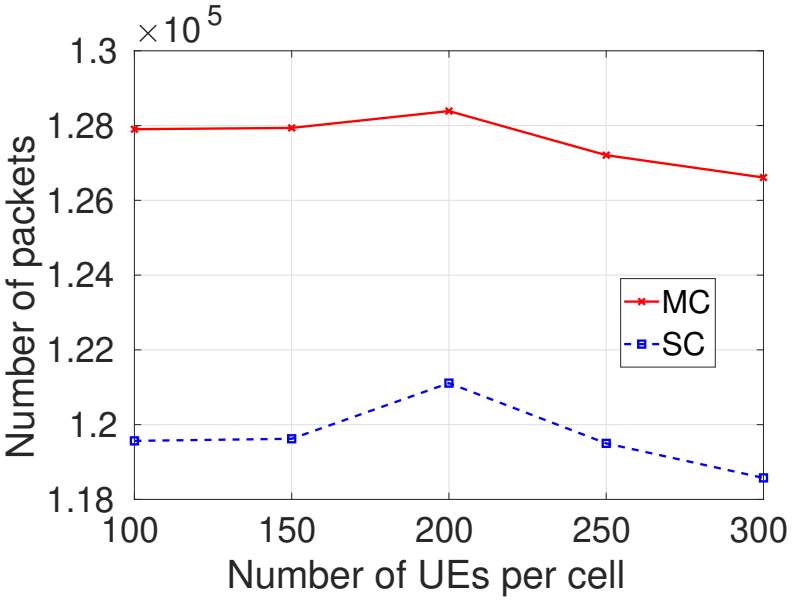


Figure 6.9: Average number of packets received successfully under MC and SC multicast for a real-time video stream

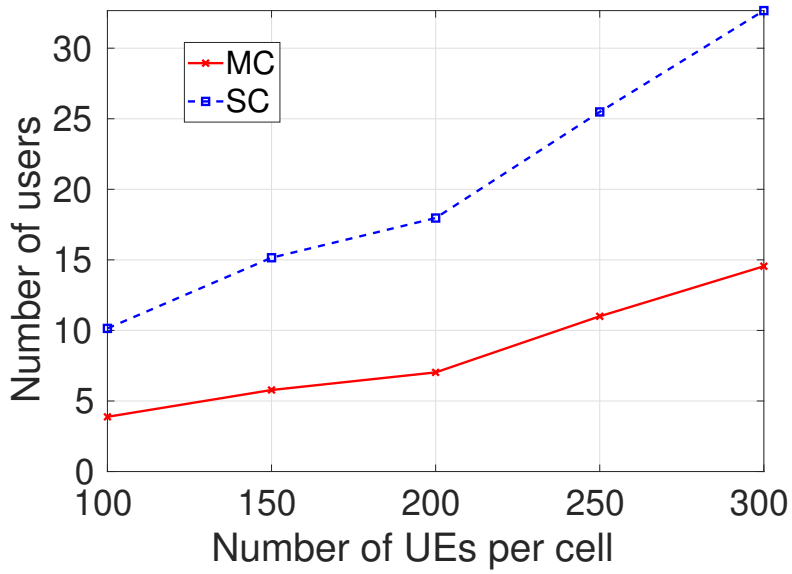


Figure 6.10: Average number of unserved UEs under MC and SC multicast for a real-time video stream

multicast performs remarkably better than MBSFN. It succeeds in delivering a much larger number of packets successfully and is able to serve significantly more UEs than MBSFN. While many UEs remain unserved under MBSFN, nearly all of them are served under MC multicast. These results confirm our claims that MC multicast can provide the

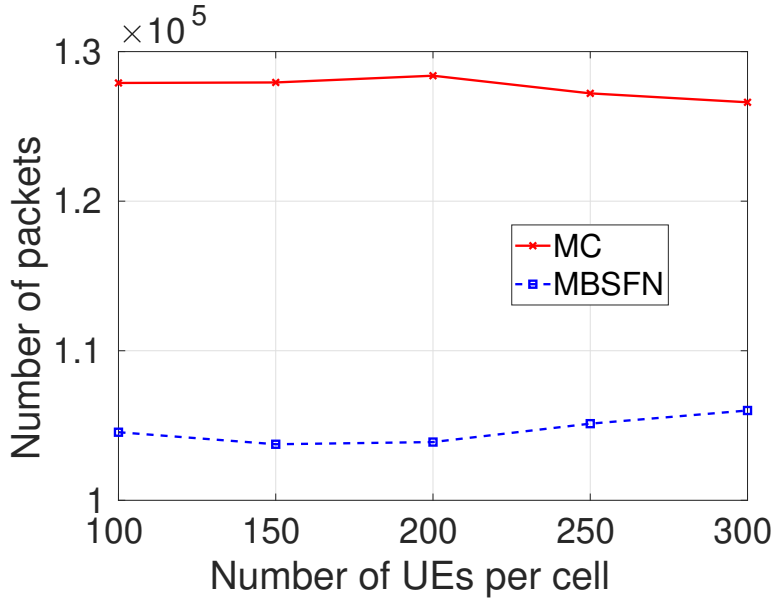


Figure 6.11: Average number of packets received successfully under MC multicast and MBSFN for a real-time video stream

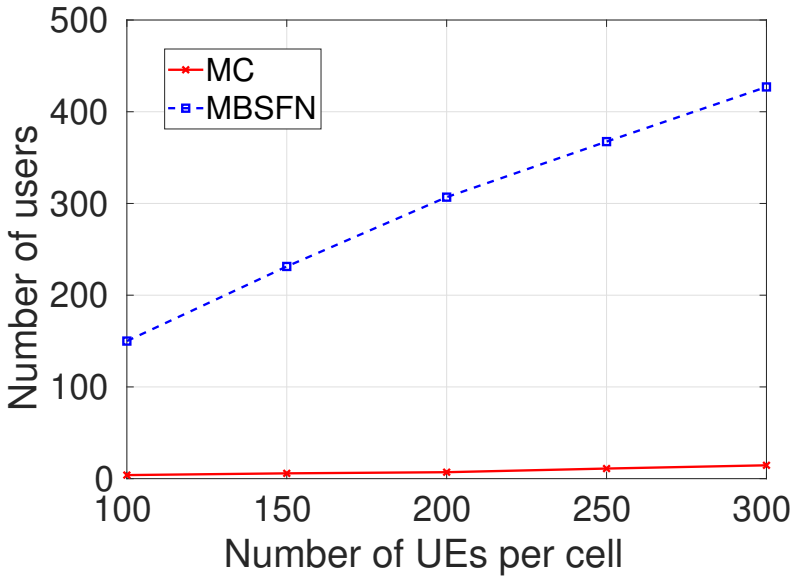


Figure 6.12: Average number unserved users under MC multicast and MBSFN for a real-time video stream

benefits of MBSFNs while eliminating the need for synchronization. In fact, as observed in Figures 6.11 and 6.12, MC multicast outperforms MBSFN by large margins.

These simulation results clearly indicate that using MC with multicast provides a significant performance enhancement in multicast systems. The flexibility of potentially

connecting to multiple eNBs results in more users being served each sub-frame. Thus, MC multicast has great potential for use in video streaming services. It can help alleviate the burden on network resources while serving the increasing video streaming traffic.

6.8 Conclusions

In this chapter, we have proposed the use of multi-connectivity in multicast transmissions. We have proposed procedures for enabling the use of multi-connectivity in MBMS. We have formulated the problem of resource allocation in multi-connected multicast systems with the aim of maximizing the number of users served. We have proved this to be an NP-hard problem. We have proposed a centralized greedy approximation algorithm for resource allocation that provides an approximation ratio of $\left(1 - \frac{1}{e}\right)$. No polynomial-time algorithm can provide a better approximation. We have also proposed a distributed heuristic resource allocation algorithm for multi-connected multicast systems. Through extensive simulations, we have established that the use of multi-connectivity in multicast transmissions significantly improves the system performance. Multi-connectivity enables serving a much larger number of users. We have also studied the performance of multi-connectivity in serving real-time video streaming applications. To generate video specific traffic patterns in these simulations, we have used traces from actual videos [16]. We have also compared the performance of multi-connectivity multicast to that of MBSFNs. Our simulation results indicate that multi-connectivity outperforms MBSFNs by large margins, while eliminating the need for strict synchronization and extended cyclic prefixes.

Chapter 7

Summary of Results and Future Directions

7.1 Summary of Results

Due to unprecedented growth of bandwidth intensive video traffic, there is a pressing need for resource allocation algorithms and streaming mechanisms that can efficiently utilize the available bandwidth. A large part of this video traffic comprises content like streaming of TV shows, live news feeds, live telecast of sports events, movie premieres, live feeds of major world events. All these applications involve the same content being requested by a large audience simultaneously. Multicast transmission provides an efficient means of serving such applications. In this thesis, we have addressed the problems of grouping and resource allocation for multicast transmissions, primarily for video streaming. We have discussed most of the problems in this work in the context of an LTE system. However, all the algorithms proposed are generic and can be easily modified for use in any wireless mobile communication system.

In services like online premieres, video quality takes precedence over other parameters. For such services, we consider a lossless multicast system in Chapter 3 in which all users need to be served at a specific rate in each sub-frame. The corresponding resource allocation problem is a binary integer linear program that minimizes the number of PRBs used for multicast transmissions while ensuring that the rate requirements of all multicast users are met. We have proved that this is an NP-hard problem, and hence,

no polynomial-time algorithms exist for determining its optimal solution. As discussed in Chapter 2, minimizing the resource utilization of multicast services is essential for their successful co-existence with a plethora of other services in the network. To the best of our knowledge, this is the first work that addresses this minimization problem.

Since the optimal solution of the resource allocation problem cannot be determined in polynomial-time, we have designed a randomized scheme that estimates the optimal solution. This randomized scheme is based on Simulated Annealing, a Markov Chain Monte Carlo technique. The scheme traverses the solution space of feasible resource allocations and eventually converges to the optimal solution with high probability. However, due to time considerations, this iterative scheme cannot be used for resource allocation in practice. So, we have designed two online heuristic resource allocation schemes, a greedy, and an LP-relaxation based scheme. These schemes run in polynomial-time and provide solutions close to the optimal. The solution provided by the randomized scheme acts as a benchmark for evaluating the performance of these heuristic schemes. The LP-relaxation scheme results in feasible resource allocations that save nearly as many PRBs as that saved by the randomized scheme in about one-fifth of the time taken by it. Extensive simulations have been conducted to compare the performance of the proposed policies with the existing state of the art. We have shown that our policies provide significant gains over traditional unicast transmissions in terms of conserving PRBs and maximizing the number of users served. To further establish the practical applicability of our policies, we have also evaluated their performance using traces from actual video streams. We have shown that our policies successfully meet the rate requirements of these video streams.

As discussed in Chapter 1, due to varying channel states of users, grouping users based only on the content required by them may result in the performance of multicast being worse than that of unicast transmissions. To prevent this, channel states of users need to be taken into consideration while grouping. In Chapter 3, we have formulated the grouping problem to minimize the resource usage under any resource allocation policy. Variation of the channel states of users over time and frequency makes the grouping problem complex. In fact, we have proved that the optimal grouping problem is also NP-hard and no polynomial-time algorithms exist for determining its optimal solution. Therefore, we have designed a heuristic algorithm for grouping, the hybrid grouping policy.

This policy groups users based on their content requirements and average channel states. Extensive simulations carried out in an LTE environment establish the effectiveness of the proposed grouping policy. The grouping and resource allocation schemes proposed in Chapter 3 can act as an enhancement to MBMS to make its multicast operations more efficient and versatile.

In lossless multicast transmissions, the system is constrained to serve all users in each sub-frame. Hence, data for a multicast group can never be transmitted at a rate higher than what can be successfully decoded by the weakest user in that group. This makes the system performance dependent on the weakest users in it. It may also result in dissatisfaction of users with good channel conditions who could receive significantly better rates using unicast. To address these issues, we propose the use of loss tolerant multicast in Chapter 4. Video streams can tolerate some amount of packet loss without any significant degradation in quality. We leverage this property to design resource allocation algorithms for multicast video streaming that allow for some controlled packet losses. The loss tolerant multicast system is, therefore, not constrained to serve all users in each sub-frame. In the loss tolerant model for MBMS in Chapter 4, users have some tolerance for packet loss. The loss tolerance of a user is dictated by factors like the type of video being streamed, the subscribed data plan, the type of device being used for streaming.

For this loss tolerant MBMS system, we have proposed two loss optimal online resource allocation policies that can meet the loss requirements of all users. The proposed policies are named Loss Optimal Resource Allocation (LORA) and priority LORA (p-LORA). We have proved the throughput optimality of both these policies, which means that these policies can meet the loss requirements of the system as long as any online or offline policy can do so. p-LORA improves upon LORA by introducing a prioritization mechanism that prevents consecutive packet losses to ensure that no user is left unserved for long periods at a stretch. We have also generalized the EXP-Q rule [18], a well-known throughput optimal policy, for use with multiple channels and multicast transmission. We have used this modified EXP-Q rule as a benchmark for evaluating the performance of the proposed policies. Simulations have been performed to study and compare the loss performance of the proposed schemes with that of the modified EXP-Q rule. Since these policies are specifically designed for video streaming, we have used traces from

actual videos in these simulations. We have observed that the proposed policies succeed in meeting the loss requirements of all users, whereas the modified EXP-Q rule fails to do so. p-LORA results in the least packet loss of the three policies. It also provides the best performance in terms of the burstiness of the losses encountered. Since PSNR is considered to be the parameter of choice for defining the quality of videos, we have also compared the PSNR of the received video streams under the three policies. LORA and p-LORA result in significantly better PSNR than the modified EXP-Q rule. Using these policies for resource allocation in multicast video streaming can significantly improve the network bandwidth utilization.

In Chapter 5, we have addressed the need for a generalized resource allocation algorithm that can adjust to optimize any parameter of the system. Such algorithms are essential for present day cellular mobile systems that typically cater to a wide variety of users, services, and devices. Fulfilling the requirements of a diverse set of users and services requires optimizing several different parameters. In light of these requirements, we have designed a generalized resource allocation algorithm based on the Vickrey-Clarke-Groves (VCG) mechanism. This algorithm is capable of meeting QoS requirements of a heterogeneous mix of users and services. It provides a unified framework for serving unicast and multicast users and can be used irrespective of the optimization objective of resource allocation. It can simultaneously allocate resources to meet the demands of different kinds of traffic that may have different QoS requirements. Each user has a certain valuation for the system resources. This valuation is a user's private information, and we have not assumed any structure on it.

The auction based algorithm takes the bids conveyed by the users as inputs and outputs a feasible resource allocation. It also determines the prices to be paid by the users in accordance with the quality experienced by them. We have proved that the proposed algorithm is strategy-proof and so, there is no incentive for users to not bid their true valuations. This ensures that the system resources are efficiently utilized for maximizing the social welfare. As discussed in Chapter 5, VCG mechanisms, in general, are NP-hard to implement. However, in this case, the proposed algorithm can be implemented in polynomial-time. We have proposed an efficient polynomial-time implementation of our algorithm using maximum weight bipartite matching. We have compared the performance

of our algorithm with that of a throughput maximizing greedy algorithm. We have shown that, even though our algorithm makes allocation decisions based only on the valuations reported by the users, without any prior knowledge of their requirements, it succeeds in meeting the QoS requirements of all users. Our simulation results establish the effectiveness of the proposed algorithm for meeting the service requirements of a heterogeneous mix of users irrespective of the nature of their valuations.

In Chapter 6, we have proposed the use of multi-connectivity in multicast transmissions. Multi-connectivity in multicast enables a user to simultaneously receive the same multicast content from multiple eNBs. It provides a method for attaining the benefits of MBSFNs without their need for extended cyclic prefix and strict synchronization between the eNBs. 3GPP architecture for MBMS defines an additional SYNC protocol layer to synchronize the MBMS content delivered to the eNBs [117]. As a result, the content received at the eNBs served by the same MBMS-GW is in sync. Multi-connectivity multicast takes advantage of this to enable users to independently receive the same multicast content from multiple eNBs. Without the need for strict synchronization between the eNBs, each eNB can optimize its resource allocation independently. The users receive the same content in different PRBs, and the resulting diversity significantly improves the received SNR of the multicast users. We have also defined detailed procedures required for establishing multi-connectivity in an MBMS system.

We have formulated the optimal resource allocation problem in a multi connected multicast system with the objective of maximizing the number of users who successfully receive MBMS content. We have proved that this is an NP-hard problem. Therefore, we have proposed a greedy resource allocation algorithm that provides an approximation factor of $(1 - \frac{1}{e})$. This is, in fact, the best possible approximation ratio for this problem. We have also proposed an uncoordinated resource allocation algorithm for multi connected multicast systems. We have shown via simulations that even an uncoordinated policy yields significant performance gains over a single connected system. Our impact assessment also reveals that multi-connectivity provides only marginal gains over a dual connected system. Hence, the proposed schemes can be implemented with dual connectivity in the existing framework of LTE and 5G while providing performance improvements close to a multi connected system.

7.2 Future Research Directions

Resource allocation algorithms proposed in Chapters 4 and 5 assign a single PRB to each user/multicast group in a sub-frame. As a future direction, we can develop more flexible algorithms that can allocate any number of PRBs to the users/multicast groups. The problem of allocating multiple PRBs in these systems is quite complex because of the variation of channel states of users across PRBs. For the auction based algorithm in Chapter 5, we can explore the use of repeated auctions for allocating resources in consecutive sub-frames.

The problems discussed in this thesis consider a homogeneous LTE system. However, Heterogeneous Networks (HetNets) are rapidly becoming an important part of the cellular mobile networks. We can, therefore, develop algorithms for implementing efficient multicast transmissions in HetNets as well. Since HetNets also use Wireless Fidelity (WiFi) access points alongside LTE and 5G base stations, we can also develop procedures and algorithms for multicasting content via WiFi access points.

The primary aim of this thesis is to develop methods and techniques for using the spectrum more efficiently so that the increasing demand for bandwidth can be supported. Having established the effectiveness of multicast in this regard, we can extend this work to include Device to Device (D2D) multicast. Multicast using underlay D2D communications can be used for further improving spectral efficiency. Future work in this direction can address the problems of interference management and power control to ensure that cellular unicast and multicast operations remain unaffected by D2D multicast. Clustering D2D users for multicast operations will also have to be addressed.

In the current standards for MBMS, multicast can only be used for transmitting MBMS sessions available in the network. However, with increasing popularity of online streaming, multicast transmissions can find many diverse use cases. Consider, for instance, an episode of a popular series available only from a streaming platform in a particular region. A large number of users would be streaming the episode simultaneously. Currently, the service provider has no way of knowing that the same content being streamed by several users. This information is only known to the content provider. For such use cases, frameworks can be developed to enable secure cooperation between content and service providers. This will enable content providers to share the details of the

content being streamed with service providers. Service providers can then combine these individual streams and serve them efficiently using multicast transmissions. To enable such interactions, the associated security and economic issues also need to be addressed. Policies can be designed to ensure that this is beneficial for both content as well as service providers.

The algorithms proposed in this work have the potential to improve the performance of multicast operations significantly. Our extensive simulations demonstrate that implementing these algorithms in cellular mobile networks can prove extremely beneficial for optimizing video data streaming.

Bibliography

[1] 3GPP TR 36.931 v.9.0.0 Rel. 9, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) requirements for LTE Pico Node B," 2011. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/136900_136999/136931/09.00.00_60/tr_136931v090000p.pdf.

[2] 3GPP TR 38.901 v.15.0.0 Rel. 15, "5G; Study on channel model for frequencies from 0.5 to 100 GHz," 2018. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/138900_138999/138901/15.00.00_60/tr_138901v150000p.pdf.

[3] Cisco, "Global Mobile Data Traffic Forecast Update, 2017–2022," *Cisco White paper*. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html.

[4] TRAI, "Wireless data services in India, An analytical report," 2019. [Online]. Available: https://main.trai.gov.in/sites/default/files/Wireless_Data_Service_Report_21082019_0.pdf.

[5] N. Heuveldop, "Ericsson Mobility Report," 2017. [Online]. Available: https://www.ericsson.com/assets/local/mobility-report/documents/2017/ericsson-mobility-report-june-2017.pdf.

[6] "LTE-Broadcast (eMBMS) Market Update," 2018. [Online]. Available: https://gsacom.com/paper/lte-broadcast-embms-market-update-2/.

[7] 3GPP TS 36.213 v.15.5.0 Rel. 15, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures," 2019. [Online]. Avail-

able: https://www.etsi.org/deliver/etsi_ts/136200_136299/136213/15.05.00_60/ts_136213v150500p.pdf.

[8] 3GPP R1-081483, "Conveying MCS and TB size via PDCCH," 2008. [Online]. Available: https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_52b/Docs/.

[9] H. Velde, O. Hus, and M. Baker, "Broadcast Operation," in *LTE-The UMTS Long Term Evolution From Theory to Practice*, pp. 293–305, Chichester: John Wiley & Sons Ltd, 2 ed., 2011.

[10] 3GPP TS 23.246 v.14.1.0 Rel. 14, "Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description," 2017. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/123200_123299/123246/14.01.00_60/ts_123246v140100p.pdf.

[11] 3GPP TS 36.304 v.15.4.0 Rel. 15, "LTE; E-UTRA; User Equipment (UE) procedures in idle mode," 2019. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/136300_136399/136304/15.04.00_60/ts_136304v150400p.pdf.

[12] 3GPP TS 36.212 v.14.4.0 Rel. 14, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding," 2017. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/136200_136299/136212/14.04.00_60/ts_136212v140400p.pdf.

[13] Y. L. Chang, T. L. Lin, and P. C. Cosman, "Network-Based H.264/AVC Whole-Frame Loss Visibility Model and Frame Dropping Methods," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3353–3363, 2012.

[14] N. Sharma and A. Madhukumar, "Genetic algorithm aided proportional fair resource allocation in multicast OFDM systems," *IEEE Transactions on Broadcasting*, vol. 61, no. 1, pp. 16–29, 2015.

[15] D. Lee, J. So, and S. R. Lee, "Power allocation and subcarrier assignment for joint delivery of unicast and broadcast transmissions in OFDM systems," *Journal of Communications and Networks*, vol. 18, no. 3, pp. 375–386, 2016.

[16] P. Seeling and M. Reisslein, "Video transport evaluation with H. 264 video traces," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 4, pp. 1142–1165, 2011.

[17] G. Van der Auwera, P. T. David, and M. Reisslein, "Traffic and quality characterization of single-layer video streams encoded with the H.264/MPEG-4 advanced video coding standard and scalable video coding extension," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 698–718, 2008.

[18] S. Shakkottai and A. L. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule," *Translations of the American Mathematical Society-Series 2*, vol. 207, pp. 185–202, 2002.

[19] G. Araniti, V. Scordamaglia, A. Molinaro, A. Iera, G. Interdonato, and F. Span, "Optimizing point-to-multipoint transmissions in High Speed Packet Access networks," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp. 1–5, 2011.

[20] G. Araniti, M. Condoluci, and A. Iera, "Adaptive multicast scheduling for HSDPA networks in mobile scenarios," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp. 1–5, 2012.

[21] L. Militano, D. Niyato, M. Condoluci, G. Araniti, A. Iera, and G. M. Bisci, "Radio resource management for group-oriented services in LTE-A," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 8, pp. 3725–3739, 2015.

[22] L. Militano, M. Condoluci, G. Araniti, and A. Iera, "Bargaining solutions for multicast subgroup formation in LTE," in *IEEE Vehicular Technology Conference fall*, pp. 1–5, 2012.

[23] L. Militano, M. Condoluci, G. Araniti, and A. Iera, "Multicast service delivery solutions in LTE-Advanced systems," in *IEEE International Conference on Communications*, pp. 5954–5958, 2013.

[24] M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and J. Cosmas, "On the impact of frequency selectivity on multicast subgroup formation in 4G networks," in *IEEE In-*

ternational Symposium on Broadband Multimedia Systems and Broadcasting, pp. 1–6, 2013.

[25] T. Liu, H. Xia, and C. Feng, "A QoS-based multi-rate multicast scheme over heterogeneous cellular network," in *IEEE International Symposium on Wireless Communication Systems*, pp. 292–296, 2016.

[26] G. Araniti, M. Condoluci, L. Militano, and A. Iera, "Adaptive Resource Allocation to Multicast Services in LTE Systems," *IEEE Transactions on Broadcasting*, vol. 59, no. 4, pp. 658–664, 2013.

[27] T. P. Low, M. O. Pun, Y. W. P. Hong, and C. C. J. Kuo, "Optimized opportunistic multicast scheduling (OMS) over wireless cellular networks," *IEEE Transactions on Wireless Communications*, vol. 9, no. 2, pp. 791–801, 2010.

[28] P. Polacek, T.-Y. Yang, and C.-W. Huang, "Opportunistic multicasting for single frequency networks," *Wiley Wireless Communications and Mobile Computing*, vol. 16, no. 15, pp. 2253–2262, 2016.

[29] S. Lu, Y. Cai, L. Zhang, J. Li, P. Skov, C. Wang, and Z. He, "Channel-Aware Frequency Domain Packet Scheduling for MBMS in LTE," in *IEEE Vehicular Technology Conference spring*, pp. 1–5, 2009.

[30] I. C. Wong, Z. Shen, B. L. Evans, and J. G. Andrews, "A low complexity algorithm for proportional resource allocation in OFDMA systems," in *IEEE Workshop on Signal Processing Systems*, pp. 1–6, 2004.

[31] G. Araniti, A. Orsino, J. Cosmas, A. Molinaro, and A. Iera, "A low computational-cost subgrouping multicast scheme for emerging 5G-satellite networks," in *IEEE Broadband Multimedia Systems and Broadcasting*, pp. 1–6, 2016.

[32] A. Orsino, G. Araniti, P. Scopelliti, I. Gudkova, K. Samouylov, and A. Iera, "Optimal subgroup configuration for multicast services over 5G-satellite systems," in *IEEE Broadband Multimedia Systems and Broadcasting*, pp. 1–6, 2017.

[33] A. de la Fuente, G. Femenias, F. Riera-Palou, and A. G. Armada, "Subband CQI Feedback-Based Multicast Resource Allocation in MIMO-OFDMA Networks," *IEEE Transactions on Broadcasting*, vol. 64, no. 4, pp. 846–864, 2018.

[34] G. Araniti, M. Condoluci, M. Cotronei, A. Iera, and A. Molinaro, "A solution to the multicast subgroup formation problem in LTE systems," *IEEE Wireless Communications Letters*, vol. 4, no. 2, pp. 149–152, 2015.

[35] J. F. Monserrat, J. Calabuig, A. Fernandez-Aguilella, and D. Gomez-Barquero, "Joint delivery of unicast and E-MBMS services in LTE networks," *IEEE Transactions on Broadcasting*, vol. 58, no. 2, pp. 157–167, 2012.

[36] J. Chen, M. Chiang, J. Erman, G. Li, K. K. Ramakrishnan, and R. K. Sinha, "Fair and optimal resource allocation for LTE multicast (eMBMS): Group partitioning and dynamics," in *IEEE Conference on Computer Communications*, pp. 1266–1274, 2015.

[37] A. de la Fuente, A. G. Armada, and R. P. Leal, "Joint multicast/unicast scheduling with dynamic optimization for LTE multicast service," in *European Wireless Conference*, pp. 1–6, 2014.

[38] 3GPP TS 26.247 v.10.0.0 Rel. 10, "Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH)," 2011. [Online]. Available: `https://www.etsi.org/deliver/etsi_ts/126200_126299/126247/10.00.00_60/ts_126247v100000p.pdf`.

[39] J. Park, J. N. Hwang, Q. Li, Y. Xu, and W. Huang, "Optimal DASH-multicasting over LTE," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4487–4500, 2018.

[40] T. Paila, R. Walsh, M. Luby, V. Roca, and R. Lehtonen, "FLUTE-file delivery over unidirectional transport," 2012. Internet Engineering Task Force Request for Comments: 6726.

[41] C.-N. Lee and H.-T. Lai, "Pricing based resource allocation scheme for video multicast service in LTE networks," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–5, IEEE, 2016.

[42] A. Tassi, I. Chatzigeorgiou, and D. Vukobratović, "Resource-allocation frameworks for network-coded layered multimedia multicast services," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 2, pp. 141–155, 2015.

[43] A. Tassi, I. Chatzigeorgiou, D. Vukobratović, and A. L. Jones, "Optimized network-coded scalable video multicasting over eMBMS networks," in *IEEE International Conference on Communications*, pp. 3069–3075, 2015.

[44] R. O. Afolabi, A. Dadlani, and K. Kim, "Multicast scheduling and resource allocation algorithms for OFDMA-based systems: A survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 240–254, 2013.

[45] R. Kaliski, C. C. Chou, H. Y. Meng, and H. Y. Wei, "Dynamic Resource Allocation Framework for MooD (MBMS Operation On-Demand)," *IEEE Transactions on Broadcasting*, vol. 62, no. 4, pp. 903–917, 2016.

[46] H.-Y. Meng, C.-C. Chou, R. Kaliski, and H.-Y. Wei, "An on-demand QoE resource allocation algorithm for multi-flow LTE eMBMS," in *IEEE Wireless and Optical Communication Conference*, pp. 93–97, 2015.

[47] O. Karimi, J. Liu, and Z. Wang, "Power-Efficient Resource Utilization in Cellular Multimedia Multicast," in *IEEE International Conference on Mobile Ad-hoc and Sensor Networks*, pp. 134–143, 2015.

[48] A. Orsino, P. Scopelliti, and M. Condoluci, "A Multi-Criteria Approach for Multicast Resource Allocation over LTE and Beyond Cellular Systems," in *European Wireless Conference*, pp. 1–6, 2016.

[49] C.-L. Hwang, Y.-J. Lai, and T.-Y. Liu, "A new approach for multiple objective decision making," *Elsevier Computers & Operations Research*, vol. 20, no. 8, pp. 889–899, 1993.

[50] G. Araniti, M. Condoluci, A. Orsino, A. Iera, A. Molinaro, and J. Cosmas, "Evaluating the performance of multicast resource allocation policies over LTE systems," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp. 1–6, 2015.

[51] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[52] "More Efficient Mobile Encodes for Netflix Downloads," 2016. [Online]. Available: `https://medium.com/netflix-techblog/more-efficient-mobile-encodes-for-netflix-downloads-625d7b082909`.

[53] "VP9: Faster, better, buffer-free YouTube videos," 2015. [Online]. Available: `https://youtube-eng.googleblog.com/2015/04/vp9-faster-better-buffer-free-youtube.html`.

[54] M. Condoluci, G. Araniti, A. Molinaro, and A. Iera, "Multicast resource allocation enhanced by channel state feedbacks for multiple scalable video coding streams in LTE networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 2907–2921, 2016.

[55] C.-W. Huang, S.-M. Huang, P.-H. Wu, S.-J. Lin, and J.-N. Hwang, "OLM: Opportunistic layered multicasting for scalable IPTV over mobile WiMAX," *IEEE Transactions on Mobile Computing*, vol. 11, no. 3, pp. 453–463, 2012.

[56] M. Tang, S. Wang, L. Gao, J. Huang, and L. Sun, "MOMD: A multi-object multi-dimensional auction for crowdsourced mobile video streaming," in *IEEE Conference on Computer Communications*, pp. 1–9, 2017.

[57] M. Tang, L. Gao, H. Pang, J. Huang, and L. Sun, "A multi-dimensional auction mechanism for mobile crowdsourced video streaming," in *IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, pp. 1–8, 2016.

[58] C.-H. Ko, C.-C. Chou, H.-Y. Meng, and H.-Y. Wei, "Strategy-proof resource allocation mechanism for multi-flow wireless multicast," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3143–3156, 2015.

[59] G. C. Sekhar, S. Parakh, and A. K. Jagannatham, "Auction based optimal subcarrier allocation for H. 264 scalable video transmission in 4G OFDMA systems," in *IEEE India Conference*, pp. 18–23, 2012.

[60] S. Parakh and A. K. Jagannatham, "VCG auction based optimal allocation for scalable video communication in 4G WiMAX," in *National Conference on Communications*, pp. 1–5, IEEE, 2012.

[61] S. Parakh and A. K. Jagannatham, "Optimal resource allocation and VCG auction-based pricing for H. 264 scalable video quality maximization in 4G wireless systems," *Advances in Multimedia, Hindawi Publishing Corp.*, vol. 2012, p. 3, 2012.

[62] A. Sinha and A. Anastasopoulos, "Generalized proportional allocation mechanism design for multi-rate multicast service on the internet," in *Allerton Conference on Communication, Control, and Computing*, pp. 146–153, IEEE, 2013.

[63] C.-Y. Wang, Y. Chen, H.-Y. Wei, and K. R. Liu, "Optimal pricing in stochastic scalable video coding multicasting system," in *IEEE Conference on Computer Communications*, pp. 540–544, 2013.

[64] A. Bradai, T. Ahmed, R. Boutaba, and R. Ahmed, "Efficient content delivery scheme for layered video streaming in large-scale networks," *Elsevier Journal of Network and Computer Applications*, vol. 45, pp. 1–14, 2014.

[65] F. Wu and N. Vaidya, "A strategy-proof radio spectrum auction mechanism in non-cooperative wireless networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 5, pp. 885–894, 2013.

[66] L. Sun, H. Shan, A. Huang, L. Cai, and H. He, "Channel Allocation for Adaptive Video Streaming in Vehicular Networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 734–747, 2017.

[67] H. Meshgi, D. Zhao, and R. Zheng, "Optimal resource allocation in multicast device-to-device communications underlaying lte networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 8357–8371, 2017.

[68] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP," in *ACM Conference on Multimedia systems*, pp. 157–168, 2011.

[69] N. Saxena, S. Singh, A. Roy, and D. H. Ail, "NEST: novel eMBMS scheduling technique," *Springer Wireless Networks*, vol. 22, no. 6, pp. 1837–1850, 2016.

[70] A. Roy and N. Saxena, *Data scheduling and transmission strategies in asymmetric telecommunication environments*. CRC Press, 2008.

[71] R. Sivaraj, M. Arslan, K. Sundaresan, S. Rangarajan, and P. Mohapatra, "BoLTE: Efficient network-wide LTE broadcasting," in *IEEE International Conference on Network Protocols*, pp. 1–10, 2017.

[72] G. Călinescu, H. Karloff, and Y. Rabani, "An improved approximation algorithm for multiway cut," in *ACM Symposium on Theory of Computing*, pp. 48–52, 1998.

[73] K. Sundaresan and S. Rangarajan, "Scheduling algorithms for video multicasting with channel diversity in wireless OFDMA networks," in *ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 20:1–20:11, 2011.

[74] K. Sundaresan and S. Rangarajan, "Cooperation versus multiplexing: Multicast scheduling algorithms for OFDMA relay networks," *IEEE/ACM Transactions on Networking*, vol. 22, no. 3, pp. 756–769, 2014.

[75] N.-M. Cheung, A. Ortega, and G. Cheung, "Distributed source coding techniques for interactive multiview video streaming," in *Picture Coding Symposium*, pp. 1–4, IEEE, 2009.

[76] Z. Liu, G. Cheung, and Y. Ji, "Unified distributed source coding frames for interactive multiview video streaming," in *IEEE International Conference on Communications*, pp. 2048–2053, 2012.

[77] Z. Liu, G. Cheung, and Y. Ji, "Optimizing distributed source coding for interactive multiview video streaming over lossy networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 10, pp. 1781–1794, 2013.

[78] J. K. Sundararajan, D. Shah, and M. Médard, "ARQ for network coding," in *IEEE International Symposium on Information Theory*, pp. 1651–1655, 2008.

[79] P. Sadeghi, R. Shams, and D. Traskov, "An optimal adaptive network coding scheme for minimizing decoding delay in broadcast erasure channels," *EURASIP Journal on Wireless Communications and Networking*, vol. 2010, pp. 1–14, 2010.

[80] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4413–4430, 2006.

[81] R. Ahlswede, N. Cai, S.-Y. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on information theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[82] G. Joshi, Y. Kochman, and G. W. Wornell, "Throughput-smoothness trade-offs in multicasting of an ordered packet stream," in *International Symposium on Network Coding*, pp. 1–6, IEEE, 2014.

[83] G. Joshi, Y. Kochman, and G. Wornell, "On Throughput-Smoothness Trade-offs in Streaming Communication," *arXiv preprint arXiv:1511.08143*, 2015.

[84] G. Joshi, Y. Kochman, and G. W. Wornell, "On playback delay in streaming communication," in *IEEE International Symposium on Information Theory Proceedings*, pp. 2856–2860, 2012.

[85] L. Keller, E. Drinea, and C. Fragouli, "Online broadcasting with network coding," in *Workshop on Network Coding, Theory and Applications*, pp. 1–6, IEEE, 2008.

[86] J. K. Sundararajan, P. Sadeghi, and M. Médard, "A feedback-based adaptive broadcast coding scheme for reducing in-order delivery delay," in *Workshop on Network Coding, Theory, and Applications*, pp. 1–6, IEEE, 2009.

[87] A. Badr, A. Khisti, W.-T. Tan, and J. Apostolopoulos, "Streaming codes for channels with burst and isolated erasures," in *IEEE International Conference on Computer Communications.*, pp. 2850–2858, 2013.

[88] A. Badr, D. Lui, and A. Khisti, "Streaming codes for multicast over burst erasure channels," *IEEE Transactions on Information Theory*, vol. 61, no. 8, pp. 4181–4208, 2015.

[89] C. Rosa, K. Pedersen, H. Wang, P.-H. Michaelsen, S. Barbera, E. Malkamaki, T. Henttonen, and B. Sébire, "Dual connectivity for LTE small cell evolution: Functionality and performance aspects," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 137–143, 2016.

[90] 3GPP TS 37.340 v.15.3.0 Rel. 15, "5G; NR; Multi-connectivity; Overall description," 2018. [Online]. Available: `https://www.etsi.org/deliver/etsi_ts/137300_137399/137340/15.03.00_60/ts_137340v150300p.pdf`.

[91] R. Odarchenko, R. L. Aguiar, B. Altman, and Y. Sulema, "Multilink approach for the content delivery in 5G networks," in *International Scientific-Practical Conference Problems of Infocommunications. Science and Technology*, pp. 140–144, 2018.

[92] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Mobility modeling and performance evaluation of multi-connectivity in 5G intra-frequency networks," in *IEEE Global Communications Conference Workshops*, pp. 1–6, 2015.

[93] M.-S. Pan, T.-M. Lin, C.-Y. Chiu, and C.-Y. Wang, "Downlink traffic scheduling for LTE-A small cell networks with dual connectivity enhancement," *IEEE Communications Letters*, vol. 20, no. 4, pp. 796–799, 2016.

[94] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved handover through dual connectivity in 5G mmWave mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2069–2084, 2017.

[95] H. Wang, C. Rosa, and K. I. Pedersen, "Dual connectivity for LTE-advanced heterogeneous networks," *Springer Wireless Networks*, vol. 22, no. 4, pp. 1315–1328, 2016.

[96] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera, "Multicasting over emerging 5G networks: Challenges and perspectives," *IEEE Network*, vol. 31, no. 2, pp. 80–89, 2017.

[97] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson, "NP-Completeness," in *Introduction to Algorithms*, ch. 34, MIT Press, 3rd ed., 2009.

[98] M. R. Garey and D. S. Johnson, "Using NP-Completeness to Analyze Problems," in *Computers and Intractability A guide to the theory of NP-completeness*, vol. 29, ch. 4, W.H Freeman and Company, New York, 2002.

[99] R. M. Karp, "On the computational complexity of combinatorial problems," *Wiley Networks*, vol. 5, no. 1, pp. 45–68, 1975.

[100] S. M. Ross, "Markov Chain Monte Carlo Methods," in *Simulation*, pp. 271–302, Academic Press, 5 ed., 2013.

[101] M. Harchol-Balter, "Time-Reversibility and Burke's Theorem," in *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*, ch. 16, Cambridge University Press, 1 ed., 2013.

[102] B. Hajek, "Cooling schedules for optimal annealing," *Mathematics of Operations Research*, vol. 13, no. 2, pp. 311–329, 1988.

[103] MATLAB Users Guide, "`https://in.mathworks.com/products/matlab/`."

[104] M. Mehta, *Radio Resource and Mobility Management Techniques in Heterogeneous Cellular Network*. PhD thesis, Department of Electrical Engineering, IIT Bombay, Mumbai, India, 2014.

[105] M. Series, "Guidelines for evaluation of radio interface technologies for IMT-Advanced," *Report ITU*, vol. 638, 2009.

[106] T. Begluk, J. B. Husić, and S. Baraković, "Machine learning-based qoe prediction for video streaming over LTE network," in *International Symposium INFOTEH-JAHORINA*, pp. 1–5, 2018.

[107] M. Vaser and S. Forconi, "QoS KPI and QoE KQI relationship for LTE video streaming and VoLTE services," in *International Conference on Next Generation Mobile Applications, Services and Technologies*, pp. 318–323, 2015.

[108] S. Aroussi, T. Bouabana-Tebibel, and A. Mellouk, "Empirical QoE/QoS correlation model based on multiple parameters for VoD flows," in *IEEE Global Communications Conference*, pp. 1963–1968, 2012.

[109] 3GPP TS 38.214 v.15.5.0 Rel. 15, "5G; NR; Physical layer procedures for data," 2019. [Online]. Available: `https://www.etsi.org/deliver/etsi_ts/138200_138299/138214/15.05.00_60/ts_138214v150500p.pdf`.

[110] 3GPP TS 38.211 v.15.5.0 Rel. 15, "5G; NR; Physical channels and modulation," 2019. [Online]. Available: `https://www.etsi.org/deliver/etsi_ts/138200_138299/138211/15.05.00_60/ts_138211v150500p.pdf`.

[111] W. Rudin, *Principles of Mathematical Analysis (International Series in Pure & Applied Mathematics)*. McGraw-Hill Publishing Co., 1976.

[112] G. Fayolle, V. A. Malyshev, M. V. Menshikov, *et al.*, *Topics in the constructive theory of countable Markov chains*. Cambridge University Press, 1995.

[113] S. Shakkottai, R. Srikant, and A. L. Stolyar, "Pathwise optimality of the exponential scheduling rule for wireless channels," *Advances in Applied Probability*, vol. 36, no. 4, pp. 1021–1045, 2004.

[114] V. Krishna, *Auction Theory*. Elsevier, 2 ed., 2010.

[115] P. Milgrom, *Putting Auction Theory to Work*. Cambridge University Press, 1 ed., 2004.

[116] 3GPP TS 36.331 v.15.5.1 Rel. 15, "LTE; E-UTRA; Radio Resource Control (RRC)," 2019. [Online]. Available: `https://www.etsi.org/deliver/etsi_ts/136300_136399/136331/15.05.01_60/ts_136331v150501p.pdf`.

[117] 3GPP TS 36.300 v.15.5.0 Rel. 15, "LTE; E-UTRA and E-UTRAN; Overall description; Stage 2," 2019. [Online]. Available: `https://www.etsi.org/deliver/etsi_ts/136300_136399/136300/15.05.00_60/ts_136300v150500p.pdf`.

[118] D. S. Hochbaum, "Approximating Covering and Packing Problems: Set Cover, Vertex Cover, Independent Set, and Related Problems," in *Approximation Algorithms for NP-Hard Problems*, PWS Publishing Company, 1997.

[119] U. Feige, "A threshold of ln n for approximating set cover," *Journal of the ACM*, vol. 45, no. 4, pp. 634–652, 1998.

List of Publications

Journal Publications

J1 S. u. Zuhra, P. Chaporkar and A. Karandikar, "Towards Optimal Grouping and Resource Allocation for Multicast Streaming in LTE," in *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12239 - 12255, 2019, DOI: 10.1109/TVT.2019.2945987.

International Conference Publications

C1 S. u. Zuhra, P. Chaporkar and A. Karandikar, "Efficient Grouping and Resource Allocation for Multicast Transmission in LTE", in *IEEE WCNC*, pp. 1-6, 2017.

C2 S. u. Zuhra, P. Chaporkar and A. Karandikar, "Auction Based Resource Allocation and Pricing for Heterogeneous User Demands in eMBMS", in *IEEE WCNC*, pp. 1-6, 2019.

Preprints/Submitted/Under Preparation

J2 S. u. Zuhra, P. Chaporkar and A. Karandikar, "Resource Allocation for Loss Tolerant Video Streaming in eMBMS", Under review, *IEEE Transactions on Mobile Computing*.

J3 S. u. Zuhra, P. Chaporkar and A. Karandikar, "Multi-Connectivity for Multicast Video Streaming", Under preparation (Journal).

Patents

P1 A. Karandikar, P. Chaporkar, P. K. Jha, S. u. Zuhra, "Methods and Systems for Using Multi-Connectivity for Multicast Transmissions in a Communication System", Patent filed, December 2019 (India) (201921051123).