Multicast Resource Management for Next Generation Mobile Communication Systems

Luan Huang

Submitted for the Degree of Doctor of Philosophy from the University of Surrey

Centre for Communication Systems Research
School of Electronics and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, UK

June 2006

© Luan Huang 2006
Abstract

For the next generation of mobile communication systems, a heterogeneous infrastructure comprising multiple radio access technologies is envisioned, in which these technologies cooperate with and complement each other. Additionally, mobile network operators have begun to investigate methods of extending existing networks with multicast capability. It is believed that by means of multicast transmission and access network cooperation between digital broadcast and mobile networks, operators will be able to offer cost-efficient and ubiquitous access to a wide variety of applications, value-added services and multimedia content. The exploitation of such an infrastructure calls for effective resource management techniques to enable the efficient traffic distribution for multicast services in interworked network environments. For many years the traffic distributions strategy for cooperative networks and for multicast services have been investigated separately. The existing methods are either unable to efficiently utilize the shared resources between networks, or, are not scalable with regard to increasing traffic demands.

This thesis provides a thorough study of multicast service distribution in cooperative networks consisting of digital broadcast and mobile networks, where users request multicast content in different locations which are jointly covered by both networks. Furthermore, the desired content is not provided to users until sufficient requests have been aggregated. Two main problems are thoroughly investigated in such a scenario: 1) determining “when” desired multicast content is eligible for scheduling; 2) the selection of an appropriate radio access technology for the delivery of multicast content for each multicast group, which ensures efficient utilization of network resources and also supports the provision of QoS guarantees. This thesis develops a new resource management framework to jointly address these two problems. The resource management framework takes various parameters related to users, services and underlying networks as inputs, and maximizes the overall “system profit” through new algorithms. The “system profit” is a parameter which comprises system throughput and the cost of service delivery. It is observed that the new algorithms can improve system profit by up to 33% compared to conventional algorithms under comparable conditions. Furthermore, an effective resource management architecture is proposed based on the resource management algorithms framework in interworked DVB and UMTS networks. This architecture supports existing business models, whilst minimizing the necessary modifications to existing infrastructures.
Acknowledgments

I would like to thank my supervisors, Professor Rahim Tafazolli and Professor Barry Evans, as well as Dr. Kar Ann Chew, Dr. Shyamalie Thilakawardana, and Dr. Haitham Cruickshank for their constant support and kind advices.

I would also like to thank all my colleagues from the project IoN, Mobile VCE, and in particular Max Gasparroni from Vodafone and Patricia Wells from Telecom MODUS. Their technical assistance was significantly helpful to the development of the work presented here.

Most of all, I would like to thank my husband Mr. Yajian Liu and my parents for the unconditional love and unending support through time and distance.
Contents

Abstract.......................................................................................................................... ii
Acknowledgments .......................................................................................................... iii
Contents ........................................................................................................................... iv
List of Figures .................................................................................................................. viii
List of Tables ................................................................................................................... xi
Glossary of Terms............................................................................................................ xii
List of Symbols............................................................................................................... xv

1 Introduction.................................................................................................................. 1
   1.1 Motivations................................................................................................................. 1
   1.2 Structure of Thesis.................................................................................................... 3
   1.3 Major Achievements............................................................................................... 5

2 Interworked DVB and UMTS and State of Art on Resource Management Schemes..... 7
   2.1 Mobile and Broadcasting Interworking................................................................. 7
      2.1.1 Composite Radio Environment........................................................................ 7
      2.1.2 Mobile Telecommunications Systems............................................................ 9
      2.1.3 Broadcasting Systems..................................................................................... 12
      2.1.4 Advantages of Mobile and Broadcasting Interworking..................................... 14
      2.1.5 Related Projects.............................................................................................. 16
   2.2 Resource Management........................................................................................... 19
      2.2.1 Resource Management in UMTS................................................................. 19
      2.2.2 Resource Management in Multicast-Capable Wireless Network................. 21
      2.2.3 Resource Management in Multi-Radio Environment.................................... 23
      2.2.4 Multicast-on-Demand Scheduling............................................................... 25

3 Multicast Resource Management ................................................................................. 29
   3.1 Service Model.......................................................................................................... 29
   3.2 Problem Identification and the Scope of Research................................................... 30
      3.2.1 Batching for Multimedia-On-Demand Scheduling........................................ 30
      3.2.2 Multicast User Assignment in Heterogeneous Wireless Networks................ 31
C.2 Testbed Setup .................................................................................................................. 167
C.3 Proof of Concept ............................................................................................................. 169
List of Figures

Figure 2-1 Composite radio environment .......................................................... 8
Figure 2-2 Reference architecture to support the MBMS bearer service ............... 11
Figure 2-3 DVB-H transmit and receive ......................................................... 13
Figure 2-4 Typical architecture of IPDC over DVB-H ..................................... 13
Figure 2-5 IoN baseline architecture .............................................................. 18
Figure 2-6 Typical locations of resource management algorithms in a UMTS network ........................................................................................................... 20
Figure 2-7 A typical multimedia multicasting system ....................................... 26
Figure 3-1 Distributed resource management .................................................. 33
Figure 4-1 High-level description of the MUA-HWN problem ......................... 42
Figure 4-2 Optimisation model for MUA-HWN problem ............................... 46
Figure 4-3 An example of MUA-HWN problem ............................................ 47
Figure 4-4 Flowchart of algorithm for MUA-HWN problem ......................... 51
Figure 4-5 Heuristic of user assignment ....................................................... 54
Figure 4-6 System profit vs. number of contents with 100% multimode users ...... 57
Figure 4-7 User loss probability vs. number of contents with 100% multimode users .......................................................... 57
Figure 4-8 System profit improvement of the MP algorithm over the alternative algorithms ........................................................................................................ 59
Figure 4-9 System profit under three TH cases and two ND cases when the number of contents = 20 ........................................................................ 60
Figure 4-10 User loss probability under three TH cases and two ND cases when the number of contents = 20 .......................................................... 61
Figure 4-11 System profit vs. number of requests per content per cell with 100% multimode users under two ND cases when the number of contents = 20 ........................................................................................................... 62
Figure 4-12 User loss probability vs. number of requests per content per cell with 100% multimode users under two ND cases when the number of contents = 20 ........................................................................................................... 62
Figure 4-13 Overall system profit vs. the ratio $c_f/c_c$ with 100% multimode users in the case of sufficient and insufficient available resources .......... 64
Figure 4-14 Overall loss probability vs. the ratio $c_f/c_c$ with 100% multimode users in the case of sufficient and insufficient available resources .......... 64
Figure 4-15 Percentage of served users in cellular and broadcast networks vs. the ratio $c_f/c_c$ with 100% multimode users in the case of sufficient available resources .......... 65
Figure 4-16 Percentage of served users in cellular and broadcast networks vs. the ratio $c_f/c_c$ with 100% multimode users in the case of insufficient available resources .......... 65
Figure 5-1 System model of MDS ................................................................. 70
Figure 5-2 Multicast-on-demand service scheduling flow chart in a loss system .................................. 71
Figure 5-3 Flow chart of MDS operation in a delay system ................................................................. 73
Figure 5-4 Analytical and simulation results ......................................................................................... 79
Figure 5-5 Impact of available multicast channels $C$ in a cell, on system performance .................. 80
Figure 5-6 Impact of user request arrival rate $\lambda$, on system performance ................................ 82
Figure 5-7 Performance of optimal timeout-based scheme and traditional schemes as the available
multicast channels in a cell vary ............................................................................................................. 86
Figure 5-8 Performance of optimal timeout-based scheme and traditional schemes as the user
request arrival rate varies ...................................................................................................................... 88
Figure 6-1 Flowchart of the MRM operation ......................................................................................... 92
Figure 6-2 Top-level algorithm functional blocks for resource management ................................. 94
Figure 6-3 NP-driven business model ................................................................................................. 98
Figure 6-4 SP-driven business model ................................................................................................. 99
Figure 6-5 High-level representation of the cooperative mobile and broadcasting networks ..... 100
Figure 6-6 Phases of multicast-on-demand service provision in IoN ............................................... 102
Figure 6-7 Signalling for service scheduling in IoN ........................................................................... 105
Figure 6-8 Signalling for multicast user assignment in IoN ............................................................... 106
Figure 6-9 Overall system profit of MRM algorithms versus user waiting time threshold in a
single 3G/MBMS network .................................................................................................................... 109
Figure 6-10 Overall user loss probability of MRM algorithms versus user waiting time threshold
in a single 3G/MBMS network ............................................................................................................ 110
Figure 6-11 Overall average delay of MRM algorithms versus user waiting time threshold in a
single 3G/MBMS network .................................................................................................................... 110
Figure 6-12 Overall system profit of MRM algorithms versus user waiting time threshold in a
single DVB network ............................................................................................................................. 112
Figure 6-13 Overall user loss probability of MRM algorithms versus user waiting time threshold
in a single DVB network ....................................................................................................................... 112
Figure 6-14 Overall average delay of MRM algorithms versus user waiting time threshold in a
single DVB network ............................................................................................................................. 113
Figure 6-15 Overall system profit of MRM algorithms versus user waiting time threshold in
interworked 3G/MBMS and DVB networks ......................................................................................... 115
Figure 6-16 Overall user loss probability of MRM algorithms versus user waiting time threshold
in interworked 3G/MBMS and DVB networks .................................................................................... 115
Figure 6-17 Overall average delay of MRM algorithms versus user waiting time threshold in
interworked 3G/MBMS and DVB networks ......................................................................................... 116
Figure 6-18 Overall system profit improvement for MRM algorithms over the Delay aware MBMS by default MRM algorithm .......................................................... 117
Figure 6-19 Overall user loss probability improvement for MRM algorithms over the Delay aware MBMS by default MRM algorithm .......................................................... 118
Figure 6-20 Overall average delay improvement for MRM algorithms over the Delay aware MBMS by default MRM algorithm .......................................................... 118
Figure 6-21 The overall system profit versus the resource cost ratio for MRM schemes .......... 121
Figure 6-22 Percentage of served users in 3G/MBMS versus the resource cost ratio for MRM schemes ........................................................................................................... 122
Figure 6-23 Percentage of served users in DVB versus the resource cost ratio for MRM schemes ........................................................................................................... 123
Figure 6-24 Overall average delay versus resource cost ratio for MRM schemes ................. 125
Figure 6-25 Overall system profit vs. available resources in DVB ........................................ 128
Figure 6-26 Overall user loss probability vs. available resources in DVB .............................. 129
Figure 6-27 Percentage of served users in 3G/MBMS vs. available resources in DVB ........ 130
Figure 6-28 Percentage of served users in DVB vs. available resources in DVB ................ 131
Figure 6-29 Overall system profit vs. resource cost ratio ..................................................... 132
Figure 6-30 Overall user loss probability vs. resource cost ratio ......................................... 133
Figure 6-31 Percentage of served users in 3G/MBMS vs. resource cost ratio ..................... 133
Figure 6-32 Percentage of served users in DVB vs. resource cost ratio ............................. 134
Figure A-1 Three modelling domains of OPNET ............................................................... 153
Figure A-2 Simulation block diagram ................................................................................ 155
Figure A-3 the concept of basic cell unit ............................................................................ 157
Figure A-4 Prototype implementation of RM ...................................................................... 165
Figure A-5 Network structure of wireless network testbed ............................................... 167
Figure A-6 Network architecture for RM and GM implementation demonstration ............ 168
Figure A-7 GUI for multicast user service configuration ................................................... 169
Figure A-8 GUI for resource management configuration .................................................. 170
Figure A-9 Resource status before resource management ................................................ 170
Figure A-10 test result for the size-based batching in the resource management ............... 171
Figure A-11 test results for UMTS terminals after network selection in the resource management ........................................................................................................... 172
Figure A-12 test results for multimode terminals after network selection in the resource management ........................................................................................................... 173
Figure A-13 test result for the remaining resources after the network selection ................ 174
List of Tables

Table 1 Default inputs of analytical and simulation models ................................................................. 77
Table 2 Mapping table in optimal timeout-based batching scheme ....................................................... 83
Table 3 Performance comparison of the Delay aware + MP algorithm in a single 3G/MBMS, a single DVB and interworked 3G/MBMS and DVB ........................................................................... 119
Table 4 Network deployment cases ...................................................................................................... 158
Table 5 Terminal heterogeneity cases .................................................................................................. 158
Table 6 Simulation parameters ............................................................................................................. 163
Table 7 Default simulation parameters ................................................................................................. 163
# Glossary of Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>3GPP</td>
<td>3rd Generation Partnership Project</td>
</tr>
<tr>
<td>AAA</td>
<td>Authentication Authorization Accounting</td>
</tr>
<tr>
<td>ABC</td>
<td>Always Best Connected</td>
</tr>
<tr>
<td>AS</td>
<td>Access Selection</td>
</tr>
<tr>
<td>B3G</td>
<td>Beyond 3G</td>
</tr>
<tr>
<td>BCU</td>
<td>Basic Cell Unit</td>
</tr>
<tr>
<td>BML</td>
<td>Max_Batch with Minimum Loss</td>
</tr>
<tr>
<td>BMQ</td>
<td>Max_Batch Maximum Queue Length</td>
</tr>
<tr>
<td>BM-SC</td>
<td>Broadcast Multicast Service Centre</td>
</tr>
<tr>
<td>BSC</td>
<td>Base Station Controller</td>
</tr>
<tr>
<td>CDMA</td>
<td>Code Division Multiple Access</td>
</tr>
<tr>
<td>CISMUNDUS</td>
<td>Convergence of IP-based Services for Mobile Users and Networks in DVB-T and UMTS Systems</td>
</tr>
<tr>
<td>DAB</td>
<td>Delay Aware Broadcasting</td>
</tr>
<tr>
<td>D-AMPS</td>
<td>Digital Advanced Mobile Services</td>
</tr>
<tr>
<td>DCS</td>
<td>Digital Cellular System</td>
</tr>
<tr>
<td>DMP</td>
<td>Delay-aware Maximum Profit</td>
</tr>
<tr>
<td>DPS</td>
<td>Device Presence System</td>
</tr>
<tr>
<td>DRiVE</td>
<td>Dynamic Radio for IP Services in Vehicular Environments</td>
</tr>
<tr>
<td>DVB</td>
<td>Digital Video Broadcasting</td>
</tr>
<tr>
<td>DVB-H</td>
<td>Digital Video Broadcasting transmission system for Handheld terminal</td>
</tr>
<tr>
<td>FCFS</td>
<td>First Come First Serve</td>
</tr>
<tr>
<td>FDD</td>
<td>Frequency Division Duplex</td>
</tr>
<tr>
<td>FDMA</td>
<td>Frequency Division Multiple Access</td>
</tr>
<tr>
<td>GAP</td>
<td>Generalized Assignment Problem</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>GGSN</td>
<td>Gateway GPRS Support Node</td>
</tr>
<tr>
<td>GM</td>
<td>Group Manager</td>
</tr>
<tr>
<td>GMoD</td>
<td>Group-based Multimedia-on-Demand</td>
</tr>
<tr>
<td>GPRS</td>
<td>General Packet Radio Service</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile Communications</td>
</tr>
<tr>
<td>GW</td>
<td>Gateway</td>
</tr>
<tr>
<td>IGW</td>
<td>Interworking of Networks Gateway</td>
</tr>
<tr>
<td>IoN</td>
<td>Interworking of Networks</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>IPDC</td>
<td>Internet Protocol DataCasting</td>
</tr>
<tr>
<td>IS-54</td>
<td>Interim Standard-54</td>
</tr>
<tr>
<td>LM</td>
<td>Local Monitor</td>
</tr>
<tr>
<td>MBMS</td>
<td>Multicast Broadcast Multimedia Service</td>
</tr>
<tr>
<td>MCCH</td>
<td>MBMS Control Channel</td>
</tr>
<tr>
<td>MDKP</td>
<td>Multi-Dimension Knapsack Problem</td>
</tr>
<tr>
<td>MDS</td>
<td>Multicast-on-Demand Scheduling</td>
</tr>
<tr>
<td>MEMO</td>
<td>Multimedia Environment for Mobile</td>
</tr>
<tr>
<td>MFQL</td>
<td>Maximum Factor Queue Length</td>
</tr>
<tr>
<td>MoD</td>
<td>Multimedia-on-Demand</td>
</tr>
<tr>
<td>MP</td>
<td>Maximum Profit</td>
</tr>
<tr>
<td>MQL</td>
<td>Maximum Queue Length</td>
</tr>
<tr>
<td>MRM</td>
<td>Multicast Resource Management</td>
</tr>
<tr>
<td>MRRM</td>
<td>Multi-Radio Resource Management</td>
</tr>
<tr>
<td>MSCH</td>
<td>MBMS Point-to-Multipoint Scheduling Channel</td>
</tr>
<tr>
<td>MTCH</td>
<td>MBMS Traffic Channel</td>
</tr>
<tr>
<td>MUA</td>
<td>Multicast User Assignment</td>
</tr>
<tr>
<td>MUA-HWN</td>
<td>Multicast User Assignment in Heterogeneous Wireless Networks</td>
</tr>
<tr>
<td>Term</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>MVCE</td>
<td>Virtual Centre of Excellence in Mobile &amp; Personal Communications</td>
</tr>
<tr>
<td>ND</td>
<td>Network Deployment</td>
</tr>
<tr>
<td>NQS</td>
<td>Network and QoS Selection</td>
</tr>
<tr>
<td>OverDRiVE</td>
<td>Spectrum Efficient Uni- and Multicast Over Dynamic Radio Networks in Vehicular Environment</td>
</tr>
<tr>
<td>PDS</td>
<td>Personal Digital Cellular System</td>
</tr>
<tr>
<td>PPV</td>
<td>Pay Per View</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RCC</td>
<td>Resource Cost Calculation</td>
</tr>
<tr>
<td>RM</td>
<td>Resource Manager</td>
</tr>
<tr>
<td>RMT</td>
<td>Resource Management Triggering</td>
</tr>
<tr>
<td>RNC</td>
<td>Radio Network Controller</td>
</tr>
<tr>
<td>SGSN</td>
<td>Serving GPRS Support Node</td>
</tr>
<tr>
<td>SS</td>
<td>Service Scheduling</td>
</tr>
<tr>
<td>TDMA</td>
<td>Time Division Multiple Access</td>
</tr>
<tr>
<td>TH</td>
<td>Terminal Heterogeneity</td>
</tr>
<tr>
<td>TMP</td>
<td>Timeout-based Maximum Profit</td>
</tr>
<tr>
<td>UA</td>
<td>User Assignment</td>
</tr>
<tr>
<td>UC</td>
<td>User Centric</td>
</tr>
<tr>
<td>UMTS</td>
<td>Universal Mobile Telecommunication System</td>
</tr>
</tbody>
</table>
List of Symbols

- $\lambda$: Request arrival rate
- $\tau$: The mean waiting time of a user in a batch
- $d$: The average duration of the contents
- $C$: The number of available multicast channels
- $W$: The batching duration in time
- $N$: The batching duration in group size
- $M$: The number of multicast contents
- $R(u)$: User $u$'s reneging function
- $P$: Satisfaction ratio
- $P_{blk}$: Blocking probability
- $P_{ren}$: Reneging probability
- $\alpha_i$: The probability of a request for the content $i$
- $\overline{N}_i$: The average batch size in the number of requests of the content $i$
- $\overline{T}_i$: The average period between two consecutive copy of content $i$ delivery time
- $v_{ij}$: The value of item $j$ if assigned to the knapsack $i$
- $a_{ij}$: The weight of item $j$ if assigned to the knapsack $i$
- $b_i$: The capacity of the knapsack $i$
- $S_c$: The set of transmission data rates of the content $c$
- $V_c$: The set of prices of the content $c$
- $H$: The set of multicast-capable cells
- $m_h$: The maximum available bandwidth capacity in the cell $h$
- $n_h$: The type of the network to which the cell $h$ belongs
- $rc_h$: the unit resource cost of the cell $h$
- $U$: The set of users
- $O_u$: the candidate cell set of the user $u$
$g_u$ the content ID subscribed by the user $u$

$k_u$ the capacity of user $u$'s terminal

$\mu_u(d_u)$ the utility function of a user $u$ served with the actual received bit rate $d_u$

$R_u$ A set of transmission rates selectable by the user $u$

$p_{u,h,r}$ The user $u$'s profit given the serving cell $h$ from $O_u$, and the transmission rate $r$ from $R_u$

$w_{u,h,r}$ The user $u$'s resource consumption given the serving cell $h$ from $O_u$, and the transmission rate $r$ from $R_u$
Chapter 1

1 Introduction

1.1 Motivations

The next generation mobile systems, such as the beyond 3G (B3G), are envisioned to be all-IP multi-access networks together with multimode terminals and smart access networks mechanisms [1][2] [3][4] [5]. The composite radio concept envisions heterogeneous infrastructures comprising diverse wireless systems e.g., 2G, 3G, DVB, WLAN, and hence enabling various transmission modes, such as, unicast, multicast and broadcast, in a cooperative and complementary manner. Multimode terminals will be required to enable such cooperation of networks. It is believed that through this cooperation, operators will be able to offer efficient wireless access to a wide variety of IP-based services, in terms of the cost and the quality of service (QoS) while still retaining control of their respective networks. The user-centred concept is widely accepted in a B3G environment. In this concept, the parameters and configuration options regarding the service delivery issue, e.g., QoS levels and preferred networks, are provisioned by the network in accordance to users' preferences.

In recent years, there have been increasing demands from users for bandwidth-intensive multimedia services such as video-on-demand and multimedia games [6]. These services are expected to bring big revenues to both network operators and service providers. However, the scarce radio resource in the mobile network is the major bottleneck preventing cost-efficient provisioning of these bandwidth-intensive services. To overcome this problem, the 3rd Generation Partnership Project (3GPP) is currently standardizing the Multicast Broadcast Multimedia Service (MBMS) feature to incorporate multicast technology in a 3G cellular network [7][8]. Multicast transmission has been regarded as a cost-effective mechanism for providing value-added multimedia services in bandwidth-limited systems. In addition to MBMS, the DVB project is standardising the Digital Video Broadcasting transmission system for Handheld terminals (DVB-H) [9], and Internet Protocol DataCasting (IPDC) [10] to enable efficient datagram delivery to mobile terminals. This gives another choice to use multicast technology for cost-effective multimedia service delivery to mobile users.

In this context, a heterogeneous infrastructure comprising multiple multicast-capable wireless networks could be one of the most promising structures in future mobile systems. Therefore, the
exploitation of such infrastructure where heterogeneous wireless networks and multicast-based transmission coexist is of primary importance for future mobile communication systems.

Efficient resource management has been widely studied in cellular systems. In addition, Multi-Radio Resource Management (MRRM) has recently attracted growing attention [11]. The MRRM mechanisms, coordinating several radio accesses, fulfil a key role for providing wireless services with improved resource efficiency, coverage, and service quality. However, the general concept of the MRRM is still in the early stage of development. More comprehensive investigations are required to develop an efficient resource management solution, especially for supporting multicast services in composite wireless networks. Therefore, this thesis presents and investigates a new resource management solution for distributing multicast services over cooperative wireless networks.

Generally, three research issues arise during the implementation of resource management techniques in the context of the provision of multicast multimedia services in a composite radio environment (e.g., 3G/MBMS and DVB-H).

- Firstly, to transmit a multicast service, it is essential to aggregate enough users before transmitting the service to the users. The more users are aggregated; the more efficiently the resource is used. However, this also implies a longer delay experienced by an early user and hence a worse QoS. Therefore, it is very important in practice to strike a balance between the efficiency of resource utilization and the quality of service in terms of the user delay;

- Secondly, in a composite network environment, users are reachable via a wide variety of access networks. Each access network offers different QoS desired by users and consumes different radio resources. In addition to the heterogeneity of networks, users differ from one another in terms of the preference in services, location, and terminal capability etc. Considering the heterogeneity of networks and users in a composite radio environment, it is a challenging task to determine the most appropriate network and QoS to provide desired services to users. This problem becomes tougher in multicast scenarios where a potentially large group of heterogeneous users have to be served simultaneously;

- Thirdly, because the resource management needs to take care of multiple networks and operators, each of which may be an independent administrative actor which controls the resources in its own network using different strategies, it is very important to identify the coordination degree, relationships and interfaces among underlying networks and operators when performing a joint resource management over them. Generally, two potential methods are available, namely distributed resource management and centralized resource management. Centralized control is beneficial to the infrastructure comprising a
number of networks belonging to a same administrative entity for efficient operation, while the distributed control is beneficial to large networks for scalability reasons, or when a central coordination is simply not desirable due to for instance, the fact that the involved administrative entities do not trust each other. The most suitable resource management architecture (i.e., distributed resource management or centralized resource management) needs to be investigated in this research. Furthermore, in the case of the distributed control, which will be affected by the type and amount of information that can be exchanged, the time scale at which information exchange is feasible, therefore, the possible degree of coordination needs to be investigated.

These issues show that efficient algorithms are crucial for effective resource management in the context of providing multicast services in a composite radio environment. Efficient multicast resource management (MRM) algorithms are developed in this thesis to deal with this problem. These algorithms are categorized into two types, namely multicast-on-demand scheduling (MDS) that addresses the first research issue, and multicast user assignment (MUA) that addresses the second research issue. The objective is to optimise the resource efficiency, coverage, and service quality, by serving multicast contents at the right time to group of users simultaneously through the best combination of access networks and with the best combination of transmission data rates.

In addition, a distributed resource management architecture is proposed to implement the algorithms, to cater for the interworking of networks (IoN) that is considered as a feasible solution to implement the composite radio concept in the near future. An IoN example is the interworking of UMTS and DVB networks. The principle reason why these networks are chosen is due to their multicasting support capability. Nevertheless, the research here is generic and aims to be applicable to any architecture comprising multicast-enabled wireless technologies (e.g., 3G/MBMS, WLAN, DVB-T/H/S, or future new technologies), by exploring the MRM problem from a high layer independent of the specific underlying access network.

1.2 Structure of Thesis

The thesis is organized into seven chapters.

Chapter 2 discusses the issues of radio network interworking/cooperation in more detail. This reviews the concept of "composite radio" environment that is envisaged to be the architecture of future mobile communications, e.g., B3G and 4G. As an example of composite radio, the cooperative mobile and broadcasting networks attract more attention than other structures of composite radios due to their efficiency of providing bandwidth-intensive services to wide audiences. Hence, the cooperative mobile and broadcasting networks will be discussed in depth in this chapter. The advantages of combining mobile and broadcasting networks are outlined, in
Chapter 1. Introduction

terms of how they can act as complementary systems. In addition, the issues of multicasting in wireless networks (e.g., cellular and DVB) will be discussed due to its distribution efficiency. Following this, some resource management issues are outlined. This includes traditional resource management in cellular networks, multicast resource management in wireless multicast networks, and multi-radio resource management for composite networks. In particular, with respect to the multicast resource management, the aggregation methods that are essential for multicasting will be given a more detailed discussion.

Chapter 3 specifies three research issues in multicast resource management in terms of the algorithm design and architecture definition. Firstly, the service model considered in the research will be introduced. It basically is a group-based multimedia-on-demand service (GMoD). Following this, two research issues will be specified, namely multicast-on-demand scheduling and multicast user assignment. The motivations behind these research issues will be presented and the challenges will be discussed. In addition, a distributed resource management that has been proposed in the composite radio environment will be explained. This resource management architecture is constructed based on an IoN infrastructure, consisting of multiple access networks.

Chapter 4 explains the multicast user assignment algorithms that have been developed in the context of rate-adaptive multimedia applications, heterogeneous access networks, and multicast transmission. By the aggregation of users’ requests, there can be many user requests staying in the system. Each request stays in a particular batching queue according to the desired content. The multicast user assignment aims to carry out two-dimensional resource allocation for these users, namely, a serving cell and a transmission bandwidth of the desired multicast content. This chapter will firstly introduce the background knowledge and justify the assumptions adopted in the multicast user assignment problem. The detailed modelling of the multicast user assignment in heterogeneous wireless networks problem is then described in a high level and formal manner. The solution to this problem aims to maximize the total profit in the system. A general profit formulation is considered, which can accommodate different measures such as throughput, user satisfaction, and fairness, each of which corresponds to a specific optimisation objective. Several different alternative algorithms maximizing the total profit are outlined. Following this, the simulator that has been constructed to evaluate the performance of the algorithms is described. Finally, in this chapter, we study the performance of the proposed solutions under a wide variety of traffic loads, wide network and user heterogeneity, and comparing them to traditional solutions.

Chapter 5 discusses, evaluates and analyses the algorithms that have been developed for the multicast-on-demand scheduling in wireless networks. Some basic concepts will be presented first including the system model. The basic idea of the multicast-on-demand scheduling algorithms is to batch users’ requests for the same contents until enough users are aggregated, and then serve them in one transmission. As a result, the resources can be used more efficiently. The proposed
algorithms will be explained, discussed, evaluated and analysed considering two system models, namely blocked-calls cleared system and blocked-calls delayed system. A general flowchart of the algorithm's operation is presented for each system model respectively. Following this, an analytical model that is derived for the performance analysis of the scheduling algorithm in blocked-calls cleared systems are described. This analytical model constructs a basis for a novel scheduling algorithm that will be explained and evaluated by comparing it to conventional algorithms using system level simulation in the remainder of this chapter.

Chapter 6 introduces a new resource management framework for implementing the overall multicast resource management algorithms. It specifies the method of combining the multicast-on-demand scheduling algorithms and the multicast user assignment algorithms in a multiple networks scenario. Two functional blocks are designed to perform the respective resource management algorithms, namely, the Service Scheduling function and the Network & QoS Selection function. The first part of this chapter introduces the detailed operation of the overall multicast resource management algorithms, and the functional block architecture that is developed by taking the IoN infrastructure of cellular and broadcasting networks as a case study. Following this, a procedure example for the GMoD service provisioning is illustrated, as well as the required signalling flows. In the second part of this chapter, the proposed multicast resource management framework is evaluated by system level simulation. Two aspects of investigations are carried out: a) the performance evaluation of the multicast-on-demand scheduling algorithms developed for block-calls delayed systems; b) effect on the performance of the multicast resource management framework when varying the user’s delay tolerance and the resource cost. Following this, this chapter explores the possible degree of cooperation among multiple networks/operators, which results from the different levels of the business relationship between operators when performing the distributed resource management. In particular, it looks at how the different degrees of cooperation affect the overall performance. The cooperation degree is mainly specified by the type and amount of information that can be exchanged between networks. Three degrees of cooperation are presented and discussed, namely non-cooperative, half-cooperative and fully-cooperative. Finally, these three cooperation degrees are evaluated by simulation.

Finally, in Chapter 7, conclusions are drawn on the overall work described in this thesis, and suggestions are made for future study.

1.3 Major Achievements

Several new areas of work were undertaken and are described in this thesis. Resource management for the provision of multicast services in heterogeneous wireless networks is a new
field, and little technical work has been undertaken or published in this area. Specifically, the following aspects of our investigation are novel:

- **Multicast user assignment algorithms for multicast resource management**: This work extends the current user assignment problem from the unicast environment to the multicast environment and formulates the user assignment problem for multicast services as a combinatorial optimisation problem. Moreover, this problem is proven to be NP-hard. Novel algorithms are developed and proposed to solve this problem. The respective publications can be found in [108][111][113];

- **Multicast-on-demand scheduling algorithms for multicast resource management**: This work extends the current wired network scheduling algorithm for providing multicast services in wireless networks. A theoretical model is derived for analysis of the algorithm's performance in a blocked-call-cleared system. Moreover, the service scheduling algorithms for multicast transmission in a blocked-call-delayed system are studied by simulation. Based on both simulation and analytical results, novel algorithms are developed especially for resource-limited and heterogeneous wireless networks [107][112][114][115];

- **A novel multicast resource management framework**: a new multicast resource management scheme comprising both the multicast-on-demand scheduling algorithm and the multicast user assignment algorithm has been developed. The structure and interface between these two components are studied [112][114];

- **Service provisioning procedure**: A new cost-effective multicast services provisioning procedure and architecture for cooperative wireless networks has been proposed and evaluated [18][109][110];

- **Impact of the degree of cooperation between the underlying networks which are performing the multicast resource management**: the multicast resource management framework is evaluated with different degrees of cooperation between the constituent networks.
Chapter 2

2 Interworked DVB and UMTS and State of Art on Resource Management Schemes

This chapter discusses the issue of composite radio environment in more detail, particularly focusing on the potential for interworked mobile and broadcasting networks. It initially reviews the concept of the composite radio environment. Following that, the development and specific characteristics of mobile and broadcasting networks are described respectively. Then, the advantages of mobile and broadcast interworking are outlined, and particularly how the respective system can be complementary to each other. Finally, the resource management issues are discussed, which includes the traditional resource management in cellular networks, multicast resource management in wireless multicast-enabled networks, and multi-radio resource management for composite networks.

2.1 Mobile and Broadcasting Interworking

2.1.1 Composite Radio Environment

The composite radio environment is a new research area, and is envisioned as a possible architecture for future generation wireless communication systems [1]-[5]. The motivation for composite radio environment is twofold. On one hand, for a network operator, it does not need to develop new radio access techniques suitable for transmission of all traffic types, but rather to use the current systems already developed and highly specialized for the type of data they were designed to deliver. These differing radio systems are operating in a coordinated manner, and allow the most appropriate access system to be used depending on particular circumstances. On the other hand, for users and service providers, the underlying radio technology can be irrelevant, as long as the requirements of cost and quality criteria are satisfied. The users are able to enjoy seamless connectivity and ubiquitous access to a wide variety of applications over the most efficient combination of available systems.

The classic representation of a composite radio environment can be seen in Figure 2-1. In such an environment, the wireless system comprises a number of different access technologies and
different administrative domain in which the coverage of one network overlays the coverage of another. In addition, different networks have different capacity and causes different resource costs by serving users due to employed internal mechanisms and access technologies. Such an environment is also referred to as heterogeneity of networks. On the other hand, mobile terminals, such as mobile phones, PDAs, handheld gaming devices and notebook computers, are being provided with the ability to connect to a number of different radio access networks. Multimode terminals (e.g., through the software/reconfigurable radio concepts [76][77]) can enable the composite radio concept. These different terminals differ from one another in terminal capability, such as the network air interface the terminal is provided with and the maximum transmission capacity the terminal can be served with. In addition, the user location and preference are important attributes that distinguish users from one another. Such an environment is referred to as heterogeneity of users.

![Figure 2-1 Composite radio environment](image)

The composite radio system will be characterized by a horizontal communication model, where different access technologies such as cellular, cordless, wireless LAN type systems, short-range connectivity and wired systems will be combined on a common platform to complement each other in an optimum way for different service requirements and radio environments. These different (wireless and/or wireline) networks will interface to core and/or backbone network elements over the IP protocol. Thus, the prime objective for composite ratio system is to incorporate and integrate different wireless access technologies and mobile network architectures in a complementary manner so as to achieve a seamless wireless access infrastructure.

In the composite radio environment, combining the strength of mobile and broadcasting networks is expected to provide users with cost-effective interactive mobile IP streaming services, thereby opening the door for a lot of interesting mobile applications, such as push delivery of content,
video and audio on demand service, multimedia conference and multiplayer games [6][10]. In the past, the mobile networks e.g., (GSM, GPRS, UMTS etc.) and the broadcasting networks (e.g., DVB-T, DVB-H, DVB-S, etc.) as well as related services evolved independently. Firstly, the development procedure and specific characteristics of mobile and broadcasting networks are presented respectively as follows, before the advantages of interworking mobile and broadcasting networks are outlined.

2.1.2 Mobile Telecommunications Systems

Digital cellular networks are the segment of the market for mobile and wireless communications that are growing most rapidly. The present day second-generation (‘2G’) systems have been designed to fit into the traditional telephony architecture. The basic versions typically implement a circuit-switched service, focused on voice traffic, and only offer data rates up to 14.4 kbit/s. Higher data rates (up to 115.2 kbit/s) can be achieved by employing ‘2.5G’ systems, which can be engineered on top of the 2G voice services. However, full high capacity data communications (up to 2Mbit/s) is achieved only with the forthcoming 3G system. Mobile telecommunications systems can be conveniently categorised by the method of access to the allocated radio resources. The media access can use following multiple access technologies: Time division multiple access (TDMA); Code division multiple access (CDMA) and Frequency division multiple access (FDMA).

Example of the second generation (2G) system includes Global System for Mobile Communications (GSM), IS-54 (North American TDMA Digital Cellular), Digital Cellular System 1800 (DCS 1800), and Personal Communications Systems (PCS) 1900 [80]. In Europe most people use GSM, which is a type of TDMA system using frequency division duplex (FDD) for the separation of uplink and downlink channels. From 2.5G development and beyond, it is obvious that mobile services would increase drastically as it is no longer just serving the basic voice calls, but is also expected to deliver multiple high data rate services that are digitally packet-switched and Internet Protocol (IP) based. Example of 2.5G systems, such as General Packet Radio Service (GPRS) is an enhancement for GSM’s data handling capabilities [81]. GPRS provides packet mode transfer of data for applications such as Web requests, which exhibit bursty traffic patterns. Users of GPRS specify a QoS, which determines the service precedence, reliability, delay and data throughput. GPRS should adaptively allocate radio resources to fulfil the user’s requirements. The Universal Mobile Telecommunication System (UMTS) is the 3G cellular mobile system designed to offer also multimedia/Internet access to portable mobile terminals, based on Wideband Code-Division Multiple Access (WCDMA) radio access platform [82]. With the target of carrying high-capacity bi-directional multimedia services via radio, UMTS networks are typically characterised by small cells, especially in densely populated areas.
Chapter 2. Interworked DVB and UMTS and State of Art on Resource Management Schemes

The main advantage of UMTS over the previous generation of mobile systems is the capability to deliver higher bit rate multimedia services, such as Internet pages and video clips, to portable phones and other types of mobile terminal. Therefore UMTS networks and terminals are able to deliver audio services and low-resolution video services to mobile terminals.

In summary, the essential common feature of 2G, 2.5G and 3G systems is that they are designed to deal mainly with "one-to-one" connections (unicast). Typical applications are telephony, bidirectional exchange of data, and on-demand access to multimedia content. In the unicast method, data are sent one by one to individual recipients. Hence, a traffic increase implies additional costs of a network and consumes more network resources. In recent years, there has been great interest in utilizing multicast technology in wireless networks due to its capability of offering cost-effective provisioning of high quality multimedia contents [78][79]. This is attractive especially for bandwidth-limited systems, such as, mobile systems where maximizing the efficiency of radio resource utilization is always the major research challenge. Unlike unicast, in multicast one copy of data is transmitted from source to multiple recipients in a multicast group. Apparently, this leads to less resource consumption in comparison to unicast. In addition, unlike broadcast that distributes contents in a large area regardless of end-users' interests and responses, such as, digital TV service, multicast distribution is focused on localized group services. Therefore, multicast is able to offer wide variety of services customized to individual users' preferences, while avoiding sending data to irrelevant users or wasting unnecessary resources to send duplicated data. From 2001, the 3rd Generation Partnership Project (3GPP) started a specification effort to enable efficient support for multicast services in UMTS, i.e., Multimedia Broadcast Multicast Service (MBMS) [7][8]. The MBMS is a unidirectional point to multipoint bearer service in which high-bit rate multimedia data is transmitted from a single source entity to multiple recipients. MBMS would utilise resources efficiently in both core and radio access networks over the UMTS networks, with the main focus on the radio interface efficiency. Specially, multiple users should be able to share common resources when receiving identical traffic. Figure 2-2 shows the architecture to support the MBMS bear service over current UMTS architecture.
Chapter 2. Interworking DVB and UMTS and State of Art on Resource Management Schemes

Figure 2-2 Reference architecture to support the MBMS bearer service

An MBMS specific functional entity – Broadcast Multicast Service Centre (BM-SC) supports various MBMS user specific services. It may serve as an entry point for content providers offering MBMS transmissions, used to authorise and initiate MBMS bearer services within the service area and can be used to schedule and deliver MBMS transmissions. The service area is the coverage in which a particular service is available by an operator [7]. A service area may represent the entire coverage area or parts of the coverage area of the access network owned by an operator. In this thesis, the service is referred to the multicast service that offers various contents.

In addition, to provide MBMS bearer services, existing functional entities, GGSN, SGSN, RNC/BSC, perform several MBMS related functions and procedures, some of which are specific to MBMS. In core network, IP multicast can be used to transmit the multicast data from GGSN/BM-SC to SGSNs and RNCs in transport IP level. Only one tunnel would exist between the GGSN and SGSN regardless of the number of users served by the SGSN. In the UTRAN, the same data would be transported over a shared channel and data should only be sent to selected cells, which has users receiving this data. Three candidates for logical channels have been considered for MBMS [8]. They are MBMS Control Channel (MCCH), MBMS Traffic Channel (MTCH) and MBMS Point-to-Multipoint Scheduling Channel (MSCH).

The 3G beyond for future mobile communication systems tend to mean different things to different people. In European Union, the widely accepted vision sketches a heterogeneous network infrastructure comprising different wireless access systems, (e.g., GSM/GPRS, UMTS, DVB-T/S, WLAN) in a complementary manner, where the user, supported by his/her personal intelligent agent(s), enjoys untethered connectivity and ubiquitous access to applications over the most efficient combination of available systems [4].
2.1.3 Broadcasting Systems

Digital Video Broadcasting (DVB) is one of the main digital technologies used for the delivery of unidirectional broadcasting services [83]. DVB is a European standard, but is becoming the dominant worldwide standard for digital TV. There are three main variants of DVB: terrestrial based DVB-T, cable based DVB-C, and satellite based DVB-S. DVB networks make use of high power transmitters covering broad services areas (e.g., regions or nations) for distribution services only. In addition, DVB networks are characterized by high data transfer rates on the downlink. For example, the achievable data-rates of DVB-T range from 3.7–23.8 Mbit/s for a 6 MHz channel and from 4.9 – 31.7 Mbit/s for an 8 MHz channel [84]. A typical DVB program contains three elementary streams: video, audio and teletext data. MPEG-2 is employed to handle these dependent packets until they are broadcasted as an independent programme together with other independent programmes, e.g. the traditional TV broadcasting [85]. In addition to traditional TV application, the DVB data broadcasting specification [86] is designed to allow software downloads, to deliver Internet services over broadcast channels (using IP tunnelling), and provide interactive TV, etc. The profile of Multiprotocol Encapsulation supports data broadcast services that require the transmission of datagrams of communication protocols, e.g. IP datagrams, via DVB compliant networks. IP multicast based service delivery from Internet can be enabled simply by connecting broadcast operator network to Internet and running some IP (multicast) routing protocols between them.

As the latest development based on DVB-T, the DVB Project has specified a transmission system to provide an efficient way of carrying multimedia services over digital terrestrial broadcasting networks to handheld terminals (DVB-H) [9]. Although the DVB-T transmission system has proven its ability to serve fixed, portable and mobile terminals, handheld terminals (defined as a light battery powered apparatus) require specific features from the transmission system serving them. The major concern with any handheld device is that of battery life. The current and projected power consumption of DVB-T is too high to support handheld receivers that expect to last from one to several days on a single charge. The other major requirements for DVB-H were an ability to receive 15 Mbit/s in an 8 MHz channel and in a wide area single frequency network at high speed. Furthermore, all this should be possible while maintaining maximum compatibility with existing DVB-T networks and systems. In order to meet the above requirements, the DVB-H makes use of the following technology elements for the link layer and the physical layer: Time-Slicing for power saving; MPE-FEC (multi protocol encapsulation-forward error correction) for additional robustness and mobility; 4K mode for mobility and network design flexibility and additional minor changes, e.g., in signalling. In contrast to other DVB transmission systems based on the DVB transport stream (adopted from the MPEG-2 standard), DVB-H is meant for IP-based
services via MPE insertion. This service can share DVB-T multiplex with MPEG-2 services. Figure 2-3 shows the procedure of DVB-H transmission and reception.

![Figure 2-3 DVB-H transmit and receive](image)

Recently, the IP Datacast (IPDC) system has been proposed in conjunction with DVB-H, enabling mobile television by DVB Project. IP Datacast over DVB-H is an end-to-end broadcast system for delivery of any types of digital content and services using IP based mechanisms optimized for devices with limitations on computational resources and battery. An inherent part of the IPDC system is that it comprises of a unidirectional DVB broadcast path that may be combined with a bi-directional mobile/cellular interactivity path. IPDC is therefore a platform that can be used for enabling the convergence of service from broadcast/media and telecommunications domain (e.g., mobile/cellular). The typical architecture of IPDC is presented as follows.

![Figure 2-4 Typical architecture of IPDC over DVB-H](image)

The major functions provided by the service management entity are: 1) registration of service applications; 2) assignment of services to location, to bandwidth and schedules services over time; 3) security/service protection provision; 4) location services provision. Further information on IPDC can be found at [87].
2.1.4 Advantages of Mobile and Broadcasting Interworking

From the above discussion, fundamental differences between mobile and broadcasting systems can be observed. One essential difference is that a broadcasting system deals with unidirectional “one-to-many” transmissions, whereas, a mobile system mainly focuses on bi-directional “one-to-one” connections. Typical applications are telephony, bi-directional exchange of data, and on-demand access to multimedia contents. Employing different transmission modes is a critical distinction between mobile and broadcasting systems, and gives rise to the main advantage of combining these types of networks. However, there are many other important differences as well.

The main characteristics of a mobile telecommunicating system are summarized below:

- Provide unicast point-to-point services. These services are also referred as 'pull' services, since the user specifically requests the service, which is then delivered on-demand;

- Services are generally bi-directional. By using radio uplink, users can easily get access to interactive services;

- Mobile networks make use of low power transmitters covering small cells. The cost per user of a mobile network is high. Moreover, the higher the interactive traffic the network must handle, the smaller the cell size must be, and hence a traffic increase implies additional cost.

- The downlink capacity is generally low. For example UMTS offers the maximum throughput of 384-2000 kbit/s for a transmission. Only maximum four active users can use a 384 kbit/s packet data service at nearly full speed in a UMTS macro cell. Although, MBMS has already made a good progress in resource efficiency by using multicast compared to UMTS, the capacity is still not enough to satisfy users’ quick increasing demands for bandwidth-intensive applications due to the lack of fast power control mechanism. For instance, 10-15% of the total base station power would be required for providing one 64kbps multicast MBMS in the macrocellular environment [66]. Thus network can be easily overloaded when diverse multicast contents are implemented;

- Mobile networks re-use the same 5MHz frequency across cells. So, the spectrum resources can be used efficiently.

- High levels of user mobility are enabled, due to the inherent support for handovers between cells. Using efficient location management mechanism, a mobile system can keep accurate track of users in the system;

- The users in a mobile system are always authenticated, and the system always knows who is accessing its services, and the levels of security are generally high;
• The system has developed support for billing the users according to their network usage.

On the other hand, broadcasting networks have different characteristics, as summarized below:

• Service delivery is point-to-multipoint, and services are generally the ‘push’ type, i.e., the network sends the services to users even though users do not specifically request them;

• A broadcast network is unidirectional. Only limited kinds of return channels are supported, for example through the fixed network, such as PSTN/ISDN. Therefore the services are inherently asymmetric;

• Broadcasting networks make use of high power transmitters covering broad service areas (e.g., regions or nations) for distribution services only. Network cost per user is low, and moreover an increase of receivers (within the covered area) does not imply an increase of network cost;

• Broadcasting networks offer high data transfer rates on the downlink. For instance, a DVB network may broadcast multiplexed data transmission streams at a rate of the order of tens of Mbps. Therefore, users can be satisfied with high transmission rate. Besides, many simultaneous streams can be supported;

• Broadcasting networks use radio carriers of 8MHz bandwidth in a cell. Broadcasting networks are spectrally inefficient in terms of the amount of radio spectrum they require to send different data across cells, as another 8MHz frequency is required to transmit different data in different cells.

• Mobility is possible in broadcast networks, but is not supported as efficiently as in mobile networks. The location management function is unavailable in current broadcast networks. Therefore, users cannot be positioned properly. As a result, the network resource could be wasted if the service is delivered to the cell where there are no interested users;

• Broadcast networks normally do not know which users are accessing its services, although some authentication can be used, particularly for subscription services;

• Users of a broadcast network are normally not billed according to the usage of the network, except in the case of pay-per-view services.

In summary, by comparing the features of mobile and broadcasting networks, it can be seen that combining the strength of mobile and broadcasting networks while complementing the weakness of each other gives significant potential for new user services and applications. The typical new service could be bandwidth-intensive multimedia-on-demand services customized according to
individual users' locations and preferences in a resource efficient way. In general, the motivations for combining mobile and broadcasting networks are summarized below:

- From the perspective of users, their increasing needs for the personalized and customized multimedia rich services are satisfied as long as the price and quality criteria are met;
- From the perspective of service providers and content providers, they are able to increase media coverage and sell more goods and services;
- From the perspective of mobile and broadcast network operators, they are enabled to achieve value-added services and offer a cost-effective and seamless path for providing personalized multimedia rich contents that can simultaneously serve large number of clients.

2.1.5 Related Projects

Much of the pioneering work on cellular and broadcasting combination has been performed in Europe, and one of the first projects to address this issue was the multimedia environment for mobile (MEMO) project [20]. It considers the possibilities of using a GSM uplink, and a digital audio broadcasting (DAB) downlink in a coordinate manner. It demonstrated the potential for service enhancement based on mobile and broadcast convergence in a real operational platform.

With the development of mobile and broadcast networks, accordingly more advanced mechanisms are required to converge these two types of networks. Consequently, following MEMO, many other projects emerge, such as CISMUNDUS [21], DriVE [22], OverDriVE [23] and Mobile VCE [17]. These projects aim to provide a network solution for converged service provision over the convergence of mobile and broadcast networks, such as GPRS, UMTS, and broadcast networks, such as DVB-T. Different from MEMO, these projects implement the IP multicast services in the convergence systems. Moreover, more service scenarios are investigated by making a full use of complementary features of mobile and broadcast networks, although the common feature is to use the uplink provided by mobile networks. Furthermore, several studies have been carried out, which look into how best to determine what service should be sent out over the broadcast medium, and what should be unicasted over a mobile network. For instance, based on the service type, some services are more suitable to be delivered via a broadcast network, such as a traditional TV service, while some services are more suitable to be delivered via a mobile network, such as a traditional voice service. For some complementary services, such as multimedia streaming services that can be supported by both networks, a more advanced scheme is required. A simple popularity-based selection criterion was proposed by both CISMUNDUS and DriVE. In this criterion, when user demand is very high for a particular service then it becomes more efficient to multicast it over the broadcast network than to unicast over the cellular.
network. Otherwise, when only a few users wish to consume a service, then it is more efficient to deliver it by unicast over a mobile network. In addition, if the selection of network occurs when there is loss of broadcast signal, then the mobile network is able to provide a complementary service and vice versa.

Most of the projects adopt GPRS or UMTS as a case study for mobile networks when cooperating with broadcast networks. Therefore, the services delivered over mobile networks are always unicast-based. With the development of MBMS specification, which enables multicast on the basis of UMTS, the resources in mobile networks can be used more efficiently. OverDRiVE and Mobile VCE adopt 3G/MBMS as an option to cooperate with DVB, jointly providing multicast services. A group management function was proposed by OverDRiVE to optimise multicast transmission to mobile receivers over the available access networks. This function is responsible for the selection of the most appropriate network for receivers either from a user point of view to please individual users, or from a network point of view to optimize the efficiency of network resources. Although the conceptual work has been contributed, it lacks the actual mechanism for algorithms to achieve the optimized delivery. In contrast, this thesis explored a practical approach to achieve these optimizations for multicast delivery from the resource management perspective. The contributions in this thesis make up one of the major inputs to Mobile VCE.

To make a full use of the combined cellular and broadcasting networks, firstly, we need to specify an effective architecture for combining these two networks. In general, the architectures proposed so far can be categorized into two types, namely, integration adopted by most of the projects, such as CISMUNDUS, DriVE and OverDRiVE, and cooperation/interworking proposed by Mobile VCE. With the integration of access networks, a third party retains full control of underlying access networks that are currently owned by different operators, and performs the service and resource management by means of a centralised platform. This architecture requires a major change to existing infrastructure, along with developing new business models and discovering a third party worthy of full trust replacing existing operators. Therefore, this approach is not feasible in the short future due to current regulatory rules and business environment. On the contrary, the interworking of access networks allows operators to retain control on their respective networks while cooperating with one another in a complementary manner. Moreover, it allows flexible enhancement when a new access system arises in the future. This architecture reduces the investment costs and hence the risks, since existing networks of all operators could be utilized for the service provisioning. The Mobile VCE Interworking of Networks (IoN) Work Area has explored a baseline IoN architecture to allow secure multiple operators interworking, providing new value-added services that each operator individually could not provide in a cost efficient, secure, or reliable manner [17]. This architecture is adopted in our research and is illustrated in Figure 2-5.
Chapter 2. Interworked DVB and UMTS and State of Art on Resource Management Schemes

The basic goal of IoN is to allow broadcast and telecom network operators to co-operate to make the heterogeneous wireless environments transparent to users. A key point is that the networks are not being integrated; instead, they remain autonomous independently-managed entities. They are owned by and under control of their respective operators. The access networks cores are interconnected via a logical interface enabling inter-working at the network and service layers. This architecture does not suggest any alteration to the existing network architecture, other than creating interfaces among the networks involved. Interworking-related signalling between networks is carried out over a new entity known as the Interworking of Networks Gateway (IGW) residing in each network. An IGW represents an administration domain and communicates with IGW of another domain to which an interworking agreement has been established. IGWs are interconnected via logical channels known as the IoN Link. This interface enables signalling and information exchange for any interworking purposes. The logical connection between two gateways is provided by an underlying transport network with certain bandwidth and other QoS guarantees. IGW is also the gateway that allows service and content providers to exploit interworking of mobile and broadcasting networks. It provides the interface for any service/content provider who wants to provide service/content over both mobile and broadcasting networks. Therefore, IGW is a signalling gateway between networks as well as a service gateway between networks and service providers. To allow seamless provision of services cost-effectively across interworked heterogeneous networks, in particular cellular and broadcasting networks, some architecture components have been developed, which reside in the IGW. These include...
resource manager (RM), group manager (GM), security, handover management and device presence system (DPS).

This thesis is focused on the development of resource management algorithms and architecture, taking the IoN architecture as a case study. However, the developed resource management algorithms are designed to be independent of the underlying architecture, and applicable to any combined multicast-enabled access networks. In the remaining of this thesis, the design and implementation of the resource management entity (i.e., RM) and the other entities supporting the resource management (e.g., GM and DPS) is discussed in depth.

2.2 Resource Management

The subject of resource management has been widely researched in both wireless and wired systems. Traditional resource management covers admission control, service/packet scheduling, congestion control, load balancing, power control etc. In the multimedia-oriented networks, the rate-adaptive feature of multimedia applications is exploited by resource management algorithms to further improve the efficiency of resource utilization. In a multicast environment, efficient resource management mechanisms are required to adapt to the point-to-multipoint feature of multicast transmission. In the heterogeneous network environment where multiple networks coexist, the resource management aims to improve the efficiency of resource utilization and service coverage, by selecting the most appropriate combination of access networks for users that can be reachable by multiple networks.

2.2.1 Resource Management in UMTS

The evaluation of the end user needs towards multimedia applications has pushed the wireless community to conceive the so-called 3G systems, such as UMTS, where WCDMA is the predominant technology. WCDMA access networks support the provision of future 3G mobile multimedia services with different QoS guarantees and the ability to optimize the spectrum efficiency in the air interface by means of efficient radio resource management algorithms. The objective of resource management is to guarantee QoS requirements, to maintain the planned coverage area and to offer high capacity, while making a full use of the radio resources of the network: available codes, bandwidth (spectrum), and transmit power [82]. Resource management needs to satisfy divergent desires of wireless network members (users, service providers), when allocating the limited amount of resource. From the users' viewpoint, they want to maximize the quality of provided services, such as getting maximum throughput, lowest block & drop rate and frame error rate. On the other hand, from the network operators' perspective, they want the maximum revenue, which means they desire the maximum system capacity to serve as many
users as possible. Meanwhile, the service providers hope to minimize the cost, e.g. deployment and recurring costs.

In UMTS, resource management can be divided into Admission Control, Load Control, Packet Scheduling, Handover, and Power Control [82]. Admission control, load control and packet scheduling aim to guarantee the quality of service and to maximise the system throughput with a mix of different bit rates, services and quality requirements. Handover is the mechanism that transfers an ongoing call from one cell to another cell as a user moves toward the coverage area of the neighbouring cell. Power control is needed to provide the required quality of service via keeping the interference levels at minimum in the air interface by controlling the power used on both downlink and uplink.

Typical locations of resource management algorithms in a WCDMA network are shown as follows.

![Figure 2-6 Typical locations of resource management algorithms in a UMTS network](image)

Typically, most of the resource management functions are located into Radio Network Controller (RNC). In addition, terminals and Node B also participate in the resource management in term of power control and load control.

Providing multimedia services with diverse bit rates and quality constraints places stringent requirements on resource management for the UMTS WCDMA downlinks. The challenge is to come up with bandwidth-management algorithms that are adaptable and efficient for varying traffic conditions and to accommodate as many heterogeneous multimedia traffic services as possible while ensuring QoS guarantees [93]. Currently, intensive research is being carried out on bandwidth-reservation and call-admission schemes in multimedia cellular networks [15][93][94].
Chapter 2. Interworked DVB and UMTS and State of Art on Resource Management Schemes

The objective of these schemes is to reduce the handover-dropping probability and/or the new-call-blocking probability. The basic idea is to dynamically reserve a set of bandwidths, exclusively for handover calls, to keep the handover-dropping probability lower than the new-call-blocking probability. Recent advances in multimedia coding have made it able to tolerate and adapt to transient fluctuations in QoS. The adaptation is typically achieved by the use of an adjustable-rate codec or by employing hierarchical encoding of voice or video streaming [95][96]. The codec, along with adequate buffering before play-out, can allow applications to gracefully adapt to temporary bandwidth fluctuations with marginal or no degradation in overall quality. This graceful adaptive nature of multimedia applications is utilized in [12] [15] [97] [98] to propose an adaptive resource allocation scheme which utilizes bandwidth adaptation algorithms to improve resource utilization while keeping call blocking and call dropping probabilities to a minimum.

2.2.2 Resource Management in Multicast-Capable Wireless Network

The main difference between the resource management mechanisms for multicast in multicast-capable wireless networks (such as 3G/MBMS, DVB) and those for unicast in traditional mobile networks (such as UMTS) mainly comes from the unidirectional nature of the multicast-enabled systems and the point-to-multipoint service topology. Novel packet scheduling algorithms were proposed for efficient multicast services scheduling over satellite UMTS designed for efficient support of MBMS [89]. The satellite radio access network (RAN) cannot have real-time feedback from the user groups (e.g., user-side measurements). The absence of satellite return link directly restricts the system short-term resource management functions, i.e., no power control is feasible, and the packet scheduler decides on its allocations without knowledge of the state of individual channels (i.e., channel-state-dependent scheduling is not possible). Even if a return link is available, such as via terrestrial UMTS, a big amount of feedback information from the receivers to the sender could cause the network congestion in the uplink due to point-to-multipoint nature of services.

As for the admission control for multicast services, it comprises the set of actions taken by the network during the phase of service establishment and service re-negotiation to decide whether to accept or reject a user group request. A new user group request can be accepted only when there are adequate network resources available to guarantee the QoS of all existing and the requesting services. When multiple services are transmitted through one multicast transmission channel, the required resources need to satisfy the most demanding of the services (in terms of such as the bit error rate). Another research issue for resource management for multicast services is broadcast scheduling, especially targeted at push and store services. It determines what program is transmitted on the network and when. Carousels are a common way of broadcast scheduling and
delivery in a push system, where users cannot submit requests to the server either because there is no return link to enable user interactivity or because it is not efficient to use it [86]. Closely related to broadcast scheduling is the concept of cache management [90]. The use of clever caching mechanisms at the terminal can significantly reduce the mean response time and hence improve the perceived user interactivity. An alternative to push is the pull mechanism, where a return channel is used and the user places a request for an item directly through the network. Ideally, pull should offer an on-demand service. Hence, the scheduling mechanism is also called multicast-on-demand scheduling, which will be discussed in depth in Section 2.2.4.

Within the MBMS standardisation process, work has been done on the selection criteria of point-to-point (ptp) and point-to-multipoint (ptm) radio bearers [8]. It has been specified that UTRAN should provide both a ptp and ptm capability for MBMS. The ptp bearer will be a dedicated channel that is bi-directional with inner- and outer-loop power control. The ptm bearer will use common channels in the downlink only. Power levels are either constant or power control may be applied to a certain extent. In [91], it is shown that about 28% of the sector total power (43dBm) has to be allocated to a single 64kbps ptm link if full cell coverage is required. This makes ptm too expensive since the overall system capacity in the downlink is mainly limited by the power resource. Therefore, due to the lack of efficient power control mechanism for ptm bearers, in some situations, serving a user group with multiple ptp bearers is more resource efficient than using one ptm bearer. Several selection criteria have been proposed to perform the decision-making. The first proposal is a threshold decision procedure based on the number of activated users within a cell as the method for selecting between ptp and ptm. However, this simple way may not be optimum because it does not take for example user location into account. Therefore, it has been proposed in addition to the number of users, the downlink radio resource environment such as radio resource efficiency should be considered as another factor for the selection of the connection type. This work is still ongoing within the MBMS standardisation process.

It is clear that the lack of efficient power control in ptm link could lead to worse resource efficiency of multicast transmission compared to unicast transmission in some cases. For instance, when a small number of users are in a cell, unicast could be more efficient. This essentially means that MBMS is not feasible unless certain advantaged powers saving methods are used. It has been investigated that based on the functionality currently available in 3GPP UTRAN Release-5, 28% of the total base station power would be required for providing 64kbps PtM MBMS in a macro cellular environment [66] [70]. This can be reduced to 4-12% through the introduction of new techniques in the Release-6 physical layer, such as macrodiversity combining. Macrodiversity for PtM MBMS cannot operate in the same manner as the soft-handover for dedicated channels, due to the broadcast nature of the system. Two types of macrodiversity combining have been proposed for Release-6: selection combining and soft combining. It has been found that the use of selective
combining and soft combining significantly reduce the transmit power requirement for one 64kbps MBMS channel to 9.3%-12% and 4.6%-7.6% of the Node-B power respectively, in order to achieve 95% coverage at 1% BLER. Moreover, with the addition of receiver diversity techniques, 40% further power reduction is achievable. In addition, simulation results have indicated that the provision of up to 256 Kbps PtM MBMS is likely to be viable with Release-6. A dynamic power setting method was proposed in [92]. In this method, using the path loss information of the MBMS users in the cell, the transmission power can be adjusted dynamically according to the worst MBMS user instead of covering the whole cell. Note that the exact power saving method has not yet been standardized for MBMS.

2.2.3 Resource Management in Multi-Radio Environment

It is envisioned that the future radio communication environments comprise a number of different access technologies and different administrative domains in which the coverage of one network overlays the coverage of another. Such an environment is referred to as a heterogeneous network environment. In this environment, users equipped with multi-mode terminals can potentially communicate with service providers using several connections of diverse access networks that might be owned by different operators. These connections typically have different properties and are priced separately. The objective of resource management in such an environment is to give enhanced overall resource efficiency by selecting the most efficient (combination of) access network(s) based on the trade-offs between resource usage, costs, end user preferences, QoS requirements etc. Moreover, resource management should give support for coordination between different administrative domains (i.e., multiple operators), and multicast/broadcast over multiple networks [11]. Therefore, it is essential to choose appropriate transmission path/access network for content-delivery to prevent the expensive use of radio resources and user dissatisfaction.

In the literature, the decision-making problem regarding the access network is called different names, such as always best connected (ABC) problem [33], network selection problem [30][31][34], traffic distribution problem [32][35], or user assignment (UA) problem [36][37]. We use the term of UA on behalf of all the access network-related decision-making problems for the sake of easy description in the rest of this document.

Recent research efforts are mainly focused on addressing the UA problem for one-to-one mapping unicast services [30] - [37], i.e., a copy of data is transmitted only to a single user at a time. Hence, the selected transmission network for delivering service is exactly the access network where the interested user is assigned. However, multicast applications obviously apply one-to-many mapping from a service to a group of users, i.e., one transmission for the desired service is shared by a group of users simultaneously. Therefore, the access network for users is still
unknown even if the transmission network for delivering the desired service has been determined. Nevertheless, transmission network for desired service can be derived from the access networks chosen for interested users, although it does not work the other way around. In the most UA problems [30]-[32][34]-[37], a user is allowed to access to in maximum one network at one time. On the contrary, the UA problem in [33] assumes that a user is able to access to multiple networks at one time. Each network is used to transmit one of several traffic flows that compose the required data session.

A user-centric concept is proposed in [30][31][33][34] to solve the UA problem in a unicast service scenario. For a user-centric assignment, the terminal of individual users determines the access network based on the knowledge obtainable at the user side such as, the cost, preference and perceived QoS etc. However, this mechanism likely leads to inefficient resource utilization due to the competing users who have different preference in services and access networks. In particular, for multicast-based services, the one-to-many transmission characteristic makes this situation worse because of the increased heterogeneity of users interested in the same multicast content. Some user assignment mechanisms consider both the knowledge at user side, and the knowledge of service type and overall load information at network side. In [35] [36] [37], principles have been discussed for allocating multiple services and corresponding users in a multi-access wireless environment. Based on the service type, the traffic demand volume and the capacity of each subsystem, service allocation algorithms are derived to maximize the combined multiservice capacity, such as [36][37], or the total utilities, such as [35]. However, they are not targeted to real time service requests. Instead, their work handles service management related requests. The objective is to determine for (potentially new) services the appropriate QoS levels and networks, in a set of service area regions and time zones.

Another solution to the unicast UA problem is the reinforcement Learning (RL) mechanism [32] that was proposed to determine suitable access networks in response to real time requests of individual users for the purpose of maximizing the efficiency of resource utilization and the revenue of network operators. However, this complex intelligent algorithm is computationally inefficient, thereby not suitable for actual communication systems that require quick decisions. Moreover, the one-to-many transmission nature in multicast makes the UA problem more complex than in a unicast situation in terms of finding the optimal access networks for users requiring the same content, because multicast UA has to be done for all users in a group simultaneously. As opposed to multicast UA, unicast UA is done for one single user at a time. This situation is made worse with the increased size and the enlarged distribution area in the number of cells of a multicast group. Therefore, the traditional solution to unicast UA problem would not be suitable for multicast UA problem, because the decision that optimizes one transmission for a single user may not optimize one transmission for an entire group.
In addition, in recent research, the monetary cost is considered to give business incentives for selecting the access network [30][35]. A user-centric UA algorithm is proposed in [30] in a multi-access network scenario. This algorithm focuses on the maximization of consumer surplus, i.e., the difference between the value of the data to the user and the actual price charged, when selecting the best available connection for transferring non real-time data, with user specified time constraints. With reference to the mechanisms proposed in [35], as we have discussed before, it aims to find an optimal allocation of the traffic demand volume to networks so as to maximize the total system utilities that can be viewed as the profit of network operators.

Similar with the UA problem, in the area of wired networks, the router selection problem finds the best router to optimise the resource utilization and meanwhile satisfy end-to-end QoS requirements in both intra-Autonomous System (AS) and inter-AS [42]-[45]. Each router in the router selection problem could be regarded as an access network in the UA problem. Different from the UA problem that can be addressed from both network side and user side, the current solutions to the router selection problem is network-centric.

### 2.2.4 Multicast-on-Demand Scheduling

Multimedia applications will be the major bandwidth consumers of future networks. Multimedia-on-demand (MoD) is the basic technology for many important multimedia applications such as digital video libraries, distance learning, company training, and electronic commerce. Such a system typically stores a large number of multimedia files in servers. When a request for a particular content arrives, this content is accessed from some disks in a server and transmitted to the client in an isochronous stream via certain networks.

In order to offer cost-effective MoD services, especially when a big number of users have the same interests, multicasting techniques can be used, where users requesting a certain content are served with a single stream without consuming extra resources. A typical multimedia multicasting system generally consists of three components: a central multimedia server, a set of multicast-capable networks (such as cable networks and broadcasting networks), and the clients, as shown in Figure 2-7.
Chapter 2. Interworked DVB and UMTS and State of Art on Resource Management Schemes

Figure 2-7 A typical multimedia multicasting system

The multimedia server stores the video or audio files and uses efficient multicast-on-demand scheduling (MDS) mechanisms to schedule users’ requests. Similar to the broadcast scheduling, MDS determines what files are transmitted through the network and when. The multicast-capable network is used so that multiple users may share a stream (or channel).

A drawback of the multicast approach is that a user request may have to wait for more requests for the same content to arrive in order to share the data stream. The elapsed time between the arrival of a request and the time it is serviced is called the service latency. A concern is that a user may defect if the service latency is kept too long (i.e., renege), and hence is lost. Keeping the latency low without requiring excessive network resources is a great challenge for MoD with multicast techniques.

A multicast method, called batching has been proposed to improve efficiency and scalability for providing on-demand services by grouping together user requests arriving within a period of time (i.e., batching duration) and serving them together using a single multicast channel. If the required contents by users are not displayed within their delay expectation or at their specified time, they leave the system (i.e., renege) and hence are lost, constituting a decrease in revenue and service quality. Since a user will usually suffer a delay from the time when the user sends the request until the time when the content is transmitted, the MoD service is in fact the near real-time MoD, which has been extensively researched in wired network using batching mechanisms [54] - [61].

Some of the earliest batching schemes are first-come-first-serve (FCFS) [60], maximum-queue-length (MQL) [60] and maximum-factor-queue-length (MFQL) [61]. FCFS serves requests according to their arrival order. It gives poor throughput, but treats all video requests equally. MQL allocates a video channel to the longest queue as soon as this channel becomes available. It produces considerably higher throughput than FCFS. MFQL, a variation of the MQL policy, assigns a weighting factor to each queue and allocates an available channel to the queue with the highest weighted length. None of these schemes considers users’ reneging behaviour, i.e., users
are batched only if there are not enough resources available. On the other hand, the batching schemes considering user's reneging behaviour exploit the user waiting tolerance and expect to further reduce the reneging probability.

The two basic methods with the knowledge on user delay preference are known as timeout-based batching and size-based batching [54], depending on whether the users are batched for a fixed, maximum period of time, or are batched until a sufficient number of requests are collected. These two basic batching schemes constitute the foundation for the future extension of the batching scheme. For instance, two hybrid methods have been proposed by Chan [55], namely, combined-for-profit batching and combined-for-loss batching, depending on the way of constituting the two basic batching methods. Shacnai and Yu [57] proposed Max_Batch and Min_Idle. In the Max_Batch scheme, whenever a channel becomes available; a decision is made as to which queue (batch of requests) the channel should be allocated. A channel is allocated to a queue if and only if at least one of the enqueued requests has exceeded a delay threshold. Two Max_Batch policies were defined: the Max_Batch with Maximum Queue Length (BMQ) and the Max_Batch with Minimum Loss (BML) policies. BMQ allocates the available channel to the longest queue, whereas BML allocates the channel to the queue with the highest expected number of losses up to the next time a channel becomes available. In the Min_Idle scheme, videos are classified as either hot or cold according to their popularity. Only hot videos are subject to batching. Hot videos have higher priority than cold videos for channel allocation. Two sets are defined: H and C. Hot videos which have at least one pending request exceeding a delay threshold belong to the set H. The other videos belong to the set C. Whenever a channel becomes available, a video in H is scheduled according either to the longest queue criterion (IMQ) or to the highest expected number of losses (IML) criterion. If H is empty, a video in C is scheduled regardless of how long the requests have been in queue. A cold video may migrate to the set H if any of its pending requests exceeds a threshold. All these schemes are based on partial users delay information (such as knowing the minimum waiting time of the users and the distribution of the remaining waiting time of users). On the other hand, the batching schemes with the full knowledge of user delay behaviour are expected to achieve lower reneging probability.

In [58], a new batching policy, called look-ahead-maximize-batch (LAMB) is proposed, which is a variant of the Max_Batch scheme. LAMB considers a queue eligible for channel assignment only if any of its Head-of-the-Line (HoL) users is about to exceed his/her delay tolerance. Moreover, instead of minimizing the number of losses expected by the next scheduling point, as it is done in BML and in IML, LAMB minimizes the losses in a batching window, which includes the reneging time of all pending requests at that scheduling time. However, LAMB is quite complex and requires continuous computationally-intensive optimisation in the server, which is not suitable to be used in real time. To overcome the disadvantage, a simple but still effective
batching scheme was proposed by Chan [56], termed delay-aware broadcasting (DAB), given the delay preference of each user. In this scheme, the video server serves the user when it is about to renege (i.e., at its maximum waiting tolerance), along with the other requests for the same content in the queue.

To further reduce the reneging time caused by the batching method, a new method called patching was proposed by combining multicasting and unicasting [102]. Under this scheme, the first user in the system is served with multicast. The following user requests are firstly served by respective unicasting streams to play the new start-up flow from the start. At the meantime, the user buffers the future stream from the existing multicast. Once the new flow has been played back to the skew point, the catch-up flow can be terminated and the original multicast can be shared. This approach does not only enjoy the saving of bandwidth as batching but also introduces zero start-up delay. However, it requires a more complicated control system and is unfavourable for high request bursts.

On the other hand, different combinations of delivery methods (such as unicast, multicast and broadcast) are also suggested to achieve delay-resource cost trade-off. Lee [103] combined unicast and broadcast services while Poon [104] combined unicast, multicast and broadcast services together.

Furthermore, the batching scheme has been analysed from a business perspective in [55][56][59]. Efficient batching mechanisms and price models for MoD services were proposed to maximize the profit earned by the service providers.
Chapter 3

3 Multicast Resource Management

This chapter aims to introduce research issues in MRM in terms of algorithm design and architecture definition. Firstly, the service model considered in the research, GMoD (i.e., group-based multimedia-on-demand) service, will be introduced. Following this, two research issues will be specified, namely MDS (i.e., multicast-on-demand scheduling) and MUA (i.e., multicast user assignment). The motivations behind these research issues will be presented and the challenges will be discussed. In addition, the distributed resource management that has been proposed to implement the MRM in the composite radio environment will be explained. This resource management architecture is constructed based on an infrastructure of interworking of networks (IoN), consisting of multiple access networks.

3.1 Service Model

Multimedia-on-demand (MoD) service is adopted in our research. In this service mode, a return channel is used and the user places a request for a particular service directly through the network. In addition, a streaming real-time application is taken as a case study in this work as it is envisioned as an ideal driver for the next generation of Internet-based mobile systems [6]. However, streaming applications, such as delivering multimedia contents, video-on-demand, multiplayer games is far more demanding in bandwidth, loss, and reliability than traditional voice services, and is more bandwidth intensive and latency-sensitive than email and short messages. Therefore, a systematic study is required before the streaming applications are implemented in practical mobile systems.

The scarcity and large fluctuations of link bandwidth in wireless networks have motivated the development of adaptive multimedia applications in which the bandwidth of a connection can be dynamically adjusted to adapt to the highly variable communication environment [12][15]. Example of adaptive multimedia traffic includes MPEG-4 [13] and H.263+ [14] coding for audiovisual contents, which are expected to be used extensively in future cellular wireless networks. Additionally, third generation cellular wireless networks, such as UMTS are designed to support adaptive multimedia by controlling individual flows to increase or decrease their
bandwidth in response to changes in traffic load. Therefore, we exploit the rate-adaptive feature of multimedia application to further improve the efficiency of wireless resource utilization.

In our service model, a batch of multicast multimedia contents is stored in an application server owned by an operator. A copy of the multicast content is owned by a multicast group. A user interested in one of the contents can subscribe to one of the multicast groups. A new group is created when the first user sends the request for a content, and this group is released after the required service is finished. All users in the same group are authorized to receive the content. Multicast subscription is performed upon the content request of each user, which is sent via wireless or wired networks before the content is distributed to the user. Via the content request, the user is allowed to negotiate on received service bandwidth with the operator and generates a utility in response to the served bandwidth. The concept of 'utility' is originally taken from economics [39]. It means the capacity of a commodity or a service to satisfy some human want, and the goal of an economic system is to allocate resources to maximize utility. The concept of utility has been widely used for resource optimisation in wireless networks [16][30][33][34][48]. The objective is to maximize total users' utilities. Each user has a utility function $u(r)$ that maps the received bit rate $r$ to the utility that can represent the throughput (i.e., $u(r)=r$) or the user's perceptual satisfaction rating. Since in almost all wireless applications, a reliable data transmission rate is the most important factor to determine the satisfaction of users, the utility function $u(r)$ is a nondecreasing function of the data rate $r$.

3.2 Problem Identification and the Scope of Research

3.2.1 Batching for Multimedia-On-Demand Scheduling

In order to utilize the multicast technology for service delivery, the system needs to batch (also called aggregate) the requests for the same content arriving within a period of time (i.e., batching duration) as a group and then serve this group by a common multicast channel. Clearly, the longer the batching duration, the smaller the number of network channels used (and hence the lower is the network resource cost) and the larger is the batch size in terms of the number of aggregated requests (and hence the higher is the revenue per channel). However, this also implies a longer delay experienced by a user (and hence a worse QoS). Since it is quite typical for users to renge after experiencing long delays, an excessively long batching period may lead to a high loss in user requests (and thus revenue), and in the long run, loss of customers. Therefore, it is very important in practice to strike a balance between profit and quality of service in terms of user delay and user loss rate. This is one of the issues we need to address in the resource management problem for multicast service, which is called MDS (i.e., multicast-on-demand scheduling).
In general, the MDS aims to perform optimal user grouping for multicast contents and service establishment over transmission networks. Most of the previous work studied the MDS problem in a multimedia server which had limited multicast channels [54]-[61]. Although in [55], a genetic resource pool is assumed, there are no limits on the available resources to accommodate multicast traffic. However, in current cellular systems, the major bottleneck preventing the bandwidth-intensive applications from implementation is the scarce radio resource availability. As a result, the objective of RRM is to allocate the limited number or resource (power, spectrum, processing resources, and channel) to satisfy the divergent desires of wireless network members (users, service providers) [82]. Moreover, the future mobile system is envisioned to be a heterogonous infrastructure comprising diverse wireless access technologies and mobile operators [1]-[5]. Therefore, a more advanced MDS scheme is required to offer bandwidth-intensive application in bandwidth-limited wireless mobile systems, especially the future heterogonous infrastructure. The MDS problem is given a systematic study in Chapter 5 and Chapter 6.

3.2.2 Multicast User Assignment in Heterogeneous Wireless Networks

One of the principle research challenges in heterogeneous networks is to find an efficient solution to the VA problem [30]-[37]. The objective of VA is to allocate the most appropriate access network to serve a user who can communicate with a service provider (or a network operator depending on who is offering the services) using several connections of diverse access networks that might be owned by different network operators. These connections typically have different properties and are priced separately. Systematic approaches to VA have to be devised because the inappropriate selection of access networks could lead to expensive use of radio resources and user dissatisfaction. As most research efforts in the literature have been focused on addressing the VA problem for unicast service, in the subject of multicast user assignment, there has previously been little established technical work performed and published. Therefore, a new and efficient mechanism is required to solve the VA problem in multicast-based heterogeneous wireless systems. The user-centric mechanism is commonly used as a solution for unicast VA problem [30][31][33][34]. The access network is selected based on the knowledge obtainable on the user side.

The VA problem is more severe and complicated for multicast services, where a potentially large group of heterogeneous users have to be served simultaneously with shared resources. The heterogeneity of users is defined in terms of the preference of contents, access networks, delay tolerance caused by waiting for the content set-up, terminal capability in terms of the proper air interface, maximum transmission bit rate, location etc. Clearly, in such a heterogeneous environment, the user-centric approach for unicast VA could lead to the situation that the same content may be delivered unnecessarily via large number of radio channels to the same location,
thereby wasting radio resources, and losing profits. Moreover, the one-to-many transmission nature in multicast makes it more difficult to find the optimal access network for each user in a group. This is because multicast UA has to be done for all users in a group simultaneously. As opposed to multicast UA, unicast UA is done for one single user at a time. The multicast UA problem could be made more difficult to solve, when the size of user group and the distribution area in the number of cells of the group increase. Therefore, it is very important and challenging work to address the UA problem for multicast services. In addition, given the rate-adaptive multimedia application adopted in our research, the appropriate transmission rate for a real-time multicast service needs to be determined in an efficient way. Therefore, it is clear that an efficient mechanism is required to select both the transmission network and data rate to serve multicast services to heterogeneous users through multiple networks. We term the particular problem as Multicast User Assignment in Heterogeneous Wireless Networks (MUA-HWN). This problem is given a systematic study in Chapter 4.

### 3.2.3 Distributed Resource Management Architecture

Generally, two potential methods are optional for the resource management in a multiple network scenario, namely the distributed resource management and centralized resource management. In the centralized resource management, the overall resources in the system are managed by a centralized entity located in a single administrative domain, which is able to get full knowledge required for optimal resource management. In the distributed resource management, the overall resources in the system are jointly managed by underlying networks, while each network is allowed to retain control of their respective networks. To perform a jointed resource management, necessary information needs to be exchanged between networks via trustable entities in a secure way. In general, centralized control is beneficial to the infrastructure composed by a number of networks belonging to a same administrative entity for efficient operation, while the distributed control is beneficial to large networks for scalability reasons, or when a central coordination is simply not desirable due to for instance, the fact that the involved administrative entities do not fully trust each other. In the research of this thesis, distributed rather than centralized resource management is chosen in order to respect the autonomy of networks involved, to minimize any alterations to existing infrastructure and to co-operate in a secure and fair manner [25].

Given the choice of the IoN architecture as illustrated in Figure 2-5, to enable the distributed resource management, a new entity is introduced, namely, Resource Manager (RM) that performs corresponding resource management solution for the MDS and MUA problems. This distributed resource management is illustrated in Figure 3-1.
The distributed approach coordinates different networks resource usage without having a centralised entity. Instead, the RM of a network communicates and exchanges resource-related information with its counterpart residing in another network. Such an RM resides in the IoN Gateway (IGW). This approach allows operators to retain full control on their respective precious networks resources and traffic load information. The Signalling and data exchange required for resource management are carried out through IGW and the logical interface between the networks.

Each RM acts independently on the top of each access network. This allows easy future enhancement for involving new types of networks while minimizing the modification to existing network infrastructures. In addition, independent RMs are able to protect the network’s internal information because in order to coordinate resources, each operator needs to communicate via the RMs. Moreover, assigning each network a RM is fair and decreases the risk by putting all duties into one entity.

In general, four functions are designed to implement resource management, which are located in RM in each underlying network. These functions include the resource management trigger (RMT) function that triggers the operation of the resource management algorithm, the service scheduling (SS) function that implements the respective MDS algorithm, the network and QoS selection (NQS) function that implements the respective MUA algorithm, and the resource cost calculation (ReC) function that assists NQS by providing required parameters, such as cost quote. The cost quote is calculated based on the information that is collected by the local monitor (LM). Meanwhile, GM and DPS assist RM by providing required parameters for implementing resource management algorithms. These functions in each IoN GW are also applicable to their counterparts in different access networks when multiple access networks exist. Out of the resource
management functions in a RM, RMT, SS and RCC are able to work in an individual network without interacting with other cooperative networks, while NQS has to coordinate with its counterpart components in other networks via IoN GW.

Distributed resource control gives the promise of efficient sharing of network resources, but requires information exchange between networks. The signals exchanged between networks for resource management include resource metric and the associated parameters for service provision. However, such signalling mechanisms between networks are currently lacking. Therefore, the underlying signalling issues required by distributed resource management need to be investigated, focussing particularly on the requirements of control information flow between different resource management entities and its practicality. Since network operators are not expected to release their real time resource status, commercial constraints on information availability need to be considered while restricting information visibility between network elements. For instance, since the resource management over multiple networks/operators uses different sources of network information as input, e.g., radio or business related, investigating the impact of the exchangeable information on the resource management performance is non-trivial.

More details regarding the investigation of resource management architecture in the context of the IoN architecture can be found in Chapter 6.

3.3 Performance Metrics

In our analysis of MRM algorithms, we use the following performance measures:

- **System profit**: this is the aggregated profit generated by each served user. We consider a general profit concept, which accommodate two different measures, namely network throughput that is the aggregated user throughput, and monetary profit that is the overall revenue by serving users minus the cost of total resource consumption in the networks. The objective of the MRM is to maximize the system profit;

- **Satisfaction ratio**: this is the percentage of users’ requests that are served with respect to the total users’ requests;

- **User request loss probability**: this is the percentage of users’ requests that are cancelled before the service is set-up due to user’s reneging from the system, or being dropped during the horizontal handover, with respect to the total users’ requests;

- **Overall average delay**: this is the average delay from the time the request enters the batching queue until the time the request is served. Only non-reneging users are considered in the delay time measure. In other words, delay is conditioned on non-reneging.
Chapter 4

4 Multicast User Assignment in Heterogeneous Wireless Networks

Suppose there is a set of users in a heterogeneous wireless network environment, and each of them requires a particular multicast content. The problem of MUA (Multicast User Assignment) is to carry out two-dimensional assignment respectively for the serving cell of the user, and the QoS of the reception of the user.

This chapter aims to explain the MUA algorithms in the context of rate-adaptive multimedia applications, heterogeneous access networks and multicast transmission. This chapter will firstly introduce the concept and justify the assumptions adopted in the MUA. The detailed modelling of the multicast user assignment in heterogeneous wireless networks problem is then described. The solution to this problem aims to maximize the total profit in the system. A general profit formulation is considered, which can accommodate different measures such as the network throughput, user's satisfaction rating, and monetary revenue etc., each of which corresponds to a specific optimisation objective. Several alternative algorithms maximizing the total profit are outlined. Following this, the simulator that has been constructed to evaluate the performance of the algorithms is described. This includes a description of the simulator structure and the parameters and models that are used. Finally in this chapter, the proposed MUA algorithms are evaluated by system level simulations.

4.1 Background Knowledge

4.1.1 Rate-Adaptive Application

There is ongoing interest in QoS provisioning in an adaptive multimedia framework, in which a QoS/bandwidth adaptive algorithm needs to be used in conjunction with the call admission control algorithm. The objective of bandwidth adaptation is to further improve the efficiency of resource utilization, in particular the radio resources, by adapting the bandwidth of a certain service according to the current resource availability, user preference, user location etc. The rate-adaptive feature of multimedia application enables the bandwidth adaptation to network...
conditions. For instance, voice applications can be encoded at a rate ranging from 2 to 128 kbps by choosing appropriate encoding mechanism or dynamically modifying the encoding parameters. Similarly, video applications can be made rate adaptive by using a layered coding method. For example, the MPEG-2 video/audio compression standard defined different layers and profiles to achieve SNR and spatial scalability [26]. The lowest layer (i.e., the base layer) consists of critical information for decoding the image sequence at its lowest visual quality. Additional layers provide increasing quality. Another promising approach for adaptation is in the use of embedded coding schemes, such as the wavelet-based JPEG-2000 image-coding standard [27]. Instead of a few discrete coding rates provided by a layered coding scheme, continuous bit rate can be achieved. Similarly, MPEG-4, which is the new generation multimedia communication-coding standard, has the fine-granular scalability (FGS) mode [28].

There has been a large amount of research in the mechanism of bandwidth allocation, call admission and degradation for adaptive multimedia services in wireless cellular networks. Basically, the layer multimedia is used to trade off the carried traffic and the bandwidth degradation, i.e., minimizing the overload probability of the system by temporarily reducing some receivers' subscription levels, and thereby the received data rate. These mechanisms apply to all existing connections in a cell and the new connection request, reallocating the bandwidth among them in order to fully utilize the radio resources. Degrading the bandwidth of required or existing multimedia connections in a controlled manner has been shown as an effective way to improve the overall system performance [29][51][52]. Advanced but time-consuming algorithms are proposed in [51][52] to reallocate bandwidth among existing and new multimedia connections in real time when a new request arrives. This would lead to frequent changes in the connection quality and is not desirable for applications and users. In addition, the execution time of the algorithms is another issue that needs to be justified especially for real time applications. To avoid the frequent changes and to still maintain efficient utilization of radio resources, a simple rate adaptation method is proposed in [29] to allocate bandwidth for a call only upon its admission and handover. The connection request specifies the range of bandwidths required to be supported by the network as [MinBW, MaxBW], where MinBW and MaxBW denote the minimum and maximum bandwidth requirements respectively. Adaptation first takes place while admitting a new connection request. If the network has enough resources available, the request is admitted at MaxBW; otherwise, it is admitted at a lower bandwidth. If the network is overloaded and MinBW cannot be satisfied, the request is blocked. Bandwidth adaptation also takes place at the time of handover occurrence. After the handover, the bandwidth of a connection could be changed to adapt to the amount of available resources in the targeted cell, for instance, degrading from a higher rate to a lower rate or vice versa.
For the adaptive multimedia transmission via multicast, a multiple rate or a single rate control method is widely used [41]. Some research work supporting wireless multicast adopt the single rate control method in a single cell, such as [49] [50]. The sophisticated control method supporting the multiple rate control in a single wireless cell can be found in [48]. For a multiple rate method, a session typically consists of more than one channel and the sender sends packets to the channels in the session at rates that do not depend on the receivers. Each receiver adjusts its reception rate during its participation in the session by joining and leaving channels dynamically depending on the available bandwidth to the sender independent of all other receivers. For single rate control method, a session typically consists of one channel and the sender sends packets to the channel at variable rates over time depending on feedback from receivers. Each receiver remains joined to the channel during its participation in the session. We adopt the multiple rate method to serve users in different cells. With this method, a content session is delivered via different channels with different rates, each channel referring to a physical connection in a cell. The particular transmission rate depends on the available bandwidth in the cell. However, in a single cell, we adopt the single rate method in order to reduce the management complexity.

4.1.2 Utility Maximization for Resource Management

The concept of 'utility' is originally taken from economics [39]. It means the capacity of a commodity or a service to satisfy some human want, and the goal of an economic system is to allocate resources to maximize utility. Recently, the concept of utility has been widely used for the resource management in both wireless and wired systems, such as the QoS optimisation for adaptive multimedia system [46][47][48], the optimal service routing and admission control across a future Internet supporting sessions requiring QoS guarantees [43], the wireless resource optimisation problem [38], the traffic distribution and the access network selection in composite radio environments [33][34][35]. Most of these problems have a common objective that is to maximize the system profit that is calculated as a difference between the aggregated session utilities/user utilities and the total cost of consumed resources. Note that, for some system profit maximization problems, the resource cost is not considered or it is integrated with the session/user utility. In these cases, the system profit maximization is also called system utility maximization. In the rest of the thesis, these two terms are interchangeable.

Some of the system utility/profit maximization problems employ a network-centric solution, where the total resources are coordinated (by a network operator) among all users in the system so as to optimise the usage of overall resources while satisfying the QoS requirements. This is also a general objective of resource management. As opposed to the network-centric method, some algorithms in the literature, such as those in [33][34] addresses the user’s utility maximization problem with the objective of maximizing a single user’s utility from a user-centric view. Clearly,
to optimize the efficiency of overall resources in multiple networks, the network-centric control mechanism is more efficient than the user-centric control mechanism. The performance of these two mechanisms will be compared and demonstrated later on by simulations.

The system utility can be calculated as either the total session utilities \([43][46]\), or total user utilities \([34][35][38][47][48]\). For the session utility, the total session utilities can be interpreted as the system revenue or the system profit. For the user utility, each user has a utility function that maps the user’s perceptual satisfaction rating for assigned certain resources to a real number. The session utility can be easily derived from the user utility by aggregating the total utilities of interested users involved in that particular session.

To calculate the system utility, it is important to decide the utility function. As justified in \([38]\), in almost all wireless applications, a reliable data transmission rate is the most important factor to determine the satisfaction of users. Therefore, they assume the utility function \(U(r)\) as a nondecreasing function of the data rate \(r\). In particular, when \(U(r) = r\), the utility is just the throughput, which is the objective of most traditional network optimisations. When the utility is used to capture the user’s subjective feeling, such as the level of satisfaction for assigned certain resources, it cannot be obtained only through theoretical derivation. Instead, it needs to be estimated from subjective surveys.

In general, the adoption of a generic user utility function enables accommodation of different measures, such as throughput, user’s satisfaction rating and revenue. Moreover, the extra consideration of the resource cost extends the measures to include the profit that is the difference between the user utility and cost.

### 4.1.3 Knapsack Problems

In mathematics, optimization is the discipline which is concerned with finding the maxima or minima of functions, possibly subject to constraints \([72]\). An example of an optimization problem is the following: maximize the profit of a manufacturing operation while ensuring that none of the resources exceed certain limits and also satisfying as much of the demand faced as possible. The following steps are necessary for solving an optimization problem.

- Define precisely the elements of the optimization problem;
- Organize these elements in a way so that the problem appears solvable;
- Choose a suitable method for solving the problem.

Combinatorial optimization is a branch of optimization in applied mathematics and computer science, related to operation research, algorithm theory and computational complexity theory \([73]\). It is the process of finding one or more best (optimal) solutions in a well defined discrete problem.
space. Combinatorial optimization problems are concerned with the efficient allocation of limited resources to meet desired objectives when the values of some or all of the variables are restricted to be integral.

The knapsack problem (KP) is an important class of combinatorial optimization problems. It derives its name from the maximization problem of choosing as much as possible essentials that can fit into one bag (of maximum weight) you are going to carry on a trip. Many practical problems, such as capital budgeting, industrial production, menu planning, cargo loading, information systems and resource allocation, may be expressed as variants of knapsack problems. For a in-depth discussion of knapsack problems please refer to [53].

In general, a typical KP is described as follows: Suppose one wants to fill a knapsack that can hold a total weight of \( b \) with some combination of items from a list of \( n \) possible items each with weight \( a_i \) and value \( v_i \) so that the value of the items packed into the knapsack is maximized. This problem has a single linear constraint (that the weight of the items in the knapsack must not exceed \( b \)), a linear objective function which sums the values of the items in the knapsack, and the added restriction that each item must either be in the knapsack or not — a fractional amount of the item is not possible.

Various knapsack problems exist, like the classical 0-1 KP that is the essence of other variants of KPs. It is described as follows: suppose there are \( n \) items and a knapsack, with \( v_j \) the value of item \( j \); \( a_j \) the weight of item \( j \) and \( b \) the capacity of the knapsack. The problem is to pick a set of items in order to maximize the total value of the pick such that the total weight of the pick does not exceed the capacity of the knapsack. Mathematically, the problem is stated as follows:

Maximize:

\[ z = \sum_{j=1}^{n} v_j x_j \quad (4-1) \]

Subject to:

\[ \sum_{j=1}^{n} a_j x_j \leq b \quad (4-2) \]

\[ x_j = 0 \text{ or } 1, \; j \in N = \{1, \ldots, n\} \quad (4-3) \]

where \( x_j = \begin{cases} 1 & \text{if item } j \text{ is selected;} \\ 0 & \text{otherwise.} \end{cases} \quad (4-4) \)

Here, the vector of \( x_j \) for \( j = 1, 2, \ldots, n \) is variable. The problem is called the 0-1 knapsack problem because variable \( x_j \) can either take a value of 0 implying item \( j \) is not picked, or a value of 1
implying item $j$ is picked. Any pick of items which satisfy the constraint is called a **feasible solution** of the problem. The solution of the 0-1 KP is the feasible solution which maximizes the sum of the profit of the picked items.

The classical 0-1 KP can be generated for multiple resource constraints (or dimensions). A knapsack problem with multiple constraints is called multi-dimension knapsack problem (MDKP). Suppose there are $n$ items, and $m$ resources. Item $j$ has a value of $v_j$, and resource required by this item is expressed using resource vector $r_j = (r_{j1}, r_{j2}, \ldots, r_{jm})$. The amount of available resources is given by resource vector $R = (R_1, R_2, \ldots, R_m)$. Mathematically, the MDKP is stated as follows:

Maximize:

$$z = \sum_{j=1}^{n} v_j x_j$$  (4-5)

subject to:

$$\sum_{j=1}^{n} r_{jk} x_j \leq R_k \quad k = 1, \ldots, m$$  (4-6)

$$x_j = 0 \text{ or } 1, \quad j \in N = \{1, \ldots, n\}$$  (4-7)

where  

$$x_j = \begin{cases} 1 & \text{if item } j \text{ is selected;} \\ 0 & \text{otherwise.} \end{cases}$$  (4-8)

Another variation of the classical 0-1 KP is the use of multiple knapsacks $K = (K_1, K_2, \ldots, K_m)$. The generalized form of KP is the generalized assignment problem (GAP). Mathematically, the GAP is stated as follows: given $n$ items and $m$ knapsacks, with $v_{ij}$ the value of item $j$ if assigned to knapsack $i$, $a_{ij}$ the weight of item $j$ if assigned to knapsack $i$ and $b_i$ the capacity of knapsack $i$. The problem is to assign each item to exactly one knapsack so as to maximize the total value assigned without assigning to any knapsack a total weight greater than its capacity, i.e.,

Maximize:

$$z = \sum_{i=1}^{m} \sum_{j=1}^{n} v_{ij} x_{ij}$$  (4-9)

subject to:

$$\sum_{j=1}^{n} a_{ij} x_{ij} \leq b_i, \quad i \in M = \{1, \ldots, m\}$$  (4-10)
4. Multicast User Assignment in Heterogeneous Wireless Networks

\[ \sum_{i=1}^{m} x_{ij} = 1, \ j \in N = \{1, \ldots, n\} \]  
(4-11)

\[ x_{ij} = 0 \text{ or } 1, \ i \in M, \ j \in N \]  
(4-12)

where \( x_{ij} = \begin{cases} 1 & \text{if item } j \text{ is assigned to knapsack } i; \\ 0 & \text{otherwise.} \end{cases} \)  
(4-13)

The problem is frequently described in the literature as that of optimally assigning \( n \) tasks to \( m \) processors (\( n \) jobs to \( m \) agents, and so on), given the value \( v_{ij} \) and the amount of resource \( a_{ij} \) corresponding to the assignment of task \( j \) to processor \( i \), and the total resource \( b_i \) available for each processor \( i \).

Since the 0-1 KP and its variants are Non-deterministic Polynomial-time (NP) hard [53], the worst-case computation time of the optimal solution grows exponentially with the size of the problem. For this reason, there are two types of algorithms developed: complete algorithms and approximate algorithms. Complete algorithms are guaranteed to find for every finite size instance of a KP an optimal solution in bounded time. In approximate methods, the guarantee of finding optimal solutions is sacrificed for the sake of getting good solutions in a significantly reduced amount of time. Considering that the complete algorithms often lead to computation times too high for practical purposes, the use of approximate methods to solve KPs has received more and more attention in the last 30 years.

A greedy approach has been suggested for approximate algorithm of KPs [53]. This approach sorts the items in decreasing order of value-resource ratio and then proceeds to insert them into the knapsack, starting from the first element (the greatest) until there is no enough space in the knapsack to accommodate more.

4.2 Problem Description

Instead of a user-centric mechanism, we use a network-centric mechanism that enables the network operator to make the decision of the best combination of access networks and transmission rates to serve users in difference locations, by providing the operator with sufficient information, such as, users profile, content profile, and network status. Since the operator has full knowledge concerning the contents, users and underlying access network status, they are expected to be able to perform sophisticated computational intelligence technology to perform the optimal assignment of resources. In addition, as mentioned before, we study the scenario where the underlying cooperative networks are multicast capable, which means that the required service will be transmitted using a multicast channel for the purpose of the delivery efficiency. In this
context, we extend the traditional user assignment problem, usually associated with unicast services, to the user assignment problem for multicast services, which is termed *multicast user assignment in heterogeneous wireless networks* (MUA-HWN).

The solution to the MUA-HWN problem is to determine the best combination of access networks along with the transmission rate to serve the required multicast contents to interested groups of users simultaneously, so as to maximize the system profit whilst at the same time satisfying the QoS requirements of the contents and respecting wireless network resource constraints. The 'profit' adopted in our research can be interpreted in different way in order to accommodate different measures. For instance, from a technical perspective, the profit can be the network throughput or the number of served users; from an economic perspective, the profit can be interpreted as the revenue paid by served users, or the profit of a service provider/network operator, i.e., the difference between the total revenue and the total cost, by taking into account the cost of consumed resources.

Our research work attempts to be generic and aims to be applicable to any architecture comprising multicast-enabled wireless technologies (e.g., 3G/MBMS, WLAN, DVB-T/H/S, or future new technologies), by exploring the MUA-HWN problem from a high layer independent of the specific underlying access network, e.g., network structure, business model, wireless access technology etc.

In the following, the MNS-HWN problem is described in a high level and formal manner. An optimisation model is proposed and presented to solve the MUA-HWN problem. Based on the optimisation model, the MUA-HWN problem is formulated and proven to be NP-hard.

**4.2.1 High Level Problem Description**

![Figure 4-1 High-level description of the MUA-HWN problem](image-url)
Figure 4-1 illustrates an example of the MUA-HWN problem from a high level. In general, the MUA-HWN system requires two types of inputs, the user assignment request and the network status of the underlying access networks (e.g., \( n_1 \) and \( n_2 \)).

A user assignment request consists of the following information:

1. Profiles of the multicast contents that have been subscribed by users and will be distributed through networks to the users via multicast channels;

2. Profiles of the users who have subscribed to the multicast contents.

The content profile is directly available from the service or content providers. It provides the information of content quality and price. For rate adaptive multimedia applications, each multicast content is encoded with a set of discrete data rates, each of which has a particular price. On the contrary, in the case of constant-rate multimedia applications, for each content, only one data rate and one price are offered. A user can only be served with a data rate that can be supported by both his/her terminal and the network. All users served in a multicast content belong to the same multicast group.

A user’s profile provides the user’s preference of contents, by which the number of users interested in a certain content can be counted. The user’s preference is available from the user’s explicit request for subscribing to his/her preferred contents. This can happen anytime before the exact content distribution time via a variety of methods, e.g., mobile phone, internet, etc. A user’s profile also provides the user’s location in terms of the identities (IDs) of feasible residing cells in hybrid access networks. The knowledge of user’s location ensures accurate content delivery to particular cells where subscribed users are located, instead of the cells where there are no any subscribers. Hence, the resources can be used more efficiently. These cell IDs can not be obtained straight forward from users’ requests. Instead, they are discovered efficiently by using hybrid-paging schemes achieved by cooperating 3G and DVB networks [40]. The validity of a cell to a user depends on several factors such as the user’s geographic position in a cell, service area coverage and the network access interfaces a user’s terminal has. For instance, a 3G/MBMS cell is regarded as a valid residing cell to a user not only because the user is residing in the cell with the proper air interface to receive MBMS services, but also this 3G/MBMS cell belongs to the appropriate service area of the multicast application. In addition, the user profile provides information about the terminal capacity, i.e., the maximum supportable data rate. Therefore, the service rate provided to a user cannot exceed the capacity of the terminal.

Another piece of information provided by the user profile is the user utility function that maps the utilized network resources into a real number. The exact choice of a utility function depends on the detailed system design. For instance, from a network operator’s perspective, an absolute utility function, such as throughput, is preferable if the revenue is proportional to the total...
received bandwidth of all the receivers. Instead of limiting our scope to a specific utility, in line with [38] [48], we assume the utility function $u(r)$ is the function of the transmission rate $r$ assigned to a user.

The status of the networks includes the following information:

1. The network identity that distinguishes the different network underlying technology (e.g., UMTS, broadcasting) as well as the network operators (e.g., Vodafone, O2), considering the fact that in practice, one operator may own multiple access networks;

2. Cell information of each access network, e.g., resource constraints;

3. The 'cost quote' offered by each network.

An operator $n$ owns an access network that consists of a set of cells, each of which supports multicast transmission. All the cells of different networks in the system form a set $H$. The information of $i$th cell owned by the network $n$, $H_n(i)$, includes the resource usage constraints, e.g., the maximum available resources in this cell. Note that resources in wireless systems can be measured by different units, e.g., bandwidth in date rate, frequency, power, code etc., depending on the access network technology. These units normally are closely related with each other. For instance, the achievable maximum bandwidth in a cell is usually positive proportional to the maximum available frequency, power and codes. In the rest of this thesis, unless otherwise stated, resources are assumed to be the bandwidth in date rate.

Two types of information make up a 'cost quote', namely the resource cost per unit volume of the transferred data and the respective supportable QoS level $^1$, for which the resource is committed, in every cell of network $n$.

The objective of MUA-HWN is to maximize the system profit that is calculated as the difference between the aggregated user utilities for all contents and costs of consumed resources in the system. The objective is presented by an objective function that is associated with both inputs and outputs. As outputs, in response to a user assignment request, the MUA-HWN optimization operation will generate the maximum system profit and meanwhile make the following decisions:

1. Determine whether the subscribed user is admitted;

2. If the user is admitted, the MUA-HWN also determines the appropriate serving cell for the user to receive the subscribed content via multicast channel;

---

$^1$ Note that many QoS parameters could be applicable in our algorithm, such as, bandwidth in data rate, delay, jitter, packet loss ratio etc. In this document, for the purpose of simplicity, only the bandwidth is considered as the QoS parameter to match the offered set of transmission rate of contents.
3. In the case of the rate-adaptive content, the MUA-HWN also determines the appropriate transmission rate for the user.

Note that after the assignment of the user to an appropriate network and transmission rate, the assignment of the required contents is also sorted out based on the policy that a multicast content is delivered in the cell where there is at least one user assigned, with a particular data rate that the user is assigned to.

To achieve the objective, some constraints should be respected. The constraints of the MUA-HWN problem fall into the following categories.

1. The user should be served with a data rate that is compliant with the content profile; supportable by the user terminal and generates positive user utility;

2. The available resource in a cell is limited, so to accommodate a user the capacity limits of the chosen cell should not be violated;

3. A user can be served at most with one data stream (i.e., one copy of a particular content) at one time. Although currently, some kinds of terminals are capable to receive multiple streams simultaneously [4], it is not necessary in some cases, such as in MoD service where users are more likely to watch just one multimedia stream at any one time;

4. A user can be assigned to at most one cell at one time, i.e., a user's traffic is not allowed to be split in order to maintain low complexity with respect to system management. This complexity could be significantly enhanced by synchronizing the transmission of multiple simultaneous data flows through different networks. In addition, the splitting of user's traffic at one time will make the business relationship between operators more complex than the case where the desired service is allowed to transmit via only one network at one time. Nevertheless, it is worth mentioning that the transmission network could change during the service transmission in order to maintain high and stable service quality, adaptive to the time-varying network and channel conditions;

5. A content is served with at most one data rate at a time to interested users in a cell. This matches the feature of the single-rate multicast transmission [41][49][50].
4.2.2 Formal Problem Description

4.2.2.1 Problem Formulation

![Diagram](image)

Figure 4-2 Optimisation model for MUA-HWN problem

Based on the optimisation mode illustrated in Figure 4-2, the MUA-HWN problem can be described as follows.

Suppose there is a set of multicast contents $C$. Each content $c \in C$ is characterised by a transmission data rate set $S_c$ (e.g., average required bandwidth in Kbps), and corresponding content price set $V_c$ (e.g., price per Kbits).

Suppose there is a set of multicast-capable cells $H$. Each cell $h \in H$ is characterised by $(m_h, n_h, r_c_h)$, where $m_h$ denotes the maximum available bandwidth capacity in the cell $h$ (e.g., Kbps), $n_h$ the type of the network to which the cell belongs (e.g., 3G or broadcasting network), $r_c_h$ the unit resource cost of the cell $h$ (e.g., cost per Kbits).

Suppose there is a set of users $U$. Each $u \in U$ is characterised by $(O_u, g_u, k_u, \mu_u(d_u))$, where $O_u \subseteq H$ is the candidate cell set of the user based on his/her geographic location and the terminal’s supportable air interfaces; $g_u$ denotes the content ID subscribed by the user; $k_u$ represents the capacity of user’s terminal in terms of the maximum supportable transmission rate and $\mu_u(d_u)$ the utility of a user $u$ served with the actual received bit rate $d_u$. If $d_u$ is not available, $\mu_u(d_u)=0$. $d_u$ should be chosen from the supportable data rates of the user’s desired content $g_u$, subject to the user’s terminal capacity limit $k_u$. Therefore, for each user $u$, there is a particular set of transmission rates $R_u$ feasible to choose. $R_u$ contains all the rates not greater than $k_u$ in set $S_{g_u}$. Hence, the user’s utility can be interpreted as $\mu_u(r), r \in R_u$. In addition, the serving cell’s capacity limit is another factor that determines $r$.

A user $u$ generates a particular user profit $p_{u,h,r}$ and consumes a certain amount of user resource $w_{u,h,r}$, given the serving cell $h$ from $O_u$, and the transmission rate $r$ from $R_u$. The user profit is the
function associated with the user utility and resource cost of the network where the user is assigned. More details about the user profit and user resources will be given in the next paragraph. The MUA-HWN optimisation is to assign each subscribed user \( u \) to exactly one cell from \( O_u \), and serve this user with exactly one transmission data rate from \( R_u \), so as to maximize the total aggregate profits \( \sum_{u \in U} \sum_{h \in O_u} \sum_{r \in R_u} p_{u,h,r} \), which is referred to the system profit. Furthermore, each user interested in the same multicast content can be assigned to a different cell and served with different data rate, as long as the capacity \( m_h \) of that particular cell is not violated. Different factors need to be considered when making the decision, such as the set of feasible cells \( O_u \), the set of feasible transmission rates \( R_u \), the utility function \( \mu_u(r) \) for each user \( u \), the cell capacity \( m_h \) and the unit resource cost \( r_c_h \) of each cell \( h \) etc. The following figure Figure 4-3 gives examples for the possible selection of the serving cells and transmission rates.

![Figure 4-3 An example of MUA-HWN problem](image)

In this example, \( u_1 \) is interested in content \( c_1 \). We assume \( u_1 \) has a terminal with multiple air interfaces, moreover it is in the coverage of both access networks. Hence, \( O_u=[h_1, h_3] \). Furthermore, the high capacity of the terminal gives \( R_u=[128, 256, 384] \). As a result, \( u_1 \) could be assigned to cell \( h_3 \) for a higher received data rate (i.e., \( r=384 \) kbps). Or \( u_1 \) could be assigned to cell \( h_1 \) for a lower resource cost, but at the price of lower received data rate (i.e., \( r=128 \) kbps) due to the cell capacity limit. Both \( u_2 \) and \( u_3 \) are interested in the content \( c_1 \). Due to the limits of the 3G cell coverage, \( O_{u2,u3}=[h_3] \). In addition, \( R_{u2,u3}=[128] \) due to the limited terminal capacity. This indicates they can only be served with the rate of 128 kbps in cell \( h_3 \). \( u_4 \) is interested in the content \( c_2 \) and is located in the overlapping area of \( h_2 \) and \( h_3 \). Nevertheless, we assume its terminal can only connect to the cellular network, hence, \( O_{u4}=[h_2] \). Due to the high capacity of the terminal, \( R_{u4}=[32, 64, 128] \). To maximize the user's throughput, we can serve \( u_4 \) with the rate of 128 kbps.
Note that to determine the transmission data rates for these users, the exact utility function is required, as the chosen rate needs to maximize the utility as a response to the received bandwidth.

To further study the MUA-HWN problem, we need to calculate the user profit $p_{u,h,r}$ and the corresponding resources $w_{u,h,r}$ consumed by that user.

In a unicast scenario, the consumed resource $w_{u,h,r}$ of a user can be calculated straightforward from his/her allocated transmission rate $r$ once the assignment for this user is completed, i.e., $w_{u,h,r} = r$. However, in a multicast scenario where one multicast transmission covers a group of users, the average consumed resource of an individual user is inverse proportional to the number of users assigned in the same radio transmission in cell $h$, and thus, the consumed resource is unknown until all the users have been assigned. Mathematically, the consumed user resource is the function of the assigned transmission rate $r$, and the number of users who have been assigned to the same cell $h$ with the same rate $r$ of the required content $g_u$, which is denoted as $M_{g_u,h,r}$. As a result, $w_{u,h,r}$ can be calculated as follows:

$$w_{u,h,r} = \frac{r}{M_{g_u,h,r}} \quad (4-14)$$

This equation indicates that if $M_{g_u,h,r} = 1$, the content is served via unicast to only one user. Otherwise, a group of users will share the resource in a multicast transmission.

The user profit $p_{u,h,r}$ is calculated as the difference between the user utility cost and the resource cost consumed by this user. The concept of profit has different applications in practice. For instance, for a network operator who aims to maximize the network throughput, the user utility is interpreted as the user throughput and the cost is not considered; for a multimedia service provider/network operator who aims to maximize the revenue, the user utility is interpreted as the revenue paid by (the bill of) the user in return for the resource consumption and the cost is not considered; for a multimedia service provider/network operator who aims to maximize the monetary profit, the profit can be defined as total revenue minus total cost, thus the user utility is interpreted as the revenue earned from the user, and the cost needs to be taken into account.

The user utility is calculated by the utility function $\mu_u(r)$ and as we discussed before, the utility is a function of the served data rate $r$. If the content price is included into the utility function, the revenue can be easily represented. For the resource cost, we use a resource-cost mapping function $f_{r \rightarrow c}$ mapping from a specific amount of consumed resources to the monetary cost of consuming that amount of resources. We assume a linear $f_{r \rightarrow c}$, i.e., the monetary cost associated with wireless resource consumption is constant. Hence, the cost of serving the user $u$ with the date rate $r$ in the cell $h$ is $rc_h \cdot w_{u,h,r}$, where $rc_h$ is the unit resource cost of the cell $h$, and is determined by
the network operator according to the expenses of necessary network infrastructure deployment, power consumption, spectrum consumption etc. The user profit is calculated as:

\[ P_{u,h,r} = \mu_u(r) - rc_h \cdot w_{u,h,r} \]  

(4-15)

With reference to the equation (4-14), the equation (4-15) indicates that in the case of non-zero \( rc_h \), the user profit is also the function of the number of assigned users in the same multicast transmission, and thus is unknown until all the users have been assigned.

The inputs to the optimisation process include the multicast content \( C \), cell set \( H \) and user set \( U \). The outputs are the maximum total profits, together with two binary optimal allocation vectors. One is \( x_{u,h,r}, u \in U, h \in O_u, r \in R_u \), which gives the allocation of cell \( h \) from the set \( O_u \) and the data rate \( r \) from the set \( R_u \) for each user \( u \). The other one is \( y_{c,h,r}, c \in C, h \in H, r \in S_c \), which gives the allocation of cell \( h \) and the transmission rate \( r \) for each content \( c \). If \( x_{u,h,r} = 1 \), user \( u \) is assigned to cell \( h \) with rate \( r \). If \( y_{c,h,r} = 1 \) content \( c \) is served to users through cell \( h \) with rate \( r \).

Due to the inherent feature of multicasting, i.e., only one copy of the multicast content will be distributed in the cell where at least one subscribed user has been assigned. Therefore, \( y_{c,h,r} \) is a function of \( x_{u,h,r} \), i.e., \( y_{c,h,r} = 1 \) if \( M_{p,h,r} > 0 \), 0 otherwise. As explained before, \( M_{p,h,r} \) is the number of users who have been assigned to the cell \( h \) with the rate \( r \) of the desired content \( g_{u} \).

\[ M_{p,h,r} = \sum_{u \in C, h \in O_u, r \in R_u} x_{u,h,r} \]

(4-20)

Based on the discussion above, the MUA-HWN problem can be formally stated as follows:

Maximize:

\[ \text{System Profit} = \sum_{u \in U} \sum_{h \in O_u} \sum_{r \in R_u} P_{u,h,r} x_{u,h,r} \]  

(4-16)

Subject to:

\[ \sum_{u \in U} \sum_{r \in R_u} w_{u,h,r} x_{u,h,r} \leq m_h, \forall h \in O_u \]  

(4-17)

\[ \sum_{h \in O_u} x_{u,h,r} \leq 1, \forall u \in U, r \in R_u \]  

(4-18)

\[ \sum_{r \in S_c} y_{c,h,r} \leq 1, \forall c \in C, h \in H \]  

(4-19)

\[ x_{u,h,r} \in \{0,1\}, \forall u \in U, h \in O_u, r \in R_u \]  

(4-20)
Chapter 4. Multicast User Assignment in Heterogeneous Wireless Networks

Constraint (4-17) specifies the limits of cell capacity \( m_i \); Constraint (4-18) identifies each user is assigned to at most one cell; Constraint (4-19) identifies that a content is served with at most one bit rate to interested users in a cell at a time, which indicates the feature of single-rate multicast transmission; Constraint (4-20) requires the output is a binary integer.

### 4.2.2.2 NP-hard proof

In this section, we prove the MUA-HWN problem to be NP-hard by reducing it to a standard NP-hard problem – the Generalized Assignment Problem (GAP) [53].

**Generalized Assignment Problem:** Given \( I \) knapsacks, \( J \) items, with \( v_{ij} \) the profit of item \( j \) if assigned to knapsack \( i \); \( a_{ij} \) the weight of item \( j \) if assigned to knapsack \( i \) and \( b_i \) the capacity of knapsack \( i \), compute an assignment \( f : \{1, \ldots, J\} \rightarrow \{1, \ldots, I\} \) of items to knapsacks such that

\[
\begin{align*}
&\text{(C1)} & \text{The total profit of assignment} \sum_{i=1}^{I} \sum_{j=f(j)=i}^{J} v_{ij} \text{ is maximized;} \\
&\text{(C2)} & \text{The total weight assigned to each knapsack} i \text{ does not exceed the capacity} b_i, \\
& & \text{i.e.,} \sum_{j=f(j)=i}^{J} a_{ij} \leq b_i; \\
&\text{(C3)} & \text{Each item is assigned to exactly one knapsack.}
\end{align*}
\]

We reduce the MNS-HWN problem to the GAP by considering the following issues:

(I1). We map the user set \( U \) and the cell set \( H \) to be the item set and the knapsack set in the GAP respectively. Assigning a serving cell for each user can be viewed as assigning a item to a knapsack;

(I2). We assume only one data rate included in \( S_c \) for a content \( c \in C \). Hence, each user has maximum only one feasible data rate to select from \( R_u \). This indicates that if a user is served, he/she is served exactly with the sole selectable data rate of the desired content;

(I3). Each user consumes a certain amount of resource. The amount of resource required by cell \( h \) in serving a user \( u \) with the rate \( r \) is \( w_{u,h,r} \). Mathematically, given (I1) and (I2), \( w_{u,h,r} \) can be mapped to \( a_{ij} \) by setting the cell \( h=\text{knapsack} i \), the user \( u=\text{item} j \), and \( r \) is excluded. Note that the GAP problem assumes that \( a_{ij} \) is a known input while ours is unknown as a function of number of assigned users in a content (i.e., \( M_{gu,h,r} \)), which makes the problem even more difficult to solve;

(I4). Each user \( u \) generates a certain profit \( p_{u,h,r} \) if assigned to a cell \( h \), with the rate \( r \). Similar as the mapping in (I3), \( p_{u,h,r} \) can be mapped to \( v_{ij} \). Furthermore, \( p_{u,h,r} \) is unknown as a function of \( M_{gu,h,r} \), while \( v_{ij} \) is assumed to be a known input in the GAP problem;
(I5). The total profit of user assignment is maximized. This corresponds to (C1);

(I6). The capacity constraint of each cell $m_i$ corresponds to $b_i$ and needs to be respected. This corresponds to (C2);

(I7). Each user is assigned to exactly one cell. This corresponds to (C3).

With the above mapping, the MUA-HWN problem can be mathematically reduced to the GAP, which is NP-hard. Hence the MUA-HWN problem is also NP-hard. Due to the NP-hardness of the MUA-HWN problem, the time required to find its optimal solution increases exponentially with the size of the problem. This prohibitive processing complexity is intolerable in practical systems and for real time applications. Therefore, a computationally efficient heuristic algorithm rather than the complete algorithm is suitable for the MUA-HWN problem that needs to be dealt with in real-time. In addition, it is worth noting that the MUA-HWN is not exactly equivalent to the GAP. Instead, it is a variation of GAP mainly because of the unknown resource consumption of (I3). As a consequence, using the existing heuristic algorithms for GAP may not solve the MUA-HWN problem. We therefore design a new algorithm for solving this problem.

### 4.3 Algorithm Operation

#### 4.3.1 General Aspects

The general flowchart of the algorithm for the MUA-HWN problem is illustrated in Figure 4-4.

![Figure 4-4 Flowchart of algorithm for MUA-HWN problem](image-url)
Initially, suppose there is a set of batch queues, each of which corresponds to a particular multicast content. In each of the batch queues, there are a certain number of users requiring the content and awaiting the resource assignment in terms of serving cell and transmission rate. In the next step, these queues are sorted into descending order depending on certain priority criteria. The higher the order represents, the higher priority to be considered for the resource allocation. The queue with the highest priority will be chosen for the user assignment operation, during which the best serving cell and transmission rate are chosen for each user in the queue, according to a certain optimisation criteria. If there is no enough resource to satisfy the bandwidth requirement of the desired content in the network, users' requests are lost from the system.

In total there are two iterations, namely iteration 1 and iteration 2 in the algorithm for MUA-HWN problem. The iteration 1 is responsible for doing the assignment for each user in a batching queue. The iteration 2 is responsible for accommodating every content represented by a particular queue into the networks in a certain order. The main parts in the algorithm operation flowchart are the queue prioritisation and user assignment. Three heuristic algorithms are investigated to use different criteria in both of the queue prioritisation and user assignment, in order to maximize the system profit.

4.3.2 User-Centric Heuristic Algorithm

The assignment principle in this algorithm is user centric (i.e., based on parameters detected by the users device without taking into account network related information such as a cost and bandwidth), as used in conventional user assignment algorithm for unicast service [30][31][33][34]. The basic idea is to allow individual users to select the access network according to the perceived quality of received service, preference, price etc. Therefore, by applying the user-centric principle, the UC heuristic works as follows: For all users in the system, select the network where the user can obtain the highest utility. If there is no a network that can give the positive utility, this user is lost from the system. If multiple networks give the same utility, a network is chosen randomly.

4.3.3 Maximum-Profit Heuristic Algorithm

As mentioned before, in our algorithm the network-centric approach is considered to deliver services to users using the best available networks. It operates by taking inputs from both the user terminals and the networks to decide the best network and transmission rates to use for a given service. Hence, by obtaining enough information, the network operator or service provider is entitled to perform the user assignment algorithm. This heuristic algorithm takes the following steps:
Step 1: For each unassigned content $c$, calculate its profit $p_{\text{max}}^c$ that is the aggregated user profits when assigned to their best serving cells with the most appropriate transmission rates. In addition to $p_{\text{max}}^c$, the corresponding aggregate resource consumption $a_i$ is also calculated. The concept of aggregate resource consumption is proposed by Toyoda who attempted to apply a greedy method to the multi-dimension knapsack problem (MDKP) that is a generalization of the classical 0-1 KP for multiple resource constraints (or dimensions) [63]. In the MP heuristic algorithm, each of multicast cells can be viewed as a resource dimension. More details of the aggregate resource consumption and the mapping method to the MP heuristic algorithms can be found in Appendix B. As outputs of the calculation in this step, in addition to $p_{\text{max}}^c$ and $a_i$, the corresponding service cell and transmission rate, which gives the $p_{\text{max}}^c$ is also generated for each user. The inputs to this calculation include content profile, user profile, and network status, as shown in Figure 4-2. According to these results, the potential $x_{u,h,r}$ and $y_{c,h,r}$ for every user can be calculated;

Step 2: Among all the unassigned contents, the one that generates the largest nonzero $p_{\text{max}}^c/a_i$ will be selected and distributed to subscribed users throughout selected cells with the bit rate chosen according to the results of $x_{u,h,r}$ and $y_{c,h,r}$ obtained in step 1. Consequently, the $x_{u,h,r}$ and $y_{c,h,r}$ of users interested in the selected content constitute a part of the outputs of the maximum-utility heuristic algorithm;

Step 3: Once the content is assigned, the unallocated resource of each chosen transmission cell is updated. Then initialise the $x_{u,h,r}$ and $y_{c,h,r}$ of remaining users and repeat step 1 to step 3 until all the contents have been processed.

With reference to the Step 1 of the maximum-utility algorithm, a heuristic algorithm is used to calculate $p_{\text{max}}^c$ of a particular content and takes the following iteration for user assignment:

For each unassigned user $u$ subscribed to the content $c$,

1. Determine the potential serving cell $h$, transmission rate $r$, and thereby $x_{u,h,r}$ according to i) user location, terminal capability; ii) the maximum user profit; iii) the minimum resource consumption, from all the feasible cells and rates.

2. Update user profit and resource cost of all users in the content.

As explained in Section 4.2.2.2, $x_{u,h,r}$ is an unknown quantity and hence the user resource consumption and profit are both unknown until all users are assigned to appropriate cells. Therefore, we use the following equation to predict the resource consumption before assigning the
user. The amount of resource a user $u$ consumes when assigned to a cell $h$ with rate $r$ of content $c$ is:

$$\text{ResourceConsumption}(u, h, r, c) = \begin{cases} 0, & \text{if } M \geq 1 \ i.e., \ a \ user \ has \ been \ assigned \\ r/M', & \text{otherwise} \end{cases}$$ (4-21)

Where, $M'$ is the number of unassigned users who can be serviced by the content $c$ in cell $h$ with rate $r$. This equation indicates that due to the resource sharing among users in a multicast channel in a cell, the more users that are in this cell, the less resource is expected to be consumed by an individual user. Moreover, no extra resource is required for a new user if a user has been assigned to this cell during previous steps of user assignment.

The details of the first step in the iteration are presented as follows.

For each request
- Determine the set of potential cell
- Update the set of potential cell
- Do not assign the user without any feasible cells
- Assign user with only one feasible cell
- Assign single terminal user with more than one feasible cells
- Assign multi-terminal user with more than one feasible cells

Figure 4-5 Heuristic of user assignment

In general, the main procedure of user assignment can be summarized as two subprocedures:

a). Identify a set of potential network cells for every user interested in the multicast content, which is fulfilled by the first three steps in the Figure 4-5;

b). Assign the user to the appropriate network in the set of potential access networks along with the appropriate transmission rate, which is fulfilled by the last four steps in the Figure 4-5.

The first subprocedure is achieved by two steps. Firstly, a set of initial potential serving cells in each access networks is determined for every user. This initial potential cell is actually user's
residing cell covering the user, subject to the user's position, the cell coverage and the terminal capability. This kind of cell can be found from the set \( O_u \) of a user.

Secondly, the set of initial potential serving cells is updated by checking whether these cells have enough resources to support the content delivery. Meanwhile for each cell, determine the transmission rate from the set \( R_u \), which gives the maximum user profit while subject to constraints. This is determined by terminal capacity, user utility function, content quality and network status. The infeasible cell that does not generate positive user profit is deleted from the set of initial potential cells and the set is updated accordingly.

As a result, there are three possibilities in the type of the set of feasible cells, namely, empty set (ES), one-element set (OE) and multi-element set (ME). These sets correspond to three cases in the serving cell availability, namely, no serving cell available, a single serving cell available and multiple serving cells available respectively. Each type of set is treated differently in the second subprocedure that is further explained below.

Once the set of feasible serving cells has been made for all users, the second subprocedure is carried out to determine the final serving cell and the transmission rate for every user. Firstly, the user with an ES is dealt with. Since there is no feasible serving cell, this user is marked as a Blocked-User (BU) and the corresponding \( x_{u,h,r} \) is set to 0. Next, the user with OE is dealt with. Since no alternative serving cells exist for the user with OE, this user is assigned accordingly to the only available cell and served with the rate giving the highest user profit. The user is marked as an Admitted-User (AU) and the corresponding \( x_{u,h,r} \) is set to 1.

The user with ME cannot be dealt with in such a straightforward manner. This kind of user is marked as Undetermined-User (UU) and is left for the future process along with the other UU users. The idea of dealing with the UU user is to assign the user to the cell where the user profit is the largest and serve the user with the corresponding rate.

If there are more than two feasible cells all of which give the same profit, the cell where the user consumes the least radio resources is assigned to the user with corresponding rate. In the case of an equal resource consumption given by multiple cells, a cell is randomly chosen. This user is then marked as AU and the corresponding \( x_{u,h,r} \) is set to 1.

After all users in the content have been processed, the improvement of the solution is carried out, by reassigning the user in the cell where the non-positive content profit is generated. As a result, the user is reassigned to the cell that gives the highest user profit or lowest user resource consumption. This improvement is iterated until the user is assigned or blocked. A user is blocked if there is no feasible cell available where the positive content profit is generated.
As a consequence, we can calculate $y_{c,h,r}$ and the content profit $p^r$, for a particular content, based on $x_{u,h,r}$ that we have calculated for all users in a multicast group interested in that content.

4.3.4 Random Heuristic

The random heuristic is regarded as a baseline algorithm to compare with the maximum-profit heuristic algorithm.

The random heuristic algorithm takes the following steps:

Step 1: From all unassigned contents, randomly select one at a time;

Step 2: For all users subscribed to this content, randomly select one access network from all valid access networks and try to serve the user with the rate that will maximize the objective function, i.e., give the highest profit. If the user cannot be served because of a constraint violation, for instance, the available resource is not sufficient to generate positive profit, another access network is tried, selected randomly from the remaining access networks. This reselection procedure repeats until the user is served or blocked after all available access networks have been tried. Following this, the content is distributed into those cells where at least one subscribed user has been assigned;

Step 3: Repeat steps 1 and 2 until all the contents have been processed.

The random heuristic algorithm may be viewed as the solution obtained by trial-and-error or an ad hoc selection approach.

4.4 Performance Evaluation

This section will present a set of simulation used to evaluate the performance of proposed algorithms for MUA-HWN problem. The performances are evaluated by considering scenarios, with and without resource cost charged by operators. Moreover, a sufficient resource scenario, where all users' requirements can be satisfied, and an insufficient resource scenario, where not all users' requirements can be satisfied, are considered. The details of the simulator can be found in Appendix A.3.

When the resource cost is not considered in this scenario, the system profit is equal in fact to the system utility, i.e., the aggregated utility of served users. Figure 4-6 and Figure 4-7 show the system profit and user loss probability achieved under two ND cases depicted in Table 4 in Appendix A.3 with 100% multimode users. Basically, in the ND case 1, there are one cellular network and one broadcasting network. While in the ND case 2, the number of networks increases to five for both cellular and broadcasting networks. In the ND case 1, the number of networks is
so small that network resources are not sufficient to accommodate all user requests. In contrast, the ND case 2 considers more networks, therefore there are more sufficient resources to accommodate almost all the user requests.

Figure 4-6 System profit vs. number of contents with 100% multimode users

Figure 4-7 User loss probability vs. number of contents with 100% multimode users
The results show that in both ND cases, the MP heuristic algorithm improves the system performance in increasing the system profit and decreasing the user loss probability, in comparison to other algorithms. This can be explained by the following two reasons. Firstly, the MP heuristic employs an efficient profit-based priority mechanism to decide the sequence of assigning contents into networks. This strategy can avoid less profitable contents using up resources that could accommodate more highly profitable contents. On the contrary, the random heuristic algorithm assigns the content to the system randomly, which leads to the inefficient resource usage. And the UC algorithm does not consider the priority of contents. Instead, the assignment is performed only at user side. Secondly, for each content, based on the user location and network resource cost information etc., the MP heuristic algorithm properly chooses the most profitable access networks in a cost-effective way for individual users subscribed to this content. This can directly lead to the increase in system profit. However, the random heuristics selects the access networks for users randomly regardless of the resource consumption. Although the UC algorithm considers the achievable user utility when doing the assignment, the consumed resources are not considered. Hence, some users might be assigned to networks where a high amount of resources is consumed, which will consequently result in lower system profit.

In addition, as the number of available networks increases (from ND_Case 1 to ND_Case 2), the system profit and the user loss probability are improved for all studied algorithms. This is because the increased available networks give more available resources to accommodate contents so that fewer users are blocked. Moreover, it gives more choices to select efficient access networks. This indicates that the more available operators there are, the more the profits can be achieved by the operators. In addition, it can be observed that the Random algorithm outperforms the UC algorithm, especially when there are many underlying networks available (i.e., under the ND case 2). This is because that the UC algorithm allows users to select the network according to their preference without considering the consumed resources. In the simulation the network that gives the higher utility is given higher priority for users to choose. In the case of equal utility from all available networks, a network is randomly selected by a single user. Therefore, it is quite likely that in the end, users requiring the same content are dispersed over different networks, which causes a lot of resource consumption. However, in the Random algorithm, users sharing the same interests in a content are attempted to be assigned into one network so that these users can be served in one multicast transmission, which may not generate high utility for a single user, but less resources are consumed than the UC algorithm. When the available networks increase, it becomes more likely that users in one multicast content choose different networks in UC algorithm. Therefore, more improvement is achieved by the other two network-centric algorithms (i.e., MP and Random), compared to the UC algorithm.
As the number of content goes up, more user requests are generated in the system, hence, higher profit is generated, especially by the network-centric algorithms, such as MP and Random. However, the profit of UC algorithm grows quite slowly. This is because as the number of contents increases, more resources are required to accommodate all the contents. However, the MP algorithm can assign the right content to networks. Moreover, it also can assign the content to right cells of networks. Although the Random algorithm assigns contents without bias, users in one content are tried to be served in one transmission. Hence, resources can be efficiently utilized.

The efficiency of the Random algorithm is obvious, compared to the UC algorithm, when there are many networks available. But, when the number of networks reduce (such as in the case of ND case 1), the Random algorithm gives almost constant profit in relation to the number of contents. This is because it selects the contents without considering the introduced profit. Therefore, the profit cannot increase when the available resources are scarce. For the user-centric algorithm, it exhausts the limited resources quickly. In addition, the more requests that are in the system, the more resources are consumed. Therefore, it can be seen that in both ND cases, the profit generated by the UC algorithm keeps almost constant in relation to the number of contents. However, the user loss probability significantly increases when the number of contents increases.

The system profit improvement of the MP heuristic algorithm in comparison to the random and UC algorithms is listed in Figure 4-8.

![Figure 4-8 system profit improvement of the MP algorithm over the alternative algorithms](image)

It can be seen that the system profit improvement of MP algorithm over both Random and UC algorithms increases as the number of contents increases. With reference to the Random algorithm, more improvement is achieved in the scenario of limited resources (i.e., in the ND case...
1) than the scenario of sufficient resources (i.e., in the ND case 2). But for the UC algorithm, when the number of contents is small, more improvement is achieved in ND case 1. When the number of contents increases, the MP algorithm increases the profit, more significantly in the ND case 2 than in the ND case 1 where the available resources are not sufficient. Hence, more improvement is achieved. In addition, we can see that the Random algorithm outperforms the UC algorithm in terms of the system profit, because the MP algorithm obtains more improvement with respect to the UC algorithm compared to the Random algorithm. Moreover, in the ND case 2, the difference of the improvement is more obvious than in the ND case 1, especially, when the number of contents increases. This indicates that the UC algorithm becomes more unfavourable when there are more contents and networks available in the system to select.

Further simulations were run to evaluate the effectiveness of the algorithms with respect to the probability distribution of user terminal type (i.e., three TH cases in Table 5 in Appendix A.3). For brevity, we study 20 contents in the simulations to evaluate the profit and user loss probability under two ND cases.

![Figure 4-9 System profit under three TH cases and two ND cases when the number of contents = 20](image-url)

Figure 4-9 System profit under three TH cases and two ND cases when the number of contents = 20
Chapter 4. Multicast User Assignment in Heterogeneous Wireless Networks

Figure 4-10 User loss probability under three TH cases and two ND cases when the number of contents = 20

Figure 4-9 and Figure 4-10 further prove the effectiveness of the network-centric assignment mechanism compared to the traditional user-centric assignment mechanism. It can be seen that under all the TH cases, the UC algorithm obtains the lowest overall system profit and causes the highest user loss probability. Moreover, the benefits of MP algorithm over the Random algorithm are illustrated by the improvement of the performance. We also can see that as the probability of multimode terminal decreases (i.e. from TH_Case 1 to TH_Case 3), the profit reduces and the user loss probability increases. This is because, as the probability of multimode terminal decreases, networks are limited to be used for more users that match with their terminal types. This results in a reduction in the resource utilization, thereby reducing the profit and increasing the user loss probability. We also can see that when only two networks are available (i.e., ND_Case 1), the performance of Random algorithm approaches that of UC algorithm when the number of multimode terminal decreases. If there are no multimode terminals in the system (i.e., TH_Case 3), these two algorithms perform the same. This is easy to understand because in such a scenario as described, users have no alternative network to choose except of the network that matches their terminal capability. However, the MP algorithm is able to serve the most profitable contents, therefore it outperforms the other two algorithms in terms of the system profit and user loss probability.

Further simulations were run to evaluate the impact of the number of requests on the performance of algorithms. For brevity, we study 20 contents under two ND cases with 100% multimode users in the simulations to evaluate the profit and user loss probability achieved by the algorithms.
Chapter 4. Multicast User Assignment in Heterogeneous Wireless Networks

Figure 4-11 System profit vs. number of requests per content per cell with 100% multimode users under two ND cases when the number of contents=20

Figure 4-12 User loss probability vs. number of requests per content per cell with 100% multimode users under two ND cases when the number of contents=20

Figure 4-11 and Figure 4-12 further demonstrate that the MP heuristic algorithm outperforms the Random and UC algorithms in terms of the overall system profit and user loss probability. In addition, the network-centric scheme (i.e., MP and Random algorithms) outperforms the user-centric scheme (i.e., UC algorithm). Moreover, when the number of available networks increases from ND case 1 to ND case 2, all the algorithm obtain better performances.
As the average number of requests of a content in a cell increases, the user loss probability remains almost unchanged, while larger system profit is generated, and more improvement of the system profit is achieved by the network-centric scheme compared to the user-centric scheme. This is because no matter how many users require a content, these users are grouped and served in one multicast radio channel in a cell. Therefore, the consumed radio resources stay constant in relation to the number of requests of a content in a cell. But, the fact that more users are aggregated and served accounts for the higher profit. However, it can be seen that the UC algorithm obtains higher user loss probability under ND case 2 when the request arrival rate increases. This is because the more users that require the same content, the more likely the users choose different networks by UC algorithm due to the user heterogeneity. Hence, the increased resource consumption leads to increased user loss probability. But in the ND case 1, users have any two choices of access networks, hence, the user loss probability stays the same in relation to the average number of requests.

Further simulations were run to evaluate the effectiveness of the algorithms in the case of non-zero resource cost. In the simulations, the ratio of the maximum resource cost of broadcasting network $C_b$ over the resource cost of cellular network $C_c$ varies from 1 to 1000 in order to take into account a wide range of practical situations. In addition to the system profit and user loss probability, we observe the impact on resulting chosen networks by different algorithms as the resource cost ratio $c_b/c_c$ varies. Two scenarios are considered. In the first scenario, the maximum available bandwidth in both cellular and broadcasting networks are sufficient to accommodate all traffic (e.g., 2560 kbps per cell for 20 multicast contents with the highest data rates of 128 kbps), in order to maximize the impact of changing the resource cost ratio. In the second scenario, the cellular network has the maximum capacity of 1024 kbps per cell and the broadcasting network has the maximum capacity of 2048 kbps per cell. Hence, each network can not accommodate all the multicast contents on its own, with the maximum transmission rate. The system profit and user loss probability are measured and presented in Figure 4-13 and Figure 4-14 respectively. In addition, Figure 4-15 presents the percentage of served users in cellular and broadcasting networks over the total users in the system, denoted by $s_c$ and $s_b$ respectively. The resource cost of cellular network is set to 1.
Figure 4-13 Overall system profit vs. the ratio $c_p/c_c$ with 100% multimode users in the case of sufficient and insufficient available resources.

Figure 4-14 Overall loss probability vs. the ratio $c_p/c_c$ with 100% multimode users in the case of sufficient and insufficient available resources.
Figure 4-15 Percentage of served users in cellular and broadcast networks vs. the ratio $\frac{c_b}{c_c}$ with 100% multimode users in the case of sufficient available resources.

Figure 4-16 Percentage of served users in cellular and broadcast networks vs. the ratio $\frac{c_b}{c_c}$ with 100% multimode users in the case of insufficient available resources.

The Figure 4-13 and Figure 4-14 demonstrate the effectiveness of the network-centric and MP algorithms under the varying cost situation, in both sufficient in insufficient resource scenarios.
can be seen that the MP algorithm outperforms the other two algorithms in terms of the system profit and the user loss probability, while the traditional UA algorithms gives the worst performance. Clearly, in the sufficient resource scenario, higher system profit and lower user loss probability are achieved compared to the insufficient resource scenario.

As the cost ratio increases, the cost in broadcasting network goes up. When there are sufficient resources in the system, both networks are able to provide the same utility to users. In the UC algorithm, users choose the networks randomly. As shown in Figure 4-15, the percentage of served users in a cellular network is always 50%. However, because of the increased cost in the broadcasting network, the users who select the broadcasting network will be blocked by the network due to the negative profit. Therefore, as the cost ratio increases, the UC algorithm generates lower profit and causes higher user loss probability due to the increased resource cost of broadcast networks rather than the shortage of resources. Consequently, the number of served users in the broadcasting network decreases as well. When there are not sufficient resources available in the system, the percentage of served users in the cellular network is less than 50% as shown in Figure 4-16. The percentage of served users in the broadcast network initially increases, because as the resource cost in the broadcast network increases, more popular content is required to compensate the resource cost. Since the available resources are limited, larger user groups instead of small user groups are more likely to be accommodated in the system. However, when the resource cost of the broadcast network continues to increase, fewer contents are popular enough to compensate the cost. Hence, the served users in the broadcast network reduce and the total user loss probability reduces as well. Accordingly, as presented in Figure 4-14, the user loss probability of the UC scheme firstly reduces as the resource cost of the broadcast network increases. Then the user loss probability increases.

For the Random algorithm, it selects the network randomly for accommodating users regardless of the resource cost and tries to accommodate users without taking into consideration of the access networks. In a sufficient resource scenario, when the resource cost of the broadcasting network is small, the Random algorithm assigns users to either of the networks without bias, so the traffic is distributed through cellular and broadcast networks equally, as shown in Figure 4-15. In addition, the profit achieved reduces as the resource cost gets larger. When the resource cost of the broadcasting network increases and reaches a certain number, i.e., 50 in Figure 4-13, not all the users who are originally assigned to the broadcasting network can generate positive profit, and thereby these users are switched to the cellular network. As a result, more users are served in the cellular network and higher profit is generated as the resource cost of the broadcasting network continues to go up. In the insufficient resource scenario, when the resource cost of the broadcast network is small, more users are assigned to the broadcast network than the cellular network, as shown in Figure 4-16. This is because the broadcast network has more resources available to
Chapter 4. Multicast User Assignment in Heterogeneous Wireless Networks

accommodate users. When the resource cost of the broadcast network increases, since the limited available resources can not accommodate all traffic, more popular contents are assigned to the broadcast network, and less popular contents are assigned to the cellular network. Consequently, the percentage of the served users in the broadcast network increases, while the percentage of the served users in the cellular network reduces, as the resource cost increases. When the resource cost continues to increase, fewer contents are popular enough to be served in the broadcast network, compensating the increasing resource cost. Therefore, more users are switched to the cellular network. Because of the capacity constraints in the cellular network, the overall user loss probability increases and the overall system profit reduces.

Different to the VC and Random algorithms, the MP algorithm assigns users adaptive to the resource cost of constituent networks. The network with the lower resource cost is more likely to be chosen to accommodate the user. In both Figure 4-15 and Figure 4-16, it shows that the MP algorithm assigns more users into broadcast network when the resource cost in broadcast network is low. This is because that the MP algorithm assigns user to the cell where the user profit is the highest. As explained before, the user profit is calculated by \( u(a) - \frac{c \cdot a}{M} \), where \( u(a) \) is the user utility served with rate \( a \); \( c \) the resource cost in the cell where the user is assigned and \( M \) the total number of users assigned in the cell. It is easy to derive that in the case of the same rate, when \( c_b \leq \frac{M_b \cdot c_c}{M_c} \) (where \( b \) denotes the broadcast network and \( c \) denotes the cellular network), the user is assigned to the broadcast network, otherwise, it is assigned to the cellular network. Therefore, \( \frac{M_b \cdot c_c}{M_c} \) is a threshold of \( c_b \), beyond which, more users are assigned to a cellular network; otherwise, more users are assigned to a broadcast network. Because \( M_b > M_c \) due to the larger geometrical area of broadcast networks than cellular networks, the threshold of \( c_b \) is larger than \( c_c \). As a result, it can be seen in the sufficient resource scenario that before the resource cost of the broadcasting network increases and reaches a certain value (i.e., 10 in Figure 4-15), more users are assigned into the broadcast network. After that, more users are assigned to the cellular network. In the insufficient resource scenario, we also can see a threshold of \( c_b \) (i.e., 100 in Figure 4-16). The threshold in the insufficient resource scenario is larger than in the sufficient resource scenario is because the resource constraint in the cellular network is more stringent than the broadcast network. Since the MP algorithm is able to assign users adaptive to the varying resource cost, higher profit is expected to be generated compared to other algorithms that do not consider the resource cost when making the decision, in both sufficient and insufficient resource scenarios. Moreover, we can see that in the sufficient resource scenario, the MP algorithm achieves a more stable system profit.
In conclusion, in this chapter, we presented a systematic study for the important problem of MNS-HWN (i.e., multicast user assignment in heterogeneous wireless networks). The solution to this problem attempts to optimize the system profit by proper selection of access networks and transmission rates serving multiple multicast contents to users in different multicast groups simultaneously. The system profit is calculated as a difference between the total user utility and resource cost of serving these users. We employed a generic utility function for each user, which allows flexible extension of our study to other user environment. The problem was formulated as a combinatorial optimization problem and was proved NP-hard. To our best knowledge, this was the first to formulate and analyze the MNS-HWN problem, and hence becoming the basis for future research in this problem. Moreover, we derived efficient network-centric algorithms to solve the problem, namely Random heuristic algorithm and MP heuristic algorithm. These algorithms make up parts of the overall MRM algorithms.

We evaluated the performance of proposed algorithms in comparison to traditional ones, by considering scenarios, with and without resource costs charged by operators. Moreover, a sufficient resource scenario and an insufficient resource scenario were considered. Simulation results revealed the superior performance of the network-centric algorithm over the traditional user-centric algorithm in increasing the overall system profit and reducing the overall user loss probability. Moreover, it was found that the MP algorithm outperforms the Random algorithm. Furthermore, the MP algorithm performed consistently well under a wide variety of traffic loads, number of available networks, resource availability, resource cost and user terminal type distributions. These results demonstrated the significance and value of using the MP algorithm for service providers or network operators to optimize their profits in the provisioning of multicast contents in the future generation mobile systems.

This work will be furthered in the Chapter 6 by investigating the methods and interfaces that unites the MDS and MUA algorithms in a common algorithm framework. This combination makes up the entire MRM algorithms. The overall performance of the MRM algorithms will be investigated as well.

Before we evaluate the methods of combining the MDS and MUA algorithms, the MDS algorithms will now be studied first in the next chapter.
Chapter 5

5 Batching for Multimedia-on-Demand Scheduling in Wireless Networks

This chapter aims to discuss, evaluate and analyse the algorithms that have been developed for the MDS (i.e., multimedia-on-demand scheduling) problem in wireless networks. Firstly, the system model considered in MDS is presented. Two system models are studied, namely the blocked-calls cleared system and the blocked-calls delayed system. Then, a general flowchart of the algorithm's operation is presented for each system model respectively. Next, a systematic study on the MDS scheme in blocked-calls cleared systems is given by the means of both theoretical analysis and simulation. More results regarding the MDS schemes in blocked-calls delayed systems can be found in Chapter 6.

5.1 System Model

The techniques of traffic analysis can be divided into two general categories: loss systems and delay systems. The appropriate analysis category for a particular system depends on the system's treatment of overload traffic. In a loss system overload traffic is rejected without being serviced. In a delay system overload traffic is held in a queue until the facilities become available to service it. Therefore, the loss system and the delay system are also called block-calls cleared system and block-calls delayed system respectively. Conventional circuit switching operates as a loss system since normally excess traffic is blocked and not serviced without a retry on the part of the user. The message or packet switching obviously possesses the basic characteristics of a delay system. The basic measure of performance for a loss system is the blocking probability. When considering the reneging of users waiting for the establishment of a multicast channel, the reneging probability will also be considered. We call the combination of the reneging probability and the blocking probability the loss probability, where a "lost" user is a person who is rejected or has reneged from the batch. A delay system, on the other hand, is measured in terms of service delays, which is the time period from the time the user enters the system until the time the content is displayed.

The current research regarding MDS in the literature studies both the loss system and the delay system. The system model used by these two systems is illustrated in Figure 5-1.
The system provides a set of multimedia contents (e.g., movies, news, music, etc). Each of these contents may have a different popularity and duration. Let $\alpha_i$ be the probability that a request chooses content $i$, and let $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_M$ without loss of generality. Note that users' requests are generated from a service area comprising a single cell or a number of cells. The arrivals of users' requests are assumed to be Poisson process with a rate $\lambda$. And each arrival of Poisson ($\lambda$) can select one of the $M$ contents. Therefore, the Poisson ($\lambda$) can be decomposed into $M$ independent Poisson processes: Poisson ($\lambda \alpha_1$), Poisson ($\lambda \alpha_2$) ... and Poisson ($\lambda \alpha_M$), each modelling the arrival of requests for content $i$. When the system receives a first user request for a particular content, it waits for a batching period to collect more requests for the same content. All the user requests for the same content are held in a particular batching queue. After batching, the system sends a resource request for serving the content, competing for a limited multicast resource pool with other resource requests. There is no restriction on the positioning of available resources. If there is a free multicast channel available, the batch of users that sends the resource request is served simultaneously. Otherwise, these users are dealt with in accordance with the system characteristic. In a blocked-calls cleared system, the batch of users is lost from the system without being served. Whereas, in a blocked-calls delayed system, these user requests are still held in the original batching queue, until the free multicast channel becomes available to serve them, or the user quits from the waiting queue (i.e., reneging) due to the long time of waiting. More details regarding the MDS in different systems will be described as follows.

5.2 Multicast-on-Demand Scheduling in Loss Systems

Four MDS schemes have been proposed to be applied to loss systems [55][59], namely two basic batching schemes (timeout-based and size-based), and two hybrid schemes (e.g., combined-for-
profit and combined-for-loss). The general procedure involved in these schemes is outlined in Figure 5-2.

![Figure 5-2 Multicast-on demand service scheduling flow chart in a loss system](image)

In general, in a loss system, the multicast service scheduling involves two steps, namely batching and resource allocation. User requests for the same content are firstly batched for a batching duration. Note that users' requests are batched depending on the desired content, not on the particular cell where requests come from. The period of the batching could be decided by the period of user's waiting time in the batch (timeout-based scheme), the number of users' requests in the batch (sized-based scheme), or by both waiting of time and number of requests (hybrid schemes). After the batching threshold is reached, the system will decide if there are enough resources available to serve the desired content to batched users, given the content QoS requirement and the current resource availability in the cell. In a loss system, multicast channels are used according to the FCFS discipline with respect to the resource request until all channels are occupied. At one time, there is only one resource request from a batch requiring the resource allocation. The resource allocation here has the same objective as the traditional call admission control in cellular networks. The traditional call admission control is a provisioning strategy used to limit the number of call connections into the networks in order to alleviate the network congestion and provide the desired QoS to users in service. In the end, the batching queue that has performed the resource allocation is cleaned up, no matter whether all the users' requests in this queue can be served or not. The methods of identifying batching duration in different schemes are briefly described below.

**Timeout-based**: batching duration is identified by the period of the first user's waiting time \( W \) referred to the batching time in order to guarantee the maximum service delay. Requests are batched until the batching time is reached. The timer is set when a user agrees to wait \( W \) minutes and the subsequent users for the same content join the waiting line.

**Size-based**: batching duration is identified by the number of users \( N \) referred to the batching size in order to guarantee the number of users a batch serves. Requests are batched until the batching size threshold is reached. Hence, the user's waiting time depends on the frequency rate of request arrivals.
Combined-for-profit / loss: batching duration is identified by both of \( W \) and \( N \). In the combined-for-profit scheme, requests are batched until both \( N \) number of user requests is collected and the first user in the batch waits no less than \( W \) time units. This scheme aims to maximize the "profit" (e.g., the number of served users). In the combined-for-loss scheme, requests are batched until either \( N \) number of users is collected or the first user in the batch has waited \( W \) time units. This scheme aims to minimize the service delay, and thus the reneging probability.

The four batching schemes described above have been discussed and evaluated in [55][59]. However, they consider infinite network resource and study only one content, such as in [55]. Or, the user's reneging behaviour is not considered, such as in [59]. Obviously, it is not suitable for the resource-limited wireless environment, or for practical implementation. We have enhanced their study by considering the blocking probability caused by the limited wireless resource, and furthermore we consider multiple contents and the behaviour of user reneging. Based on these studies, an analysis model is derived to study the performance of the timeout-based batching scheme. Moreover, we propose a novel optimal timeout-based batching scheme to maximize the ratio of served users to the total users in the system, adaptive to the available resources and traffic loads in the system.

### 5.3 Multicast-on-Demand Scheduling in Delay Systems

There has been a lot of research work focused on MDS algorithm development in delayed systems, such as FCFS [60], MQL [60], and the variation of MQL, i.e., BMQ [57]. It has been proved in [57] that the BMQ algorithm outperforms the FCFS and MQL in reducing the user loss probability, and thus improving the system throughput. In the above schemes, the explicit delay preference of a user is regarded as unknown, in particular the FCFS and MQL algorithms. The BMQ algorithm is based on partial user delay information (such as knowing the minimum waiting threshold of all the users). Differing from them by considering full knowledge of user delay behaviour, Chan in [56] assumes that such delay preference is able to be revealed to the server upon users' arrival. Based on this assumption, they proposed a DAB (Delay-Aware Broadcasting) scheme with the objective of minimizing the user loss rate. In this scheme, a user reveals his maximum delay tolerance to the network along with the content request. The network keeps a sorted list of the reneging time of the users, and serves the one just about to leave, along with all the requests for the same content. If there is not enough resource at the reneging point, the user is not served and is lost. Compared to the FCFS algorithm, the DAB algorithm significantly reduces the overall user loss probability. Another MDS scheme considering the user delay tolerance is LAMB (Look Ahead Maximize-Batch) [58]. However, LAMB is quite complex and requires
continuous computationally intensive optimisation in the server. Hence, we do not consider this mechanism in the multicast service scheduling problem.

In general, in a delay system, the multicast service scheduling involves five steps, namely batching, queue identification, queue prioritisation, queue selection and resource allocation, as illustrated in Figure 5-3. Note that the steps of batching and resource allocation are the same as the counterpart steps in a loss system.

![Figure 5-3 Flow chart of MDS operation in a delay system](image)

User requests for the same content are firstly batched for a batching duration. Then, the queue identification determines the specific batching queue that is eligible for being allocated to networks. These eligible queues will be marked for future use. Therefore, at a time, there could be multiple resource requests, each of which comes from a batch requiring resource allocation. In the next step, these marked queues are sorted in a descending order depending on certain priority criteria. The higher order represents the higher priority to be considered for the resource allocation. The queue with the highest propriety will be chosen for the resource allocation. If there is no enough resource to satisfy the bandwidth requirement of the desired content in the network, users' requests are re-batched in the queue until they renege or free resources are released. However, users' requests that have been served are deleted from the queue.

We propose two MDS algorithms for delayed systems, namely timeout-based maximum profit (TMP) algorithm and delay-aware maximum profit (DMP) algorithm. With reference to Figure 5-3, the major steps of proposed algorithms and two conventional algorithms (i.e., BMQ and DAB) are detailed below. It is worth noting that these algorithms differ from each other by applying different criteria in the steps of queue identification, queue prioritization and resource allocation.
Chapter 5. Batching for Multimedia-on-Demand Scheduling in Wireless Networks

Queue identification:

The queue identification identifies those queues that are eligible for scheduling if they have user requests with waiting time exceeding a certain batching threshold. The objective is to maximize the number of aggregate users in the chosen batch so as to maximize the system throughout. Different criteria are employed in specifying the batching threshold. For instance, in both BMQ and TMP algorithms, the batching threshold equals the minimum waiting time threshold of users in a batching queue. The waiting time threshold is a period of time for which a user is willing to stay in the queue without reneging. For both DAB and DMP algorithms, the batching threshold is associated with user's exact delay tolerance. In the DAB algorithm, a queue is eligible for scheduling if there is a user just about to leave, which can be known from his delay tolerance. However, the DMP algorithm relaxes the eligibility criteria: a queue is eligible either if it has a user just about to leave OR if the reneging has happened.

Queue prioritisation:

The queues that are identified in the step of queue identification are prioritized. The queue with higher priority is scheduled earlier. The DAB algorithm does not use any prioritization criteria, since always there is only one queue that is eligible for scheduling. However, in the BMQ algorithm, the maximum-queue-length (MQL) criterion is applied. This criterion prioritizes queues in the order of the number of waiting requests. Obviously, the objective of MQL is to maximize the system throughout by scheduling more popular contents earlier, which has been verified in [60]. But this scheme is unfair since it favours more popular contents. Apart from MQL, one of the other potential existing criteria for queue prioritization is FCFS that selects the queue with the longest waiting time to be served the first. The advantage of this algorithm is its fairness: every content is treated equally regardless of its popularity. The drawback of this strategy is the lower system throughput. Acting effectively as a compromise between FCFS and MQL, the maximum-factored-queue-length (MFQL) scheme has been proposed to exhibit the good features of each of previous schemes while avoiding the bad [61]. However, none of these prioritization schemes considers the heterogeneity of users staying in every queue, the diversity of contents and the feature of networks that are used to deliver the contents.

To cope with the heterogeneous environment in current mobile communication systems, to overcome the drawbacks in existing schemes for queue prioritisation and to ensure an effective multicast-on-demand scheduling in wireless networks, a new queue prioritisation scheme is proposed and referred to as MP (i.e., maximum-profit), which selects the queue with the highest potential profit to be served the next. The profit is associated with the user location, terminal capability, preference, and the transmission cost. We have introduced the MP algorithm in more detail in Chapter 4.
### Resource Allocation

Similar to loss systems, this step determines if the users in the queue with the highest priority can be served in the network given the current resource availability and application requirements. The current research studies the MDS problem for a multimedia server with limited number of multicast channels, such as DAB and BMQ algorithms. Users are served if there are free channels available to serve the content with required bandwidth. Otherwise, users’ requests go back to the waiting queue and wait. For the proposed DMP and BMP algorithms, a novel solution is proposed to solve the MDS problem considering the heterogeneous mobile networks. A user will be assigned to a network where the maximum profit is generated. Moreover, we use the data-adaptive technique to further improve the efficiency of resource utilization by serving users with degraded QoS when the available resources are not enough to serve users with their original QoS.

The rest of this chapter provides an insight into the MDS mechanism for the loss systems. More details about the MDS mechanism study for the delay systems can be found in Chapter 6.

### 5.4 Analysis Model

An analytical model is developed to evaluate the performance of timeout-based batching scheme. Three metrics of the performance are formulated. They are the reneging probability $P_{\text{ren}}$, blocking probability $P_{\text{blk}}$ and satisfaction ratio $P_s$, where $P_{\text{ren}}$, $P_{\text{blk}}$ and $P_s$ are, respectively, the ratio of the number of reneged users due to the delayed service set-up, blocked users due to the insufficient resources and served users, to the total number of users in the system. Obviously, $P_s = 1 - P_{\text{blk}} - P_{\text{ren}}$.

Users are able to request $M$ independent contents according to a certain probability. Let $\alpha_i$ be the probability that a request chooses content $i$, and let $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_M$ without loss of generality.

The arrivals of users’ requests are assumed to be Poisson process with a rate of $\lambda$. And each arrival of Poisson ($\lambda$) can select one of the $M$ contents. Therefore, the Poisson ($\lambda$) can be decomposed into $M$ independent Poisson processes: Poisson ($\lambda \alpha_1$), Poisson ($\lambda \alpha_2$) ... and Poisson ($\lambda \alpha_M$), each modelling the arrival of requests for content $i$.

Each user is willing to wait for a period of time $U \geq 0$ to view the desired content. If this content is not displayed by then the request reneges. $U$ follows the distribution $R(u) = P(U < u)$ with a mean $\tau$.
Let $\lambda_i$ be the throughout of content $i$, which is the number of remaining requests without reneging in unit time during batching. $\lambda_i = \frac{N_i}{T_i}$, where $N_i$ is the average batch size of content $i$ and $T_i$ is the average period between two consecutive copy of content $i$ delivery time. As explained in [55],

$$\overline{T_i} = \frac{1}{(1 - R(W))\lambda_i} + W$$  \hspace{1cm} (4-1)

$$\overline{N_i} = 1 + \lambda_i \int_0^W (1 - R(x)) \, dx$$  \hspace{1cm} (4-2)

We aggregate $M$ arrival processes into one process. The total throughout $\lambda$ is then given by $\lambda = \sum_{i=1}^M \lambda_i = \sum_{i=1}^M \frac{N_i}{T_i}$. So, the reneging probability $P_{ren}$ is given by $P_{ren} = 1 - \frac{\lambda}{\lambda}$. Based on the equation, we can derive that $P_{ren}$ has an upper limit with $\lambda$, i.e., $\lim_{\lambda \to \infty} P_{ren}(\lambda) = 1 - \int_0^W (1 - R(x)) \, dx$.

After batching, remaining requests compete for a total of $C$ multicast radio channels for delivering multicast contents. There is no restriction on the positioning of available resource. Requests are blocked if there is no channel available.

Given the total $C$ multicast channels, the loss system can be modelled as an M/G/C/C queue. Let $P_{\text{blk}}$ be the probability that the total channels are occupied, the blocking probability $P_{\text{blk}}$ is given by $P_{\text{blk}} = P_{\text{blk}} \cdot (1 - P_{\text{ren}})$. $P_{\text{blk}}$ can be calculated using Erlang B Formula [74] and is given as following.

$$P_{\text{blk}} = \frac{1}{C!} (\lambda_s d)^c \sum_{i=0}^C \frac{1}{i!} (\lambda_s d)^i$$  \hspace{1cm} (4-3)

where $\lambda_s$ is the arrival rate of total resource requests for serving desired contents, which is given by

$$\lambda_s = \sum_{i=1}^M \frac{1}{T_i}$$  \hspace{1cm} (4-4)

Based on the equation, we can derive that $P_{\text{blk}}$ has an upper limit with $\lambda$, i.e.,

$$\lim_{\lambda \to \infty} P_{\text{blk}}(\lambda) = \frac{1}{C!} \frac{(dM / W)^c}{\sum_{i=0}^C \frac{1}{i!} (dM / W)^i}$$  \hspace{1cm} (4-5)
Since $P_{ren}$ also has an upper limit, the upper limit of $P_{blk}$ and thereby $P$, exist, 
\[ \lim_{\lambda \to \infty} P_{blk} = 1 - \lim_{\lambda \to \infty} P_{blk} \cdot (1 - \lim_{\lambda \to \infty} P_{ren}) - \lim_{\lambda \to \infty} P_{ren}. \]
These results indicate that $P_{ren}$, $P_{blk}$ and $P$ will not be dependent on $\lambda$ when $\lambda$ increases to a certain value.

Three performance measures are studied, namely satisfaction ratio, reneging probability and user probability. To study the performance, the exponential user reneging function in [55] is adopted, i.e.,
\[ R(u) = \begin{cases} 
0, & \text{if } 0 \leq u \leq U_{\min} \\
1 - e^{-(u-U_{\min})/\tau}, & \text{otherwise.} 
\end{cases} \quad (4-6) \]
where $U_{\min}$ is the minimum time for which users are always willing to wait. Zipf distribution is adopted to model the probability of choosing a particular content [67]. The probability of choosing the $k$th content is given by:
\[ P_k = \frac{1}{k^{(1-\theta)} \sum_{k=1}^{M} k^{(1-\theta)}} \quad (4-7) \]
where, $\theta$ is the parameter that determines the shape of the distribution. A small $\theta$ corresponds to a more severe discrimination of requests among contents and indicates that some contents are desired more frequently than the others. As in [56], $\theta=0.271$ is assumed in our work.

Table II shows the default parameter values used in the analysis and simulation unless otherwise stated.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>1000 requests/hour</td>
</tr>
<tr>
<td>$\tau$</td>
<td>10 minutes</td>
</tr>
<tr>
<td>$d$</td>
<td>15 minutes</td>
</tr>
<tr>
<td>$U_{\min}$</td>
<td>1 minute</td>
</tr>
<tr>
<td>$C$</td>
<td>20</td>
</tr>
<tr>
<td>$M$</td>
<td>5 (e.g., news, travel info, sport, music, weather info)</td>
</tr>
</tbody>
</table>

Table 1 Default inputs of analytical and simulation models
5.5 Simulation Model

A simulator was constructed using OPNET [62] to justify the derived analytical model and evaluate the proposed optimal timeout-based batching scheme by comparing with four conventional batching schemes, namely, two traditional basic schemes (timeout-based, size-based) and two traditional hybrid schemes (combined-for-profit, combined-for-loss). The assumption in the analytical model is adopted by the simulation model unless otherwise stated. This simulator consists of a wireless cell (e.g., DVB cell) with limited available multicast radio channels. Users are uniformly distributed over the cell. Each user randomly sends a request for a particular content at one time. The duration of the content is determined based on an Exponential distribution. We consider a mixed traffic scenario where multicast-on-demand service and the other services (e.g., TV broadcasting and datacasting) share and compete for a common and limited resource pool. As a result, the network resource available for multicast-on-demand service fluctuates over time, as denoted by \( C_{\text{multicast}} \) channels.

5.6 Analytical Results

Both the results obtained from computer simulations and analytical model are shown in Figure 5-4. We observe that the performance depends highly on the selection of batching time \( W \). Thus, finding an appropriate \( W \) is a crucial issue. \( P \), first increases to reach a maximum, and then decreases rather slowly as \( W \) increases. Maximum \( P \) is achievable given certain system status, which includes service quality constraints (e.g., \( P_{\text{ren}} \) and \( P_{\text{blk}} \) constraint), traffic information (e.g., content duration, request arrival rate), user profile (e.g., reneging function), and resource availability. The \( W \) at which the maximum \( P \) is achieved is denoted as the optimal batching time \( W_{\text{op}} \). We can see a trade-off between \( P_{\text{ren}} \) and \( P_{\text{blk}} \). \( P_{\text{ren}} \) increases rather linearly when \( W \) increases as more users renege with the longer batching time, whereas \( P_{\text{blk}} \) decreases due to the more efficient resource utilization. Also we can see that \( P_{\text{ren}} \) varies more gradual than \( P_{\text{blk}} \) until \( P_{\text{blk}} \) decreases to reach 0 as \( W \) increases. This indicates that \( P_{\text{blk}} \) is a dominant factor affecting \( P \), when resource is not enough; otherwise, \( P_{\text{ren}} \) becomes dominant. The lower bound of \( W_{\text{op}} \) is determined by the constraint of the maximum \( P_{\text{blk}} \) while the upper bound of \( W_{\text{op}} \) is determined by the constraint of the maximum \( P_{\text{ren}} \). For instance, given the constraints \( P_{\text{ren}} \leq 0.1 \) and \( P_{\text{blk}} \leq 0.05 \), we can use \( W_{\text{op}} = 4 \) minutes.
Figure 5.5 shows that the available resource $C$ has a big impact on the system performance. Figure 5.5 (a) shows that as $C$ increases, $P_{blk}$ reduces as more resource is available. However, $P_{ren}$ is completely independent of $C$, which is also proven in the analytical model. Furthermore, it can be seen from Figure 5.5 (b) that $W_{op}$ can be achieved at low value of $W$, especially when $C$ is high. When $C$ increases until $P_{blk} = 0$, $C$ has no impact on the performance. Hence, $W_{op}$ can be kept as low as possible in order to minimize $P_{ren}$. As $C$ decreases, more users have to be batched in a multicast channel in order to efficiently utilize more scarce resources, and hence $W_{op}$ increases. This indicates that when resource is sufficient, a short $W$ is required in order to minimize $P_{ren}$; otherwise, a long $W$ is required in order to maximize resource utilization efficiency.
Chapter 5. Batching for Multimedia-on-Demand Scheduling in Wireless Networks

Figure 5-5 Impact of available multicast channels $C$ in a cell, on system performance

Figure 5-6 shows that $\lambda$ has not much influence on the performance, especially when $\lambda$ is high. This result justifies the analysis in the analytical model. Figure 5-6 (a) shows that as $\lambda$ increases, $P_{blk}$ and $P_{ren}$ first increases and decreases respectively, and then both change rather slowly. Therefore, a constant $W_{up}$ can be used for a wide range of $\lambda$ especially when $\lambda$ is high, which can be seen in Figure 5-6 (b). For instance, to maximize $P_s$ while maintaining the maximum $P_{ren}$ at
around 0.2 and \( P_{blk} \) at around 0.05, we can use \( W=5 \) minutes for \( \lambda \) ranging from 1000 to 10000 requests/hour. And the minimum \( \lambda \) at which the \( P \), keeps stable can be calculated by finding the minimum \( \lambda \) from the solutions of the function \( \left| \lim_{\lambda \to 0} P_\lambda - P_\lambda(\lambda) \right| = \varepsilon \ (\varepsilon = 0.01) \). As \( \lambda \) gets lower, \( P_{ren} \) becomes a dominant factor affecting \( P \), so a shorter \( W \) is required to reduce \( P_{ren} \). For instances, in Figure 5.6 (b), when \( \lambda \) varies from 1000 to 100 request/hour, we reduce \( W_{op} \) from 5 to 2 minutes. This result indicates that in the case of a high \( \lambda \), it is not necessary to adapt \( W \) to \( \lambda \). However, when \( \lambda \) is low, a shorter \( W_{op} \) is required in order to minimize \( P_{ren} \).
5.7 Optimal Timeout-Based Scheme

Based on our analytical model, \( W_{op} \) can be calculated given a certain system status including the service quality, traffic information, user reneging function and resource availability. Therefore, we can come up with a mapping table from a given system status to \( W_{op} \), such as in Table 2. With the mapping table, the network or service providers can keep track of the system current status, in particular the available resources, to determine \( W_{op} \) at that moment. Users’ requests for contents are then scheduled and served based on the \( W_{op} \). Considering that \( W_{op} \) does not depend much on the request arrival rate \( \lambda \) when \( \lambda \) is high, the size of the mapping table can be decreased by mapping a range of \( \lambda \) to a particular \( W_{op} \).

In a resource-limited wireless communications system, \( P_{blk} \) is the key concern when evaluating the grade of service (GoS) of on-demand services because of the following reasons. It is more annoying for the users to wait but eventually not being served than to decide not to wait but renege from request. Moreover, effective methods to increase the users’ delay tolerance, such as decreasing the service price, are easier to be implemented than to increase the resources limits in the system [59]. Therefore, we set \( P_{blk} \) as a priority constraint to \( P_{ren} \). Consequently, in our optimal batching scheme, the \( W \) is optimized to maximize the satisfaction ratio \( P \), subject to constraints of the maximum \( P_{blk} \) (5%) and \( P_{ren} \) (30%).
5.8 Evaluation of Optimal Timeout-based Scheme

5.8.1 Algorithm Parameters

We evaluate the proposed optimal timeout-based batching scheme by simulation. In the simulation, the optimal batching time $w_{op}$ adapts to the time varying resource availability and request arrival rate. Table 2 (a) and Table 2 (b) show the mapping of number of available channels and request arrival rate to the optimal batching time $w_{op}$ respectively. As a result, if the number of available channels is low, the batching time is set as long as possible to improve resource utilization by merging enough users in a group. Otherwise, the system will shorten the batching time in order to avoid too many users reneging from the batch. In contrast, given a certain $C$, the batching time is kept constant for a wide range of $\lambda$. This makes the optimal scheme computational efficient since amount of calculation to adapt $w_{op}$ to $\lambda$ can be avoided.

<table>
<thead>
<tr>
<th>Number of available channels $C$</th>
<th>Optimal batching time $w_{op}$ (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td>35</td>
<td>2</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
</tr>
</tbody>
</table>

(a). Mapping table of $C$ to $w_{op}$ with $\lambda = 1000$ request/hour

<table>
<thead>
<tr>
<th>Request arrival rate $\lambda$ (request/hour)</th>
<th>Optimal batching time $w_{op}$ (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>400-1900</td>
<td>4</td>
</tr>
</tbody>
</table>

(b). Mapping table of $\lambda$ to $w_{op}$ with $C=20$

<table>
<thead>
<tr>
<th>Request arrival rate $\lambda$ (request/hour)</th>
<th>Optimal batching time $w_{op}$ (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>400-1900</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2 Mapping table in optimal timeout-based batching scheme

In the traditional batching schemes (i.e., the basic and hybrid schemes), the resource constraint is not considered. Their batching parameters, namely batching time $W$ and batching size $N$, are
chosen from its respective constituent batching schemes (i.e., $W$ in the basic timeout-based scheme and $N$ in the basic size-based scheme) "optimized" independently for a target arrival rate and remain constant for other rates[55]. Therefore, we can characterize these traditional schemes as static ones in the sense that their batching parameters do not adapt to the variation of actual system status, such as the resource availability.

We evaluate the traditional scheme in a scenario where the cell capacity is limited, moreover, the available resources are time-varying. The batching parameters of traditional schemes are configured based on the target user request arrival rate and the heaviest load of background traffic in the cell, as the methodology employed by the static resource allocation [75].

5.8.2 Simulation Results and Discussions

Figure 5-7 presents $P_r$, $P_{blk}$, and $P_{ren}$ as $C$ varies from 10 to 40 with $W=9$ minutes, $N=23$ in traditional schemes and $\lambda=1000$ request/hour. The parameters $W$ and $N$ are independently chosen given the target available channels $C=10$.

Figure 5-7 (a) shows that the proposed optimal timeout-based scheme gives higher $P_r$ than other schemes for investigated values of $C$, especially when $C$ is large. This is because the optimal scheme chooses the most appropriate batching time for each $C$ so as to maximize $P_r$. In contrast, in traditional schemes, the batching parameters are dimensioned for the worst case (i.e., smallest $C$). These parameters result in a longer batching duration, which is not appropriate for non-worst case $C$. In addition, it can be seen that $P_r$ mainly depends on $P_{ren}$. This is because $P_{blk}$ is so small that the $P_{ren}$ is dominant over the $P_{blk}$ to determine the $P_r$.

Figure 5-7 (b) shows that the optimal scheme improves $P_{ren}$ as $C$ gets larger by choosing shorter batching time. However, the traditional schemes assume constant batching parameters as $C$ varies such that the user's waiting time and thereby $P_{ren}$ are also constant. In other words, it is not optimized and improved based on $C$. In addition, it can be seen that for all investigated values of $C$, the combined-far-loss scheme outperforms the other traditional schemes in reducing $P_{ren}$, especially, at the worst case $C$ (i.e., $C=10$ channels), it outperforms the optimal scheme. This is because of the extra consideration of the batching size that is used together with the batching time to limit the batching duration in the combined-far-loss scheme. As a result, the batching duration is reduced compared to the other schemes.

Figure 5-7 (c) shows that although the two basic schemes and the combined-for-profit scheme give lower $P_{blk}$ than the optimal scheme, the optimal scheme still keeps $P_{blk}$ under 0.05 for all investigated $C$. In addition, the combined-for-loss scheme gets the highest $P_{blk}$ at the targeted $C$ due to the shortest batching duration, which weakens its improvement in $P_{ren}$. 
Chapter 5. Batching for Multimedia-on-Demand Scheduling in Wireless Networks

(a)

(b)
Figure 5-7 Performance of optimal timeout-based scheme and traditional schemes as the available multicast channels in a cell vary

Figure 5-8 presents $P_s$, $P_{blk}$, and $P_{ren}$ when $\lambda$ changes from 400 to 1900 request/hour with $W=4$ minutes, $N=14$ in traditional schemes and $C=20$. The parameters $W$ and $N$ are independently chosen given the target arrival rate $\lambda = 1000$ request/hour.

It can be seen that two timeout-based schemes including optimal scheme and basic scheme offer rather flat $P_s$, $P_{blk}$ and $P_{ren}$ as a function of $\lambda$, while in the other schemes these metrics can vary quite significantly with $\lambda$. Figure 5-8 (a) shows that the optimal scheme and the basic timeout-based scheme give more consistent and higher $P_s$ than other schemes. In contrast, as $\lambda$ goes higher than the target rate, the basic size-based scheme gives lower $P_s$. It is because in the size-based scheme, users are batched until $M$ users are collected. As $\lambda$ increases, shorter time is required to aggregate $M$ users, which increases the traffic load in the system, thus increasing $P_{blk}$, as shown in Figure 5-8 (b). As $\lambda$ drops, longer time is required to aggregate $M$ users in a batch, which leads to increased $P_{ren}$ and reduced $P_s$ as shown in Figure 5-8 (c) and (a).

The combined-for-profit scheme first operates according to the timeout-based scheme when $\lambda$ is high, and according to the size-based scheme when $\lambda$ drops. In contrast, the combined-for-loss scheme operates as the timeout-based scheme when $\lambda$ is low, and the size-based scheme when $\lambda$ increases. These results are consistent with those in [55].
Chapter 5. Batching for Multimedia-on-Demand Scheduling in Wireless Networks

(a) Satisfaction Ratio

(b) Blocking Probability

Legend:
- basic timeout-based
- basic size-based
- combined for profit
- combined for loss
- optimal timeout-based
Chapter 5. Batching for Multimedia-on-Demand Scheduling in Wireless Networks

Figure 5-8 Performance of optimal timeout-based scheme and traditional schemes as the user request arrival rate varies

In summary, this chapter studied the MDS problem for both loss systems and delay systems in the wireless environment. As a result, a general algorithm framework was developed for each system respectively. Following this, a systematic study was given to the MDS problem for loss systems. We derived an analytical model for the performance evaluation of the basic timeout-based scheme considering wireless resource constraints. We found that there is a tradeoff between the blocking probability and the reneging probability. A long batching time can improve the resource usage, thereby reducing the blocking probability. However, users would abandon their requests due to the long waiting time, thus increasing the reneging probability, which in turn reduces the efficiency of resource utilization. Also we found that there exists an optimal batching time that can strike this balance. Furthermore, it was found that the change of request arrival rate has no much impact on the system performance, such as the satisfaction ratio, blocking and reneging probability, especially when the request arrival rate is high. On the other hand, different amount of available resources require different optimal batching time. This indicates an optimal batching time could be appropriate for a range of users’ requests arrival rate, hence saving the resources of computing the optimal batching time. But, when the available resources in the system change, a different optimal batching time is required to optimize the system performance.

An optimal timeout-based scheme was proposed based on this analytical model to maximize the satisfaction ratio in the system. In this scheme, the batching duration time adapts to a certain system status including service quality constraints, the traffic load, user profiles and resource
availability. Simulation results showed improved performance using the proposed optimal scheme compared to traditional schemes, when the system status varies. In addition, we re-evaluated the traditional batching schemes but now with wireless resource limitation also being considered. Our results showed that in order to maximize the satisfaction ratio, when the request arrival rate changes, the basic timeout-based scheme outperforms the other traditional schemes, whereas the combined-for-loss is better than the other traditional schemes as the available resource changes.

The proposed optimal scheme provides a superior solution to service providers or operators to offer multicast-based MOD services over resource-limited wireless networks. Especially, under varying resource limitations, service providers and operators can still get superior performance by adapting the batching time to the changes accordingly so as to make an optimal balance between blocking users and users reneging, in other words, between profitability and quality of service.

In this chapter, the MDS was studied in a single cell. Nevertheless, it is expected that the tradeoff between the blocking probability and the reneging probability also exists in a multi-cell scenario. This is because the number of cells only affects the available resources, thereby affecting the blocking probability. The method of calculating the blocking probability in a multi-cell scenario is more complicated than the method in a single cell. If we do not consider interferences between cells, each cell is independent with each other and can be regarded as having an independent resource pool. The users’ requests that are aggregated from each cell can be assigned to that particular cell if there are enough resources. Therefore, the method used for calculating the blocking probability in a single cell can be used in multiple cells. In the case that we consider interferences between cells, the introduced impact is the amount of the maximum capacity in a cell. The cell capacity would be ‘soft’, which means that the maximum capacity in a cell is flexible and limited by the amount of interference in the air interface. For instance, the less interference is coming from neighbouring cells, the more channels are available in the middle cell. Therefore, the calculation of the blocking probability would be more complicated and is associated with all the traffic loads in the system. This is left for future study.

Another thing worth noting is that in this chapter, the MDS problem was studied in a single network scenario. Therefore, efficient MDS algorithms for the multiple networks scenario are to be developed.

As discussed in the Chapter 4, the MUA algorithms operate by dealing with the overall resources in all the participating networks in the system in order to optimize the efficiency of resource utilization, QoS received by users and service coverage throughout the heterogeneous networks. Basically, the MUA algorithms perform two-dimensional resource allocation for a batch of users (i.e., a multicast group), i.e., to determine the serving cells of individual users and the transmission bandwidth of desired multicast contents. Novel MDS schemes will be studied in the
next chapter, and combined with the MUA algorithms in the context of heterogeneous wireless networks comprising MBMS-enabled UMTS and DVB. That study will focus on a delay system, as it is more widely researched than the loss system for the MDS problem. Another reason is that in the delay system, the MDS and MUA algorithms are more closely related than in the loss system, hence it is more meaningful to explore the methods of combining MDS and MUA in the delay system, which will be further discussed in the following chapter.
Chapter 6

6 Combining Batching and Multicast User Assignment in Multiple Networks

This chapter aims to develop a complete multicast resource management mechanism that consists of both MDS (multicast-on-demand scheduling) and MUA (multicast user assignment) algorithms. This research work can be seen as an extension of previous work in Chapter 4 and Chapter 5 for the development of MDS and MUA algorithms respectively. With regard to the MDS, this chapter will explore an efficient multicast-on-demand scheduling algorithm considering heterogeneous users and access networks, in the context of delay systems. With reference to the MUA, this chapter will discuss the possibility of coordinating user assignment algorithms and scheduling algorithms for the purpose of efficiently working in a real time environment. A generic multicast resource management framework is developed, in which various algorithms can be adopted for different optimisation objectives. In addition, the distributed resource management architecture will be discussed in this chapter including function entities, signalling and service provisioning phases, for the purpose of algorithm implementation in the interwoking of cellular and broadcasting networks. Three functional blocks are designed to perform the respective resource management algorithms, namely, RM triggering (RMT) function, service scheduling (SS) function and network and QoS selection (NQS) function. These functions are coordinated in the resource manager in each underlying networks. Parameters that are exchanged between these functions are devised.

Firstly, the possible interaction between the MDS and MUA will be discussed, based on the previous investigations in Chapter 4 and Chapter 5. Based on this discussion, a generic resource management framework for multicast delivering is derived. The detailed operation of the overall multicast resource management algorithms and functional architecture are described. In addition, an example procedure for the GMoD (i.e., group-based multimedia-on-demand) service provisioning is illustrated. Following this, the overall improvement of the MRM algorithm that consists of MDS and MUA is checked by simulation. Finally in this chapter, the MRM algorithm is evaluated by studying different cooperation degrees between constituent networks.
6.1 Interaction between Multicast on-Demand Scheduling and Multicast User Assignment

With reference to the algorithm operation flowchart of the MDS and the MUA in Figure 5-3 and Figure 4-4 respectively, two algorithm steps are the same, namely queue prioritisation and queue selection. This indicates that these two steps are able to serve as an interface linking MDS and MUA algorithms. Moreover, it indicates that the MDS and the MUA can work in a complementary way. The MDS can perform batching before MUA in order to aggregate enough users so that the multicast channel can be efficiently utilized. On the other hand, MUA performs a more comprehensive resource allocation algorithm especially in a multiple wireless network scenario, determining the serving cell and the transmission data rate of individual users. Moreover, by combining with the MUA, the MDS is able to carry out more efficient prioritisation scheme by taking the status of underlying networks into account. Based on the discussion above, it can be concluded that linking MDS and MUA at the step of queue prioritisation and queue selection, while replacing the resource allocation step in the original MDS algorithm with the user assignment step in MUA algorithm is expected to efficiently perform the MRM for multicast service provision in the environment of multiple wireless networks. The flowchart of the MRM operation with combined MDS and MUA is illustrated below.

![Figure 6-1 Flowchart of the MRM operation](image-url)
Chapter 6. Combining Batching and Multicast User Assignment in Multiple Networks

The flowchart of the MRM operation aims to be flexible enough to accommodate different subalgorithms so as to achieve different optimization objectives. It mainly consists of four parts, namely batching, queue identification, queue prioritization and user assignment. By using different schemes in each step, different objective can be achieved. For instance, to maximize the network throughput, an effective batching algorithm should try to aggregate as many users as possible in a batch. And an effective queue prioritization algorithm should give priority to the batch that has the most users for scheduling. On the other hand, to maximize the monetary profit of network operators, the resource cost of different networks from available networks should be taken into account when designing algorithms for batching, queue prioritization and user assignment. The flowchart of the MRM operation works as follows.

Upon the arrival of user requests, the MDS batches user requests for multicast contents by using effective batching mechanisms, such as timeout-based batching, size-based batching, hybrid batching and delay-aware batching. An efficient algorithm of batching is expected to be able to balance the trade-off between the efficiency of resource utilization on the one hand and the reneging probability and the content set-up delay on the other hand. After batching, the batching queue that is eligible for scheduling is identified by MOS, which then triggers MUA and inputs some user- and service-relevant parameters for performing MUA. The transferred parameters include the geographical distribution of batched users, the user preference, and the content quality.

Taking into account parameters provided by MDS, plus some network-relevant information, such as the available resources, achievable QoS and corresponding costs, which are collected in MUA, the identified queues are prioritized according to a certain criteria, such as first-come-first-served, maximum-queue-first and maximum-profit-first. Subsequently, for each queue based on its priority order, MUA determines the best serving cell and the transmission data rate to serve the respective content to batched users. As a consequence, the users batched by MDS in one queue are re-batched/sub-batched to form several smaller groups according to the selection results of MUA in serving networks and transmission data rates.

For users that are assigned to the same networks, they are re-batched to become a group. Moreover, out of these users, the users that are assigned to the same transmission rate are re-batched to become a subgroup. These users are notified of the regrouping operation (in terms of serving cell and transmission rate) by the group manager (GM) via the multicast signalling channel (MSCH). Furthermore, a middleware in the user's terminal manages the network layer services by establishing or releasing the respective multicast session through the chosen air interface, as a response to the notifications. The MSCH is also used for informing users of the re-assignment, such as the vertical handover due to the changes of the access network, or
6. Combining Batching and Multicast User Assignment in Multiple Networks

degradation/improvement of the transmission data rate due to the changes of the available resources. More details about the MSCH and the middleware can be found in [68].

It is possible that some users cannot be served because there is insufficient network resource available, or there are not enough users batched for compensating the consumed cost. Consequently, these users are re-batched in the original batching queue, together with subsequent users requiring the same content, until either they renege or are served. Note that in a loss system, in the scenario described above, the users will be lost from the system, rather than being re-batched. Hence, the MUA operation is more closely related with the MDS operation in a delay system.

In general, during the interaction between MDS and MUA, MDS performs the service area-based batching only considering the user and service relevant information, but unaware of the network relevant information. While MUA re-batches users by taking into account the network relative information, as well as the service and user relevant information provided by the MDS.

6.2 Multicast Resource Management Framework

The main, high-level, functional blocks of the developed resource management mechanism can be seen in Figure 6-2.

![Figure 6-2 Top-level algorithm functional blocks for resource management](image-url)

**Inputs:** There are five main inputs to the resource management framework as shown in Figure 6-2.
Chapter 6. Combining Batching and Multicast User Assignment in Multiple Networks

- The content profile includes the content bit rate, the content duration, and the price of each transmission rate. These parameters are available from the service provider or the network operator who provides the content;

- The user profile includes the user service set-up delay tolerance, the content preference, utilities with relation to a set of content rates, the terminal capability and the location. Users' explicit requests for particular contents are expected to provide these parameters, except for the user's location that is provided by a specific entity in the network, i.e., DPS;

- The network status includes the achievable transmission rate for a particular content and the corresponding resource cost in each cell. This information is provided by the individual networks based on their internal mechanisms. Due to the commercial constraints on information availability, some operators may not provide complete real-time information on network status;

- The constraints identify rules the algorithm has to respect when assigning users into networks. For instance, a user can be served by only one network or can be served by multiple networks simultaneously; the maximum available resources in each network can not be violated.

- The optimization objective function is specified by the network operators to identify the direction of the optimization algorithm in any given situation. For instance, the objective assumed in our work is to maximize the system profit of network operators, taking into account the user utility and cost of consumed resources. More detailed explanations regarding the concept of system profit can be found in Chapter Error! Reference source not found•.

Each of the functional blocks designed to implement the multicast resource management mechanism is briefly described below.

**RM Triggering Function:** the purpose of this functional block is to activate the operation of the MRM algorithms, thereby triggering the batching of requests, and subsequently the assignment of users among different access networks when users send request for contents; or triggering a rebatching and re-assignment during the period of data transmission to users. In its simplest form, this function monitors the arrival of user requests. Upon the arrival of users' requests, the MDS algorithm is triggered to batch these requests in particular queues. Or in a more complicated form, the RMT function monitors each user's horizontal handover and the degradation of the user's received QoS. When a user requests a horizontal handover due to the moving from the serving cell to the targeted cell, or encounters the degraded quality of received service at terminals, the
MRM algorithms are triggered to be performed. This user will be assigned to a cell where the profit is the highest. The RMT function triggers the service scheduling function by sending a MDS request that contains the user- and content-relevant parameters. As a reply to the MDS request, after performing respective algorithms, the SS function sends back a MDS reply message that contains the results of serving cells and data rates selected for all the users and desired contents.

**Service Scheduling Function**: The purpose of this functional block is to perform the respective MDS algorithms. This function creates user groups by aggregating users with the same interests in one multicast transmission, furthermore determines when the desired contents are eligible for scheduling. A content is eligible until its service set-up time is postponed for a batching duration. This function takes into account part of user relevant parameters, e.g., user service set-up delay tolerance and content preference. The batching duration is identified by the batching time (i.e., time-based batching), by the batching size (i.e., size-based batching), by a hybrid of time- and size-based, or by the explicit user’s delay tolerance in the waiting time for service set-up. After the batching, the NQS function is performed to allocate serving cells and transmission rates to users who do not renege and still retain in the batch that requests an eligible content. The SS function triggers the NQS function by sending it with a MUA request that contains the parameters relevant to the eligible contents and remaining users interested in these contents. As a reply to the MUA request, after performing respective algorithms, the NQS function sends back a MUA reply message that contains the results of serving cells and data rates selected for these users and contents.

**Network and QoS Selection Function**: the purpose of this functional block is to perform the respective MUA algorithm, i.e., serve desired multicast services to users in the best way in terms of the transmission cells and data rates, by taking into account algorithm relevant parameters (i.e., constraints and objective), the network status, user and content relevant parameters provided by the SS function. It selects the most appropriate access network for admittance of a new service request, or for a handover initiation of an on-going service in order to make an efficient use of shared resources between networks and maintain undisrupted quality of service. This involves efficient resource management algorithms to make decisions based on network conditions and user demands. The decision is likely to be made as users join or leave a service or move through an area, which is triggered by the RMT function. Depending on the service and the number of users, it may be more efficient to use a specific network. Taking advantage of these efficiencies requires effective resource management algorithms as well as close co-operation between networks.

If there is no enough resource to guarantee the transmission data rate desired by a user, there are two methods determining how to deal with these users, named blocked-calls cleared strategy and
Chapter 6. Combining Batching and Multicast User Assignment in Multiple Networks

blocked-calls delayed strategy. In the former strategy, the user request is simply tuned away (blocked) and the user tries to connect again later. However, in the latter strategy, the user request goes back to the original batching queue and is dealt with in the same chosen order. When the request is eligible for scheduling, the NQS function works again until either this request is served or reneges due to the long waiting time for the service set-up.

**Output**: as the outputs of the MRM framework, the serving cell and the transmission rate of each user and content are determined.

### 6.3 Distributed Resource Management Architecture for Multicast Service Provisioning

In our work, a distributed resource management framework has been defined for providing improved service to users at a reasonable cost. The distributed approach coordinates different networks' resource usage without having a centralised entity. Instead, the RM (i.e., resource manager) of a network communicates and exchanges resource-related information with its counterpart residing in another network. This approach allows operators to retain full control on their respective precious network resources and traffic load information. Such an RM resides in the IGM (i.e., IoN Gateway). Signalling and data exchange required for resource management are carried out through the logical interface between the IGWs.

One question arises when we consider who handles the resource management. In other words, where the batching of users' requests and the scheduling for the group-based services take place. The answer depends on the business model. Two business models are widely accepted in the B3G environment, namely, Network Provider (NP)-driven model [35] and Service Provider (SP)-driven model [101]. They reflect the current trends in the mobile networks. The objective is to provide a flexible business model enabling network operators or service providers to offer innovative new heterogeneous services anytime, anywhere, in any medium or combination of media, while users are on the move. These two business models are presented as follows in Figure 6-3 and Figure 6-4 respectively.
Chapter 6. Combining Batching and Multicast User Assignment in Multiple Networks

The NP-driven business model is operator-centric, as the management of both users and service distribution is handled by network operators. The network that performs the management is called the originating network. The network operator of the originating network is denoted as \( np_0 \). The originating network is in charge of all user profiles, subscriptions and service related configurations. Billing and authentication is also based on the originating network and in this case the user will receive one bill per service. In addition to managing the users, the originating network has business agreements with a set of network operators, \( np_1, \ldots, np_N \), which are included in a set denoted as \( NP_{CP}(np_0) \). These networks are called cooperating networks of the originating network. The user has a direct business agreement with \( np_0 \) that offers the service and earns revenues from the user. In order to maximize its revenue, \( np_0 \) can either serve the user in its own network, or switch the user to one of the cooperating networks in the \( NP_{CP}(np_0) \) set to obtain better QoS, but charged for consuming the resource in the cooperating network. The selection of the network from the \( \{ np_0, NP_{CP}(np_0) \} \) set is done intelligently in the RM entity in \( np_0 \). This is assisted by information exchange with the RM counterparts in the networks of the \( NP_{CP}(np_0) \) set.
In the SP-driven business case, the user has a direct business agreement with the service provider. The service provider is a third party company that provides, manages and distributes business services directly to users from a central data centre. The service provider could cooperate with another (AAA) third party that provides authentication, authorization, and accounting for both users and service providers. In this case, the user has a direct business agreement with the AAA provider. The AAA provider enables service providers as well as network operators, to offer their services to the user without having direct business agreements between them. However, the third party AAA provider has business agreements with users, service providers as well as network operators. Clearly, the SP-driven business model is not operator centric as in the NP-driven business model. The value chain will be further explained by assuming a user has a direct business agreement with the service provider. The service provider offers the service and earns revenues from subscribed users. However, from these revenues the service provider has to pay the content provider and a set of network operators, $np_1$, ..., $np_N$, which are included in a set denoted as $NP(SP)$, for delivering contents. The resource management is dealt with in the RM in the service provider. The RM in the service provider will talk to the RM counterpart in the network of each network operator in the $NP(SP)$ set, for distributing required services through appropriate networks to end-users.

To summarize the features of the existing business models for the future B3G systems, we can come up with the following points:

- The main business actors involved in the business model include users, network operators and service providers. Aside from that, another business actor, the content provider that creates the contents, is also involved in the business relationship. It has a business role in providing the content that is delivered to the users through the network operators.
agreement with the originating network (i.e., in the NP-driven model), or with the service provider (i.e., in the SP-driven model). The relationship with the content provider is not presented because it is not involved in the procedure of the resource management;

- A user receives one bill from the dominant business actor that the user is subscribed to. This actor could be the network operator (i.e., in the NP-driven model) or the service provider (i.e., in the SP-driven model). The dominant actor cooperates with a set of network operators. These operators can own networks of various wireless access technologies. The dominant actor is in charge of the management of users’ requests and service distribution among all the involved networks. Service traffic can be distributed through any combinations of constituent network components in the system.

Compared to the SP-driven model, the NP-driven model is more pragmatic and easier to be implemented based on the current business model in network operators, because both the current business model and the NP-driven model are strictly operator centric. Therefore, for the purpose of easy implementation in the near future, the NP-driven business model is assumed in the discussion followed. Nevertheless, the MRM algorithms are designed to be applicable to both of these two business cases.

Resource management mechanism is implemented in a resource manager that is distributed between networks. In Figure 6-5, we show the detailed resource management architecture in the context of IoN architecture including each of the identified resource management functions.

Distributed resource management provides the flexibility of supporting multiple operators for the future enhancement. Multicast services are offered to users through various wireless access
networks owned by different operators, each of which owns one access network\(^2\), e.g. UMTS or DVB. One access network is an independent administrative domain cooperating with each other via a Gateway (GW) and a logical interface in the network and service layers.

In the previous section, the resource management functions have been explained in detail, such as RMT, SS and NQS. This section focuses on discussing the availability of parameters required by the RM to carry out the resource management algorithms, and explaining the interaction of the RM with other entities in IoN GW, such as the GM (group manager) and DPS (device presence system).

The information regarding contents and users is provided by GM. The user relevant parameters, such as the user service set-up delay tolerance, the content preference, utilities with respect to a set of content rates and terminal capability are collected during the procedure of subscribing users for their preferred contents, which could happen anytime before the exact content distribution time via a variety of means, e.g., mobile phone, internet, etc. Any updates of the user relevant parameters are informed by the GM to the RM. In addition to user relevant parameters, the content relevant parameters, such as the content bit rate, the content duration and prices of transmission rates are also provided by the GM that provides an interface for content or service providers to configure service-related parameters.

An important user relevant parameter is the geographical position in each of the constituent networks. This parameter is used for the user assignment in order to accurately deliver contents into the exact cell where there are subscribed users so as to utilize resources efficiently. The DPS performs the location discovery, management and tracking of user devices that have subscribed to contents across heterogeneous networks. Cooperating with broadcasting networks allows mobile networks to determine users' current residing cells accurately and efficiently. Three original schemes for the mobile user location discovery in cooperative cellular and broadcasting networks have been proposed in [40].

Resource cost calculation (RCC) calculates a list of real time ‘cost quotes’ of delivering the given content with a set of possible QoS levels in each cell of the access network where the RCC resides. The cost quote is expressed by three-tuple (cell ID, QoS set (transmission rate, delay, jitter, packet loss ratio etc.), current resource usage, unit cost), for instance (3G/MBMS cell 1, (128kbps, 64 kbps), 256kbps, 1 penny per Mbits). This example means the content can be transmitted in MBMS cell 1 with the rate of 128 or 64 kbps at a price of 1 penny per Mbits. The amount of resources that are currently consumed is 256kbps. The transmission rate set in the cost quote is 0

---

\(^2\) It is possible that one network operator owns more than one access network. However, in this study, for the purpose of simplicity in results' presentation, it is assumed that one network operator owns one access network.
if under the current network conditions, there are no enough resources to serve the content with the required rate. Note that there could be many QoS parameters as presented above. Here, for simplicity, we only consider one QoS parameter, i.e., transmission rate.

To calculate the cost quote, the network is required to monitor the status of the underlying network segments on an individual cell basis, e.g., the current traffic loads in a cell, the predicted load required by satisfying a certain QoS requirement, the maximum resource limitation in a cell. These parameters are collected by the LM that interfaces with the GGSN in a cellular network and interfaces with the IP/DVB gateway manager in a DVB network. Furthermore, based on the information collected, each network calculates the unit cost and the supportable QoS levels of required contents.

6.4 Service Provision Procedure

6.4.1 Service Provision Procedure Phases

The multicast resource management enables the cost-effective multimedia-on-demand service provisioning in wireless networks. To give a clear picture regarding the role of resource management components during the multicast-on-demand service provisioning in IoN, Figure 6-6 is used to show the phases of this service provisioning including some multicast resource management steps. This procedure has been adopted from the MBMS specification [7] with some alterations made to cater for the on-demand service provision in the IoN environment.

![Figure 6-6 Phases of multicast-on-demand service provision in IoN](image-url)
The phases of subscription, joining and leaving are performed individually per user. The other phases are associated with a service, i.e. for all users interested in that particular service. The sequence of phases may repeat depending on the needs of service/content delivering. In addition, some phases are supported by the group management mechanism, such as, the subscription, service announcement and service notification. Details regarding the group management can be found in [68]. The phases are listed below and briefly described.

**Subscription**: this phase establishes the relationship between the user and the network operator, which allows the user to receive the related multicast service. Service subscription is the agreement of a user to receive services offered by the operator. During this phase a unique group identifier is created and assigned to subscribed users. Meanwhile, a MSCH channel is configured by the GM in the control plane for the transmission of the control signalling during the future phases.

**Service Announcement**: this phase allows users to discover or be informed about the range of multicast user services available. It is used to distribute to users the information about the service parameters required for the service activation (e.g. IP multicast addresses), other service related parameters (e.g. offered service rates and prices), and the MSCH channel parameter required for the establishment of the control plan.

**Joining**: Joining (i.e. service activation by the user) is the process by which a subscriber joins (becomes a member of) a multicast group that corresponds a particular multicast content in the data plane and meanwhile joins a multicast group that corresponds a MSCH channel in the control plane. During the joining, the user indicates to the network that he/she wants to receive a particular multicast content, his/her delay tolerance, and gives a rating with respect to the offered transmission rates from the provider. Meanwhile, a control plane for the user is established for receiving the MSCH channel.

**Multicast Resource Management**: this phase is developed specially for provisioning multicast on-demand services in IoN environment. The multicast resource management algorithms are carried out in this phase that takes the following three steps.

1. **Multicast-on-Demand Scheduling**: the SS function aggregates requests for the same multicast user service in the same group for the batching purpose;

2. **Resource Cost Request**: the RM in the originating network collects information from other counterpart RMs for the user assignment purpose. During this information exchange, the RM calculates the cost quote in its own network and collects cost quotes from the counterpart RMs in other cooperative networks;
3. *Multicast User Assignment*: After getting all the required information, the NQS function performs the user assignment algorithm to select the best access networks and the most appropriate transmission rates for all the subscribed users that have joined a multicast group.

**Session Start**: This phase is triggered for the multicast bearer resource establishment in the chosen access network for multicast data transfer. If it fails for any reasons, another access network will be tried starting from the one with the lowest resource cost until the service is set up or no available access networks are left.

**Service Notification**: The GM is triggered by RM after the transmission network and data rate have been decided for each subscribed user. Using the MSCH channel, the GM informs the users about forthcoming multicast data transfer and configures the terminals for data reception at proper air interface.

**Data Transfer**: This is the phase when multicast data are transferred to subscribed users. At the same time, during this phase, in order to maintain the received quality of services at users, vertical handover might be triggered due to the change of user location or resource condition. As a result, the multicast resource management phase is repeated to reselect suitable transmission networks and data rates for users. To implement these changes, the GM function sends control signalling via MSCH channel to the terminal that in turn implements this handover.

**Session Stop**: This is the point at which the network determines that there will be no more data to send for some period of time. As a result, the bearer resources associated with the session are released.

**Leaving**: Leaving (i.e. multicast deactivation by the user) is the process by which a subscriber leaves (stops being a member of) a multicast group, i.e. the user no longer wants to receive data of a specific multicast service. This phase can happen at any time of the data transmission. It is implicitly performed, when the Session Stop is performed.

### 6.4.2 Signalling

Distributed resource control gives the promise of efficient sharing of network resources, but requires information exchange between networks. The signals exchanged between networks for resource management include resource metric and the associated parameters for service provision. However, such signalling mechanisms between networks are currently lacking. Therefore, the underlying signalling issues required for distributed resource management have also been investigated, focussing particularly on the requirements of control information flow between different resource management entities and its practicality. Since network operators are not expected to release their real time resource status, commercial constraints on information availability and the provision of standard APIs are considered while restricting information
visibility between network elements. The interface between network resource managers and the procedure for service provision based on joint-usage of networks are identified.

This section presents signalling flows showing the support of distributed resource management in achieving multicast-on-demand service provisioning in the IoN environment, specifically focusing on the phase of multicast resource management. The other signalling flows in the phases of subscription, service announcement, joining, service notification etc., have been identified by the group management work package and can be found in [68].

Figure 6-7 and Figure 6-8 illustrate a signalling example for the multicast-on-demand scheduling and multicast user assignment in multicast resource management phases.

Users first subscribe to the GM, which enables users to receive a multicast service from IoN. Then by service announcement, GM notifies subscribed users of information about offered services and parameters for the MSCH creation. During joining, users inform the GM of their interests in a particular content and the parameters that are required for the resource management purpose. Joining can be done via different media, such as the UMTS radio uplink or wired Internet access. Meanwhile, users join a multicast group for receiving control information conveyed via the MSCH channel. Once the GM obtains user information during Joining, it triggers RM by sending a RESOURCE MANAGEMENT REQUEST MESSAGE along with parameters of multicast contents and users to perform relevant resource management algorithms.
Once a resource management request message has been obtained, RMT triggers SS to perform service scheduling. Note that RMT can trigger SS on its own during data transmission for the quality maintenance purpose, without necessarily being triggered through the user joining. During service scheduling, some batching queues are established to aggregate user requests for a certain time. Users requesting for the same multicast content are put into the same batching queue. Once time is up, SS sends a MUA REQUEST MESSAGE to NQS to require the assignment of transmission networks and QoS levels for all the eligible contents and involved users that are aggregated during the service scheduling. As a result, NQS sends back a MUA REPLY MESSAGE along with the chosen transmission networks and QoS levels. Subsequently, the RM notifies the GM of the results of the selection, and then the GM initiates the establishment of the required multicast bearers for the data plane.

![Diagram of Resource Manager and MUA Request/Reply Signalling](Image)

**Figure 6-8 Signalling for multicast user assignment in IoN**

Figure 6-8 illustrates in more detail the signalling for the multicast user assignment in multicast resource management phases.

After obtaining a MUA REQUEST MESSAGE from SS in the same resource manager, NQS performs the MUA algorithm. Firstly, NQS collects the ‘cost quotes’ from each of the underlying networks including the originating network and cooperating networks. RCC is responsible for calculating the cost quote of delivering a certain service through its network, by asking the LM the network status information on a cell basis. After obtaining the cost quotes, taking into account
other parameters including content and user relevant parameters, constraints and optimization objective, NQS performs the MUA algorithm to select the access network and QoS level (in particular transmission rate) for all the eligible contents and aggregated users. Finally, during the Session Start phase, the respective multicast bearers based on the chosen delivery networks/cells and QoS levels are established, through the corresponding components in the chosen networks. At the same time, the GM sends control signalling via MSCH to get subscribed user terminals configured so as to receive data properly on the proper air interface.

6.5 Performance Evaluations

The objective of the MRM algorithm is to maximize the system profit. This is also beneficial from the capacity perspective, as a higher profit in many cases also means that more services are served to users. Generally, in addition to the resource management algorithms, the system profit depends on the user's delay tolerance, the access technology, the traffic load and resource cost of the access technology etc. These factors determine the profit the network is able to achieve, and how often a particular network is accessed. This section presents a thorough simulation study on the proposed MRM algorithms in different conditions of the user's delay tolerance, and resource cost.

As discussed before, the MRM consists of MUA (i.e., multicast user assignment) and MDS (i.e., multicast-on-demand scheduling) schemes. The algorithms of these two schemes have been intensively studied in Chapter 4 and Chapter 5 respectively. The objective here is to study the overall performance of the MRM algorithm and the impact of each part (i.e., MDS and MUA) on the overall performance. Based on our previous studies, there are different combinations of MDS and MUA algorithms. For instance, the Timeout-based MDS plus the MP MUA algorithm, or the Delay aware MDS plus the MP MUA algorithm. These two new algorithms can be seen as the extension of the conventional pure Timeout-based batching algorithm, i.e., BMQ in [57] and the conventional pure Delay Aware batching algorithm, i.e., DAB in [56] respectively. In Section 5.3, these four algorithms have been briefly explained in the context of delay systems.

The difference between the conventional batching schemes and our schemes lies in our specialized MUA algorithm. It takes into extra consideration of heterogeneous wireless networks and users in the system. Moreover, the rate-adaptive feature of multimedia services is utilized in the MUA algorithm. It is designed to intelligently adapt the content transmission rate to the network status, and select the most appropriate network to transmit the contents. The similarity of our algorithms to the traditional algorithms is that users have to be batched for a certain period of time before they are eligible for scheduling. The batching is implemented by the MDS algorithm. The proposed MRM algorithms are explained in more detail as follows.
1. Timeout plus MP: user requests are firstly batched in a batching queue. Users for the same multicast contents are batched in the same queue. When a new request arrives or a content stream is completed, we choose the content that has requests with waiting time exceeding a prespecified batching threshold. In line with the timeout based batching scheme in [57], we set this batching threshold equal to the user's minimum waiting threshold. In order to increase the batching effect, these chosen contents are prioritized for resource allocation, based on the "maximum-profit first" criterion adopted by the MP algorithm in MUA. The proposed MP algorithm has been discussed in Chapter 4. Moreover, the MP algorithm assigns each of interested users to a combination of a serving cell a transmission rate which either generates the largest profit or consumes least resource;

2. Timeout plus MBMS/DVB by default: different from the above MP algorithm in MUA, this algorithm selects the content that has the most requests. Moreover, for each selected content, interested users are assigned to a particular access network that is selected by default (e.g., 3G/MBMS or DVB). If there are not sufficient resources available, the other network is tried, until all users are assigned or no feasible network available.

3. Delay Aware plus MP: similar to the first MRM algorithm, a user is firstly batched in a particular batching queue after requesting a content. Meanwhile, each user's delay tolerance is stored into a database. When a user is going to renege from the system due to a lone time batching, the system is informed and picks the queue where this user resides for scheduling. In addition, the queue that has reneged users are also selected to maximize the batching efficiency. Subsequently, applying "maximum-profit first" criterion, the MP algorithm determines the order of allocating resources to desired contents, and then selects the serving cells and transmission rates for each user in the desired contents. If the available resources are not sufficient to serve these users, the user that triggers this scheduling is lost from the system. The other users get back to their original batching queue and wait for the next scheduling time;

4. Delay Aware plus MBMS/DVB by default: different from the third algorithm, only the queue where a user is going to renege is eligible for the resource assignment. Moreover, this algorithm assigns users to a particular access network by default (e.g., 3G/MBMS or DVB). If there are not sufficient resources available, the other network is tried, until all users are assigned or no feasible network available.

Firstly, the MRM algorithms will be evaluated by comparing with two conventional algorithms, namely Timeout-based batching scheme (i.e., BMQ algorithm), and Delay Aware batching scheme (i.e., DAB algorithm). Since the conventional algorithms were proposed considering a
single network, we first evaluate the MRM algorithms in a single 3G/MBMS network with 7 cells and a single DVB network with one cell respectively. Following this, the MRM algorithms will be further evaluated in an interworked 3G/MBMS and DVB infrastructure. In addition, considering that MRM algorithms require information exchanging and sharing across diverse networks that might belong to different administrative domains, the impact of the availability and validity of exchangeable information on the algorithm performance will be investigated. The details of the simulator can be found in Appendix A.3.

6.5.1 Evaluation of Multicast Resource Management Algorithms in a Single Network

Firstly, we compare the proposed MRM algorithms with conventional algorithms in a single 3G/MBMS network, under different user delay tolerance. This is the batching time duration for which users stay in the batching queue without reneging. It has been observed in Chapter 5 that the user’s batching duration has a significant impact on the performance of the MDS algorithms. In this section, we investigate its impact on the performance when combining MDS algorithms with MUA algorithms. Figure 6-9, Figure 6-10 and Figure 6-11 show the performance of overall system profit, user loss probability and average delay, as the function of the user’s waiting time threshold. The system profit is the overall user throughput in the system. It is assumed that the available resources in both 3G/MBMS and DVB are 1000 Kbps per cell, and the user’s mean waiting time is one minute.

![Overall system profit of MRM algorithms versus user waiting time threshold in a single 3G/MBMS network](image)

Figure 6-9 Overall system profit of MRM algorithms versus user waiting time threshold in a single 3G/MBMS network
In general, we can see from Figure 6-9 and Figure 6-10 that as the user waiting time threshold increases, the overall performance is improved in terms of the profit and the user loss probability. This is because as the user waiting time threshold increases, users can wait longer for the content set-up without reneging. This in turn reduces the arrival rate of requests for the user assignment after batching. Therefore, fewer resources are required to serve these users. However, since users have to stay in batching queues for a longer time, a longer delay is suffered. This is illustrated in Figure 6-11.
In addition, it can be seen that in the proposed schemes where the traditional Timeout-based and Delay Aware batching algorithms are combined with the MP algorithm, the performance is significantly improved compared to the traditional Timeout-based and Delay Aware batching schemes, in terms of the overall system profit, the overall user loss probability and the average delay for the service set-up. Obviously shown in the figures, given a certain user’s waiting time threshold, combining traditional batching schemes and the MP scheme obtains higher system profit and lower user loss probability and service set-up delay. It implies that to achieve the same user loss probability as the proposed schemes, the conventional schemes need to batch users’ request for a longer time, hence resulting in a longer delay for the user’s service set-up.

The efficiency of the proposed schemes can be attributed to the following reasons. When combining the batching algorithm with the MP algorithm, the transmission rate of desired contents can adapt to the available resources in the networks. Therefore, when the available resources in the network are not sufficient to serve the users, an appropriate transmission rate is chosen. However, the conventional algorithms do not use the rate-adaptive technique. Therefore, users’ requests are refused when there are no enough resources available.

Another reason is that when prioritizing desired contents, the MP algorithm considers the user location in the network and the potential profit that could be achieved in each cell by serving users. The content with the largest potential profit will be given the highest priority for scheduling. On the other hand, the Timeout-based scheme considers only the batching size for queue prioritization, while the Delay Aware scheme does not prioritize any contents for scheduling. Since prioritizing contents can maximize the batching effect and make sure the popular or profitable content can be scheduled prior to the less popular or less profitable contents that otherwise could use up available resources, the Delay Aware scheme without using any prioritization criteria gives the smallest profit, highest user loss probability and longest delay.

In addition, we can see that the Delay Aware plus MP scheme outperforms the Timeout plus MP scheme, in terms of the overall system profit and the user loss probability. This is because, the Delay Aware algorithm is aware of the exact user’s delay tolerance, and thereby batching users exactly until users are about to renege from the batching queue. However, the Timeout-based algorithm is not aware of the explicit user’s delay tolerance. Hence, users are batched for the minimum delay tolerance among all the batched users. As a result, the Delay Aware plus MP scheme can batch users for a longer time than the Timeout plus MP scheme, hence making a more efficient usage of valuable radio resources. However, in Figure 6-11, it can be seen that the Delay Aware plus MP scheme suffers a longer delay for the service set-up.
We now study the performance of the MRM algorithms in a single DVB network with one cell, when the user's waiting time threshold varies. Figure 6-12, Figure 6-13 and Figure 6-14 show the performance of overall system profit, user loss probability and average delay.

Figure 6-12 Overall system profit of MRM algorithms versus user waiting time threshold in a single DVB network

Figure 6-13 Overall user loss probability of MRM algorithms versus user waiting time threshold in a single DVB network
Figure 6.14 Overall average delay of MRM algorithms versus user waiting time threshold in a single DVB network

Similar to the performance in a single 3G/MBMS network, as the user’s waiting time threshold increases, better performance is achieved in terms of the overall system profit and the user loss probability, however at a price of an increased delay for the service set-up.

In addition, it can be seen that combining the MP scheme, our algorithms significantly outperform the two conventional algorithms in increasing the overall system profit, reducing the overall user loss probability and reducing the average delay for the service set-up. Furthermore, combined with the MP scheme, the Delay Aware algorithm achieves higher system profit and a lower user loss probability than the Timeout-based algorithm, because it has full knowledge of the user’s delay tolerance. However, this benefit on the other hand causes a longer delay for the service set-up.

In summary, the effectiveness of the proposed MRM algorithms has been proved in a single 3G/MBMS and DVB networks. Compared to conventional schemes, our schemes can achieve significant performance gains in terms of the system profit, the user loss probability, and the service set-up delay. The employment of the rate-adaptive techniques and more efficient prioritization criteria for scheduling, makes our proposed algorithm outperform the conventional algorithms. Furthermore, it was found that combined with the MP algorithm, the Delay Aware scheme outperforms the Timeout-based scheme in terms of the system profit and the user loss probability, but paying the price of a longer delay for the service set-up. However, the pure conventional Delay Aware scheme obtains worse performance than the pure conventional Timeout-based scheme.
Comparing the results generated in a single 3G/MBMS network and in a single DVB network, it can be observed that given the same user's waiting time threshold and the available radio resources in each cell, these two networks individually are able to give quite similar performance in terms of the system profit, user loss probability and average delay. Since the 3G/MBMS network has more cells than the DVB network in a common coverage, the aggregated users in a MBMS radio channel are fewer than those in a DVB radio channel, given the same user waiting time threshold. Therefore, we can conclude that to achieve the same good performance in terms of the system profit and user loss probability, more users need to be aggregated in a DVB radio channel than in a MBMS radio channel. Since a DVB cell usually has a bigger geographical area than a 3G/MBMS cell, a DVB network can easily aggregate more users in one channel. Moreover, it can be seen that a 3G/MBMS network normally needs more radio channels to accommodate the same amount of traffic as a DVB network does, especially when users are dispersed over a large geographical area. Hence, the DVB is more resource efficient than the 3G/MBMS to serve the multicast service to a group of users widely distributed cross cells.

In the following sections, we will further study the performance of the proposed MRM algorithms in the case of multiple networks.

6.5.2 Evaluation of Multicast Resource Management Algorithms in Multiple Networks

Since conventional algorithms (i.e., pure Timeout and Delay Aware batching algorithms) do not consider the scenario of multiple networks, we have enhanced them to adapt to the multiple networks environment. A straightforward modification is to serve users into a default transmission network that originally provides the required multicast contents. These users will be handed over to another network if the serving attempt to the default network fails due to, for instance, the lack of sufficient resources. To make a fair comparison, both cellular and broadcast networks are considered as the default transmission networks respectively. As a result, we term the modified version of conventional batching algorithms as Timeout + MBMS-by-Default, Timeout + DVB-by-Default, Delay-Aware + MBMS-by-Default and Delay-Aware + DVB-by-Default. Adopting the MP algorithm in the MUA operation comes out another two algorithms, namely Timeout + MP and Delay-Aware + MP. These two schemes are compared with those of the modified version of conventional Timeout and Delay-Aware schemes. It is assumed that the available resources in both 3G/MBMS and DVB are 1000 Kbps per cell, and the user's mean waiting time is one minute.

6.5.2.1 Effect of Varying User's Waiting Time Threshold

Firstly, we study the impact of varying the user's waiting time threshold on the system performance. Figure 6-15, Figure 6-16, and Figure 6-17 present the performance of proposed and
enhanced MRM algorithms in terms of the overall system profit, user loss probability and the average delay for the service set-up respectively.

Figure 6-15 Overall system profit of MRM algorithms versus user waiting time threshold in interworked 3G/MBMS and DVB networks

Figure 6-16 Overall user loss probability of MRM algorithms versus user waiting time threshold in interworked 3G/MBMS and DVB networks
Figure 6-17 Overall average delay of MRM algorithms versus user waiting time threshold in interworked 3G/MBMS and DVB networks

It can be seen that in general, the performance in terms of the system profit and the user loss probability are improved, as the user waiting time threshold increases. But a longer delay is experienced.

Regardless of the batching algorithms in the MDS operation, compared to the schemes where the desired contents are served by default in 3G/MBMS or DVB, the schemes where the MP algorithm is applied in the MUA operation significantly improve the system profit, the user loss probability and the average delay. Aside from the advantages of the MP algorithm, which have been explained before, such as the data-adaptive techniques and the profit-based prioritization criteria, the MP algorithm is able to intelligently serves the contents with the network where the higher profit can be generated. Another thing we can observe is that regardless of the MDS operation, selecting the DVB, rather than the 3G/MBMS, as the default transmission network is able to generate higher system profit, lower user loss probability while causing shorter service set-up delay. This is because, in a common geographical service area, a 3G/MBMS network has more cells than the DVB network. When the available resources in a 3G/MBMS cell are insufficient to serve required services, the users in that cell will be handed over to the DVB cell in the same coverage. Since a 3G/MBMS cell is smaller than a DVB cell, the number of users handed over to DVB is not enough to enable an efficient usage of radio resources in DVB. Hence, the performance is degraded.

When the Delay Aware batching algorithm is combined with the MP algorithm, because it can batch more users in a transmission, it outperforms the Timeout-based batching algorithm in terms
of the user loss probability and the system profit, but by paying the price of a longer service set-up delay. When combined with the MBMS/DVB by default algorithm, the Timeout-based batching algorithm outperforms the Delay Aware batching algorithm in terms of the system profit, the user loss probability and the average service set-up delay, because the Timeout-based batching algorithm uses queue prioritisation method.

To summarize the results, in the following figure we measure the improved system profit, the improved user loss probability and the improved average service set-up delay of the studied MRM algorithms respectively, with reference to the performance of the Delay Aware algorithm with 3G/MBMS as the default transmission network. The user waiting time threshold is set to nine minutes because it gives the smallest user loss probability.

Figure 6-18 Overall system profit improvement for MRM algorithms over the Delay aware + MBMS by default MRM algorithm
Based on the results generated in a single 3G/MBMS network, in a single DVB network, and in interworked 3G/MBMS and DVB networks, it is not hard to see the benefit of combining the 3G/MBMS and DVB networks in improving the system profit, user loss probability and average delay. It can be seen in the following table that for the Delay aware plus MP algorithm to generate the similar system profit and user loss probability in interworked networks to those in an
individual network, the interworked networks require a shorter user’s waiting time threshold, thereby a shorter delay suffered by users.

<table>
<thead>
<tr>
<th></th>
<th>User waiting time threshold (minutes)</th>
<th>System Profit (Kbps)</th>
<th>User Loss Probability</th>
<th>Average Delay (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single DVB</td>
<td>15</td>
<td>67006</td>
<td>0.004</td>
<td>7.8</td>
</tr>
<tr>
<td>Single 3G/MBMS</td>
<td>15</td>
<td>66920</td>
<td>0.006</td>
<td>7.8</td>
</tr>
<tr>
<td>Interworked 3G/MBMS and DVB</td>
<td>8</td>
<td>68717</td>
<td>0</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Table 3 Performance comparison of the Delay aware + MP algorithm in a single 3G/MBMS, a single DVB and interworked 3G/MBMS and DVB

In summary, we have demonstrated the improved performance of the proposed MRM algorithms in terms of the system profit, user loss probability and the service set-up delay, in a single network as well as a multiple networks environment. The proposed algorithms are mainly comprised of the MDS operation for batching, and the MUA operation for resource allocation. The efficiency of the proposed algorithms is mainly attributed to the effectiveness of the MP algorithm in the MUA operation. Compared to conventional algorithms, the MP algorithm considers more user and network relevant information when assigning users. Furthermore, the MP algorithm is able to adapt the transmission rate of desired contents to the available radio resources.

In addition, it is found that with respect to the MDS operation, if combined with an effective MUA algorithm, such as the MP scheme, the Delay Aware batching scheme outperforms the Timeout-based batching scheme in terms of the system profit and user loss probability. However, the Delay Aware batching scheme itself causes longer delay than the timeout-based scheme. This indicates that to optimize the system profit or the user loss probability, the Delay Aware plus MP algorithm can be applied. On the other hand, for the purpose of optimizing the delay for service set-up, but still achieving high profit and low user loss probability, the Timeout plus MP algorithm can be applied.

In the MRM algorithms, users are lost from the system because the waiting time for the service set-up is beyond their expectation, or are dropped during the horizontal handover. It has been found that the former factor is dominant over the latter factor in determining the lost probability, due to the wide distribution of users across cells. When a multicast group with users widely distributed is admitted into the network, multiple channels will be established across cells, which
makes a new channel set-up unnecessary during the horizontal handover. Therefore, the dropping probability can be significantly reduced. Two situations may lead to a long waiting time. The first is the shortage of available resources, and the second one is the insufficiency of interested users to compensate the resource cost. In either of the situations, users have to wait until free resources are available, or enough users are aggregated. The loss of users will lead to a profit reduction. Furthermore, the resource cost also has direct influence on the system profit, i.e., a low resource cost results in a high profit.

We have investigated the impact of the user's delay tolerance on the system performance, given the constraint on resource availability. In the following sections, we further study the proposed MRM algorithms, by investigating the impact of varying resource costs, given a fixed user's delay tolerance.

6.5.2.2 Effect of Varying Resource Cost

In the following, we will study the impact of varying the resource cost of the underlying networks. In this scenario, we assume the resource cost of MBMS is 0.5 costs per unit of transmitted data, and change the ratio of the resource cost of DVB with respect to that of MBMS from 1 to 100.

In order to maximize the impact of varying resource cost of networks on performance, we assume there is no constraint of resource availability, i.e., a batched group can be served as long as there are enough users aggregated during batching.

Figure 6-21 shows the performance of the system profit of MRM algorithms as the resource cost ratio varies. We do not show the performance of the overall user loss probability as it is always zero for all the MRM algorithms. In addition, as we discussed before, the improved performance of the MP algorithm lies in the fact that it is able to dynamically assign users to the best network that are potentially able to generate higher profit. To further analyse the simulation results, we measure the number of served users in each network when the resource cost of DVB changes, as illustrated in Figure 6-22 and Figure 6-23.
Figure 6-21 The overall system profit versus the resource cost ratio for MRM schemes

In general, we can see that regardless of the MDS algorithm, namely the Delay Aware and the Timeout-based batching schemes, the MP MUA algorithm outperforms the other algorithms, namely the MBMS-by-default and DVB-by-default. Moreover, the MP algorithm is able to be adaptive to the varying resource cost of DVB, as it traces out the outer "envelope" of the other two MUA schemes. As the resource cost of DVB is low, the MP algorithm performs similar to the DVB-by-default algorithm where all the contents are served to DVB by default. When the resource cost of DVB increases, the MP algorithm performs similar to the MBMS-by-default algorithm where all the contents are served to 3G/MBMS by default.

In addition, it can be seen that when the resource cost of DVB is low, the DVB-by-default scheme outperforms the MBMS-by-default scheme. However, when the resource cost of DVB increases, the system profit generated by the DVB-by-default scheme firstly decreases and becomes smaller than that generated by the MBMS-by-default scheme. Then the profit of the DVB-by-default scheme goes up and becomes the same as the profit of the MBMS-by-default scheme.

Another thing we can observe is that regardless of the MUA operation, in the MDS operation, the Delay Aware algorithm outperforms the Timeout-based algorithm and gives more system profit. This is because the Delay Aware algorithm adopts the exact user's delay tolerance as the scheduling point when the batched users are assigned to networks. On the other hand the Timeout-based algorithm uses the minimum user's delay tolerance as the scheduling point.
Therefore, given the same traffic load, more users can be batched by the Delay Aware algorithm than by the Timeout-based algorithm and fewer radio resources are consumed. Hence, the Delay Aware algorithm is able to generate higher profit than the Timeout-based algorithm with the same MUA mechanism.

To further explore the facts behind the results, we measure the distribution of assigned users in individual networks, which is represented by the percentage of served users in 3G/MBMS and DVB respectively. It is presented in the following figures.

Figure 6-22 Percentage of served users in 3G/MBMS versus the resource cost ratio for MRM schemes
It can be seen that regardless of the MDS operation, for the MP algorithm in the MUA operation, users are firstly served in DVB when the resource cost of DVB is small, because DVB can accommodate the same number of users with lower cost compared to 3G/MBMS. Therefore, the MP algorithm works similar to the algorithm where contents are served to DVB by default. We can roughly calculate the cost in a unit time of serving a content in a network as 

$$cost = r \cdot n \cdot rc,$$

where $r$ is the transmission rate of served contents; $n$ the number of transmission channels; $rc$ the resource cost of a unit transmission data. Since DVB covers a bigger geographical area than 3G/MBMS does (e.g., one DVB cell overlaps with seven 3G/MBMS cells), lower cost is caused in DVB than in 3G/MBMS to transmit the same rate of service when the resource cost of the unit transmission data is the same. Even if DVB has a larger resource cost than 3G/MBMS, it is also possible that DVB causes lower cost if its geographical coverage is big enough compared to 3G/MBMS. Therefore, by calculating the cost consumed by serving a certain service, the MP algorithm can select the “cheapest” network to serve users.

For the MBMS-by-default and DVB-by-default algorithms, they are ignorant of the cost relevant knowledge of underlying networks. As a result, an attempt is always made to serve users in the default network first, namely 3G/MBMS or DVB, leaving the other network unloaded.

When the resource cost of DVB increases, the MP algorithm will accommodate more users in 3G/MBMS due to the lower resource cost. Therefore, it is expected to act similar to the algorithm where contents are served to MBSM by default. It also can be seen that when the resource cost of
DVB increases, the algorithm where contents are served to DVB by default puts more traffic into 3G/MBMS than DVB. This is because these contents can not be served in DVB any more due to the increased resource cost of DVB. As a result, the contents are forced to be handed over to 3G/MBMS where the contents can be served due to the lower resource cost. Therefore, we can see that in Figure 6-21, for the DVB-by-default algorithm, the system profit firstly reduces due to the increased resource cost, and then increases due to the handover of traffic to 3G/MBMS. For the MP algorithm, because it is able to intentionally select the best network by comparing the resource cost of two networks, it adapts to the varying resource cost more quickly than the DVB-by-default algorithm. Hence, the MP algorithm outperforms the DVB-by-default algorithm.

Another thing we can see is that for the MBMS-by-default algorithm, it performs independently of the change of the resource cost in DVB. As a result, all the users are served to 3G/MBMS and the same system profit is generated as the resource cost of DVB increases.

For the MDS operation, it can be seen that the Delay Aware and the Timeout-based schemes perform the same for the MBMS-by-default algorithm in terms of the percentage of served users in both 3G/MBMS and DVB, as all the scheduled users can be served in 3G/MBMS. For the MP and DVB-by-default algorithms, when the traffic loads are distributed through both networks (i.e., when the resource cost is 5 for the MP algorithm, and the resource cost is 10 and 50 for the DVB-by-default algorithm), by batching more users in a transmission, the Delay Aware scheme assigns more users into DVB and fewer users into 3G/MBMS than the Timeout-based scheme.

We continue to study the performance of the overall average delay, as illustrated in Figure 6-24.
It can be seen that the average delay generated by all the MRM algorithms remains constant in relation to the resource cost of DVB. This is because three factors have impact on the overall average delay, namely the available resources, the resource cost and the MDS scheme. With reference to the former two factors, regardless of the MDS scheme, in the MUA scheme, users who can not be served in DVB due to the increased resource cost can nevertheless be served in 3G/MBMS which has constant resource cost and amount of available resources. Therefore, in this studied scenario, the MDS scheme is the only factor that affects the overall average delay.

We can see that regardless of the MUA operation, the Timeout-based MDS algorithm causes a shorter delay than the Delay Aware algorithm, because they use a different mechanism to decide the scheduling point. The Delay Aware algorithm batches users for a longer time than the Timeout-based algorithm.

In summary, we have seen that the resource cost has significant effect on the performance. The lower resource cost can lead to larger profits. Moreover, the traffic distribution throughout 3G/MBMS and DVB are affected as well. It is found that in the MUA operation, the MP algorithm can select the best network to serve users adaptively to the varying resource cost, so that users are always served in the network that gives high profit. Therefore, compared to the algorithms where the contents are always tried to be served through the default networks (i.e., 3G/MBMS or DVB), the MP algorithm improves the system profit. We also find that the Delay Aware MDS algorithm can further improve the performance by aggregating more users in a transmission.

6.5.3 Aspects of Cooperation Degrees when Performing Distributed Resource Management

As discussed in Section 6.3, a distributed resource management architecture is proposed based on the interworking of access networks. In such an architecture, the resource management functionality is distributed between access networks. The distributed resource control gives the promise of efficient sharing of network resources, but requires information exchange between networks. Since network operators are not expected to release their real time resource status, the commercial constraints on information availability will affect the type and amount of information that can be exchanged, the time scale at which information exchange is feasible, and thus the possible degree of cooperation.

This section aims to explore the possible degree of cooperation among multiple networks/operators when implementing a distributed resource management. In particular, this section looks at how the different degree of cooperation can affect the overall performance. Three
degrees of cooperation are presented and discussed, namely *non-cooperative*, *half-cooperative* and *fully-cooperative*. Finally, in this section, these cooperation degrees are evaluated by simulation. Since during the previous study, it was found that in general the combination of the Delay Aware algorithm in the MDS operation plus the MP algorithm in the MUA operation outperforms the other algorithm combinations in terms of the system profit and user loss probability, this algorithm combination is studied in the following sections.

### 6.5.3.1 Network Cooperation Degree

The MRM algorithms are expected to work in a distributed resource management architecture in which each underlying network is an independent administrative domain that not only performs its internal resource management mechanisms, such as admission control, QoS management, load balancing inside the network, but also coordinates with other networks to share the radio resources to a certain extent. The efficiency of the MRM algorithms, e.g., the Delay Aware plus MP algorithm, requires adequate information exchanging between networks, such as the resource cost and the achievable QoS level in each of the coordinated networks. However, since network operators are not expected to release their real time resource status, it is quite possible that these pieces of information that are necessary for the efficiency of the resource management may not be available or may be only partly available across interworked networks. So the commercial constraints on information availability will be considered in this section. The objective is to investigate the impact of the type and amount of information that can be exchanged across networks, on the system performance, such as the achievable system profit.

In the distributed resource management across multiple networks, the network that faces the users’ requests and originally provides the desired contents is called an originating network, while the other networks involved in the system are called cooperating networks. The originating network could be the cellular network or the broadcasting network.

To carry out the MP algorithm in the distributed resource management architecture, the originating network needs to exchange a ‘cost quote’ in real time with its cooperating networks via the gateway. This cost quote includes three types of parameters, namely the current resource usage, the achievable QoS, and respective resource cost in each cell of the cooperating network.

Based on whether resources are shared between networks, and furthermore whether the ‘cost quote’ is exchangeable between networks, three degrees of cooperation between networks are identified, namely non-cooperative, half-cooperative and fully-cooperative.

#### Non-cooperative MRM

The non-cooperative case acts as a reference case compared to other cooperation degrees. In this case, there is no cooperation between operators. Users can connect only to the originating network.
that they have been subscribed to. This case can be regarded as the current situation where the cellular operator and broadcasting operator work independently as different administrative domains, owning different customers and services. In each individual network, the MRM algorithm still applies. For the MDS operation, the Delay Aware batching scheme applies to determine the scheduling point, based on the information of subscribed users in the network. For the MUA operation, the MP scheme applies to assign users the network resources in terms of the transmission rate. Since there is no cooperation between networks, users can be assigned only to the originating network that they subscribe to. Therefore, the MP algorithm calculates the potential maximum profit based on the information of the originating network for content prioritization and user assignment. Two cases of non-cooperative MRM are studied based on the type of the originating network, namely non-cooperative-MBMS (i.e., 3G/MBMS is the originating network) and non-cooperative-DVB (i.e., DVB is the originating network).

**Half and fully Cooperative MRM**

In contrast to the non-cooperative case, the interworking of networks provides a platform enabling the operators cooperating with each other to some extent. More specifically, two levels of cooperation can be defined depending on how much information is exchangeable across networks to perform MRM algorithms introduced in previous sections.

- **Half Cooperative**: the originating network can serve contents through any of the constituent networks in the system. To schedule users’ requests, the user information including the location, terminal capability, preference and the delay tolerance is collected by the originating network. Nevertheless, because the composition of networks may be hindered by competition or disagreements between actors, the cooperating networks do not reveal the network information, such as the cost quote. Users are firstly assigned to the originating network and re-directed to one of the cooperating networks chosen randomly from available cooperating networks, whenever the load congestion happens. For the MDS operation, the Delay Aware batching scheme is applied to determine the scheduling point for the user assignment. Following this, in the MUA operation, the required contents are prioritised only based on the information available in the originating network, as there is no resource relevant information exchanged across networks. Two cases of half cooperative MRM are studied based in the type of the originating network, namely half-cooperative-MBMS (i.e., 3G/MBMS is the originating network) and half-cooperative-DVB (i.e., DVB is the originating network);

- **Fully Cooperative**: different from the half cooperative case, the fully-cooperative relationship allows operators to establish business/trust relations and exchange data for efficiently sharing resources. Therefore, the same information is available across
operators as within individual operators. Consequently, the Delay Aware and the MP algorithms can be implemented completely and effectively in a fully-cooperative situation by obtaining full information (i.e., the cost quote) from all the constituent networks. The fully cooperative case is expected to bring improved overall system profit and capacity utilization of constituent networks than other cooperative cases, due to unbiased and complete input information.

6.5.3.2 Performance Evaluation

6.5.3.2.1 Effect of Varying Available Resources in DVB

In Figure 6-25 and Figure 6-26, the profit and user loss probability are presented for all the cooperative cases respectively. Firstly, we study the performance when changing the available resources in DVB. We change the available resources in DVB from 512 to 3072 Kbps, and set the available resources in 3G/MBMS to 1792 Kbps.
In general, it can be seen that as the available resources in DVB increases, all the schemes achieve larger system profit and lower user loss probability, except for the non-cooperative MBMS scheme where the traffic can only be distributed into the 3G/MBMS network. However, the other schemes are eligible to utilize the resources in DVB if required.

It can be seen that the fully-cooperative case brings the highest overall system profit and the lowest user loss probability, while the lowest overall system profit and the highest user loss probability are generated in the non-cooperative case. These results are in fact expected. Moreover, in the half-cooperative case, the DVB-first assignment obtains similar performance as the fully-cooperative case, when DVB gets more resources available, such as 2560 Kbps. This is because when the available resources in DVB change, the fully-cooperative scheme is aware of the resource relevant information of all the involved networks. Hence, it is likely to distribute the most profitable contents into the best network, either 3G/MBMS or DVB. When the available resources in DVB increase, the most traffic will be distributed through DVB for most of the time. Therefore, it has a similar performance to the half-cooperative-DVB case where the DVB is the default network to serve the desired contents. When the available resources in DVB reduce, the fully-cooperative scheme is able to make a more efficient use of 3G/MBMS resources by serving the most profitable contents in 3G/MBMS. The result indicates that the fully-cooperative scheme is in favour of a dynamic environment where the available resources are varying over time. For instance, in the scenario where the multicast traffic shares and competes for limited available resources with other traffic, such as the TV program in DVB or voice service in UMTS, the fully-
cooperative scheme can distribute traffic into the best network adaptively to the varying resource availability.

In addition, it can be observed that in the half-cooperative case, the DVB-first assignment scheme outperforms the MBMS-first assignment scheme, especially when the available resource in DVB is high. When the available resource in DVB reduces, the MBMS-first scheme firstly approaches slowly, and then outperforms the DVB-first scheme, in terms of the system profit and user loss probability.

In the non-cooperative case, the 3G/MBMS network firstly generates higher profit and lower user loss probability than the DVB network due to more available resources. However, when the available resources in DVB increase, the DVB network generates more profits and lower user loss probability than the 3G/MBMS network.

To further investigate the impact of different degrees of the network cooperation, we measure percentage of served users in both networks, which are presented in Figure 6-27 and Figure 6-28 respectively.

![Figure 6-27 Percentage of served users in 3G/MBMS vs. available resources in DVB](image-url)
It can be seen that for the non-cooperative scheme, all the traffic is served through only one network. While for the other schemes, the whole traffic is distributed through both networks. The half-cooperative-MBMS scheme assigns the most traffic into 3G/MBMS. However, as the available resources in DVB increases, this scheme hands more traffic over to DVB from 3G/MBMS. On the contrary, the half-cooperative-DVB scheme assigns more traffic into DVB, and less traffic into 3G/MBMS. With regard to the fully-cooperative scheme, we can see that this scheme almost serves the same amount of traffic into DVB as the half-cooperative-DVB scheme does; however, more traffic is served in 3G/MBMS. This is because the fully-cooperative scheme will first try to put contents into the DVB that is more resource efficient if it has enough resources. Otherwise, the 3G/MBMS will be used. This selection policy is similar to the policy used by the half-cooperative-DVB scheme. Hence, they have similar performance in terms of the percentage of served users in DVB. However, since the fully-cooperative scheme is aware of the resource information in the 3G/MBMS network, therefore, it makes a more efficient utilization of the available resources in 3G/MBMS than the half-cooperative-DVB scheme, and hence serving more users.

Varying the available resources in 3G/MBMS network is expected to have the same impact as varying the available resources in DVB network on the performance of different degrees of network cooperation. This is because for both cooperative and non-cooperative cases, we have studied two schemes where both 3G/MBMS and DVB are considered as the originating network to transmit the desired contents by default.
6.5.3.2.2 Effect of Varying Resource Cost of DVB

We continue studying the impact of different degrees of network cooperation in the scenario where the resource cost in DVB varies, while the resource cost in 3G/MBMS is fixed at 0.5. In order to maximize the impact of varying resource cost on the performance, we set the available resources in both networks as sufficient as possible to accommodate all multicast traffic. Therefore, user requests are lost only because they are not willing to wait until enough users are aggregated to compensate the resource cost that is consumed for serving these users. In Figure 6-29 and Figure 6-30, we show the overall system profit and user loss probability when varying the ratio of the resource cost of DVB with respect to that of 3G/MBMS. To help to explain the results, we show in Figure 6-31 and Figure 6-32 the percentage of served users in 3G/MBMS and DVB with respect to the total users in the system respectively.

![Figure 6-29 Overall system profit vs. resource cost ratio](image-url)
Figure 6-30 Overall user loss probability vs. resource cost ratio

Figure 6-31 Percentage of served users in 3G/MBMS vs. resource cost ratio
It can be seen from Figure 6-30 that no schemes lose any users from the systems, except for the non-cooperative-DVB scheme. This is because when the resource cost of DVB increases, users have to wait for a longer time to aggregate enough users. Therefore, some users might renege from the system and are lost. However, for the half-cooperative and fully-cooperative schemes, the 3G/MBMS is an alternative network to be used. Hence, users will be handed over to 3G/MBMS and are served eventually. For the same reason, it can be seen that the non-cooperative-DVB scheme obtains smaller system profit as the resource cost of DVB increases. However, since the resource cost of 3G/MBMS does not change, the non-cooperative-MBMS scheme achieves constant system profit regardless of the change of the resource cost of DVB.

In addition, it can be seen that the fully-cooperative scheme outperforms the other schemes in terms of the overall system profit. Moreover, the fully-cooperative scheme is able to be adaptive to the varying resource cost of DVB. It traces out the outer "envelope" of the two half-cooperative schemes. When the resource cost of DVB is small, the non-cooperative-DVB scheme and the half-cooperative-DVB scheme obtain higher profit than the non-cooperative-MBMS scheme and the half-cooperative-MBMS scheme, because the DVB usually consumes fewer radio resources than 3G/MBMS when serving the same users. Since the fully-cooperative scheme is aware of the resource cost of DVB and its efficiency of serving multicast services, more users are assigned to DVB, thereby giving the similar performance as the half-cooperative-DVB scheme. As the resource cost of DVB increases, more users that originally attempt to be assigned to DVB are blocked because it becomes more difficult to compensate the increasing resource cost. Hence, the
non-cooperative-DVB scheme and the half-cooperative-DVB scheme obtain lower profit and larger user loss probability. However, the profit of the half-cooperative-MBMS and non-cooperative-MBMS schemes keep constant in relation to the resource cost of DVB. But in the end, the half-cooperative schemes obtain the same performance as shown in Figure 6-29, because all users will be assigned to 3G/MBMS due to the high resource cost of DVB. Since the fully-cooperative scheme is aware of the increasing of the resource cost of DVB, it accordingly assigns users to 3G/MBMS that causes smaller resource cost. Hence, the fully-cooperative scheme obtains similar performance as the half-cooperative-MBMS scheme as the resource cost of DVB increases.

In conclusion, this chapter has investigated the overall multicast resource management algorithm that consists of Multicast-on-Demand Scheduling algorithm and Multicast User Assignment algorithm. A general resource management framework has been proposed, including the resource management triggering function, the service scheduling function for MDS and the network and QoS selection function for MUA. This resource management framework enables the flexible adoption of different schemes to achieve different optimization objectives, which are defined by the operators who implement the resource management mechanism. A distributed resource management architecture has been investigated to implement the resource management mechanism. In addition, simulation was conducted to evaluate the proposed resource management algorithms. They are composed of different MDS algorithms (i.e., Delay Aware batching scheme, Timeout-based batching scheme) and MUA algorithms (i.e., MP assignment scheme and network-by-default-based assignment scheme). It was found that the combination of Delay Aware batching scheme plus the MP assignment scheme could give the best performance in terms of the overall system profit and user loss probability. Moreover, the combination of Timeout-based batching scheme plus the MP assignment scheme can achieve the best service set-up delay while giving a good performance of the system profit and user loss probability. Finally, the impact of cooperation degree between networks on the performance of resource management algorithm was investigated. It was found that through full cooperation between networks, the best performance in the system profit and user loss probability can be achieved.
Chapter 7

7 Conclusions and Future Work

The concept of beyond 3G envisions heterogeneous infrastructures comprising diverse wireless systems e.g., 2G, 3G, DVB, WLAN, and hence enabling various transmission approaches, e.g., unicast, multicast and broadcast, in a cooperative and complementary manner. In addition, mobile network operators have started exploring effective methods to extend the existing networks with multicast capability due to its transmission efficiency, such as 3G/MBMS. Additionally, standardization efforts have also been spent in broadcasting networks for the DVB-H and IPDC within the DVB project. This gives another choice for delivering multicast services to mobile users. Intensive research has been carried out in the subjects of beyond 3G and wireless multicast. Out of these research studies, resource management is one of the most important subjects that are still being investigated. This thesis gives a thorough study in the multicast resource management for heterogeneous wireless networks for the first time.

In the multicast resource management for heterogeneous wireless networks, the work set out here addressed two major problems, namely the MDS (multicast-on-demand scheduling) problem and the MUA (multicast user assignment) problem. The first problem is to determine “when” desired content is eligible for scheduling. The second problem is to select the most appropriate access network and transmission rate for the delivery of the desired multicast content for each eligible multicast group.

In the remainder of this chapter, all the achievements during this research are summarized. In addition, the direction for future work is suggested.

7.1 Summary of Achievements

The research presented here can be broken down into following three main aspects:

Multicast Resource Management Algorithm Development

The first achievement of the work here was the development of the MRM algorithm for cost-effective multimedia service provisioning over heterogeneous networks. The algorithm was based upon two main blocks, comprising the MDS and the MUA. Within each of these blocks, there
were several different sub-algorithms developed and evaluated. Therefore, in total, there were a large variety of permutations of the overall algorithm proposed to perform the MRM.

The Multicast-on-Demand Scheduling algorithms were developed for both loss and delay systems. A generic control procedure of algorithm operation has been developed for each system model respectively. We derived an analytical model for a timeout-based scheme in the loss system. Based on analytical results, we found that an optimal timeout threshold exists to maximize the percentage of number of served users in the system, while adhering to the constraints of user blocking probability and user reneging probability. Based on this finding, an optimal timeout-based scheme was proposed to maximize the percentage of the served users in the system, in a dynamic wireless environment where the system status, such as the request arrival rate and the available resources are varying over time. Different from traditional static schemes, this optimal scheme is able to adapt the timeout threshold to the system status. Simulation results showed the proposed optimal scheme can improve the percentage of the served users up to 60%, compared to traditional schemes when the system status varies. In addition, we re-evaluated the traditional schemes, but this time with the wireless resource limitation being considered. Our results showed that when the request arrival rate changes, the basic timeout-based scheme outperforms the other traditional schemes in terms of the percentage of served users in the system, while the combined-for-loss is better than the other traditional schemes when the available resources change.

To cater for the future wireless networks, we explored efficient MDS algorithms suitable for heterogeneous networks. Novel schemes have been proposed. Different from traditional schemes that deal with the MDS in a multimedia server without knowledge regarding underlying deliver networks, the proposed schemes take inputs from both multimedia servers and underlying networks when scheduling multicast contents. Furthermore, the rate adaptation technique was utilized to further improve the efficiency of resource utilization. The performance was evaluated in two simulation scenarios, with varying users' delay tolerance and varying resource costs respectively. The simulation results showed the proposed MDS scheme can archive up to 33% improved system profit, 100% improved user loss probability and 22% improved service set-up delay in comparison to traditional schemes.

The problem of Multicast User Assignment in Heterogeneous Wireless Networks (MUA-HWN) was mathematically formulated as an optimization problem. The solution to this problem is to determine the best combination of access networks along with the transmission rates to serve required multicast contents to interested groups of users simultaneously, so as to maximize the system profit whilst satisfying the QoS requirements of the contents and respecting wireless resource constraints. The inputs to the problem include content profiles, user profiles and the network status. The outputs include the system profit, the access network and the transmission
Chapter 7. Conclusions and Future Work

rate chosen for each user. A generic concept of ‘profit’ was adopted in our research in order to accommodate different measures, such as network throughput, number of served users, and monetary profit. The MUA-HWN problem was proved to be NP-hard, and thus, the computational complexity for finding the optimal solution increases exponentially with the size of the problem. To develop efficient algorithms for practical systems, several approximate algorithms were proposed to find sub-optimal solutions. Rather than using a user-centric control technique that has been widely used in traditional algorithms to solve the problem, we adopted network-centric control technique that entitles the network operators to calculate the outputs of the MUA-HWN problem, by providing operators with all the inputs. Simulation results revealed the superior performance of our schemes over traditional schemes in increasing the overall system profit up to 300% and reducing the overall user loss probability up to 94%. Furthermore, our algorithms perform consistently better under a wide variety of traffic loads, number of available networks, resource costs and user terminal type distributions.

Furthermore, we developed a multicast resource management algorithm framework that in general consists of MDS and MUA algorithms, in order to achieve efficient multicast-on-demand service provisioning in heterogeneous wireless networks in real time. This framework was designed to be flexible enough to accommodate different sub-algorithms so as to achieve different optimization objectives. In general it consists of four parts, namely batching, queue identification, queue prioritization and user assignment. By using different schemes in each step, different objectives can be achieved. Based on the resource management framework, different combinations of sub-algorithms have been evaluated by simulation. For batching and queue identification, it was found that the Delay Aware scheme that utilizes the accurate information of users’ delay tolerances outperforms the Timeout-based scheme that is only aware of the minimum user delay tolerance, in terms of system profit and user loss probability, especially when combining with efficient MUA algorithms, but paying the price of longer service set-up delay. In addition, different schemes for queue prioritization and user assignment were investigated. It was found that the Maximum Profit scheme outperforms the other schemes in terms of system profit, because this scheme considers potential profits and resource costs as the basis of decision-making.

Multicast Resource Management Architecture and Implementation

A distributed resource management architecture was proposed to respect the sovereign of networks involved, to efficiently utilize resources of underlying networks, to minimize the alterations to the existing infrastructure and to enable networks cooperating in a secure and fair manner. We devised a new entity, called resource manager for each constituent access network. The resource manager of a network communicates and exchanges resource-related information with its counterpart residing in another network, to coordinate different networks’ resource usage without having a centralised entity. Three functional blocks were designed to perform the
Chapter 7. Conclusions and Future Work

respective resource management algorithms, namely, RM triggering function, service scheduling function and network and QoS selection function. These functions are coordinated in the resource manager in each underlying networks. In addition, we have explored the procedure of the multimedia-on-demand service provisioning over multiple networks. Distributed resource control gives the promise of efficient sharing of networks resource, but requires information exchange between networks. We investigated the underlying signalling issues required for the distributed resource management, focusing particularly on the requirements of control information flow between different resource management entities and its practicality. In addition, the viability of the resource management mechanism developed in this work was evaluated in a real testbed. Several experiments were conducted to verify the correct operation of the resource management functionalities.

Cooperation Degree Investigation

The proposed resource management algorithms require information sharing between networks. This implies that in a distributed resource management architecture, implementing resource management algorithms requires a certain amount of information exchanges between networks. Since network operators are not expected to release their real-time resource status, the impact of the information availability was investigated. Three degrees of cooperation between networks were identified, namely full-cooperation, half-cooperation, and non-cooperation. Through full cooperation between networks, network operators are allowed to establish business/trust relations and exchange data for efficiently sharing resources. However, in the case of half-cooperation, operators do not reveal or exchange network information, but share the overall resources in the system. When there is no cooperation between networks, different network operators work independently as different administrative domain, owning different customers and services, and the resources in all the networks cannot be shared. It was found from the simulation results that the full-cooperation of networks can improve the system profit up to 41% and 114% in comparison to the half cooperation and the non-cooperation, respectively.

7.2 Future Work

Future work that could be carried out into the subject of multicast resource management may focus on three aspects.

- Refinement and improvement of the models used in the research presented in this thesis;
- Investigation of different multicast resource management methods;
- Practical requirements for implementing the multicast resource management mechanisms.
Chapter 7. Conclusions and Future Work

7.2.1 Refinement and Improvement of Models

The main aspect that could be refined and improved in this work is the traffic models and cost model that calculates the resource cost used in the simulations. It has been shown that one of advantages of the proposed multicast resource management algorithms lies on the fact that the system performance can be optimized in terms of system profit and user loss probability, taking into account the traffic information and resource costs of underlying access networks. Since our research is the first study in a new area, the traffic model and the cost model that is used to calculate the resource cost are two best guesses at what might be seen in reality. Therefore, there is scope for further research into the evaluation of different patterns of time-varying traffic load demands and the cost of accommodating the traffic load. Furthermore, the utility function used in the simulations could be enhanced. For instance, more utility functions could be examined in order to accommodate different measure. When the utility function is used to represent user's perceived satisfaction rating of a certain service with a certain transmission rate, this subjective parameter could be obtained through experiments with real users; another simulation parameter, user tolerance to the service set-up delay could also be obtained in the same manner. The accuracy of modelling the UMTS system could also be improved, such as modelling an interference limited capacity, in order to precisely observe the impact of the radio resources on the resource management operation.

7.2.2 Investigation of Different Multicast Resource Management Methods

In the research presented in this thesis, we introduced two research issues with respect to the multicast resource management, namely multicast-on-demand scheduling and multicast user assignment, focusing on service-level scheduling and traffic distribution in heterogeneous wireless networks respectively. In terms of new multicast resource management methods, it would be interesting to investigate the packet-level scheduling techniques in multicast-enabled wireless networks. Different from service-level scheduling, packet-level scheduling is performed in a shorter-term and is more easily influenced by the wireless channel variation and traffic burstiness. It has been proven in traditional packet scheduling techniques for unicast services in UMTS that the real-time channel state information can give significant performance in terms of system throughput. To obtain this information, a return link is required from a user to the network so as to report the channel information from time to time. In a multicast transmission, the transmission of feedback from users could exhaust the radio resources in the uplink due to the potential large group size. Therefore, the impact of both the dynamic channel variation and radio resource consumption in the uplink need to be considered when designing a packet-level scheduling for multicast transmission. This can be further investigated and then integrated into the service-level scheduling studied in this work.
7.2.3 Practical Requirements for Implementation

As for the practical requirements for implementing the proposed multicast resource management schemes in real systems, the signalling load is one of the major concerns. It would be an interesting issue to evaluate the potential signalling overhead of the proposed multicast resource management schemes, and further fine-tune them so as to achieve good system performance on one hand, while maintaining the signalling overhead under a satisfactory level on the other hand.
Publications

During the course of the research presented here, the following journal and conference papers have been submitted or published. A list of these publications is as follows.


References


[7]. 3GPP, “Multimedia Broadcast/Multicast Service, Stage 1,” TS 22.146, v6.6.0, Sept. 2004


[10]. “IPDC in DVB-H: Technical Requirements” CBMS1026 v1.0.0 Rev.1/TM 3095 Rev. 2


[21]. dea.brunel.ac.uk/project/Cismundus/home.htm

[22]. www.comnets.rwth-aachen.de/~o_drive/

[23]. www.ist-drive.org


[62]. www.opnet.com


[68]. A. Gluhak, “Specification and evaluation of a group management support for heterogeneous wireless networks,” MVCE Core 3, D-IoN-5.2, November, 2005


[71]. Implementation Guidelines for DVB-H Services - TR 102 377

[72]. Lucas Pun, "Introduction to Optimization Practice," John Wiley & Sons Inc


[84]. ETSI, "Digital Video Broadcasting (DVB); Implementation guidelines for DVB terrestrial services; Transmission aspects," TR 101 190, v1.1.1, December, 1997.


[87]. www.dvb-h-online.org


[92]. 3GPP TSG RAN WG1 #28, R1-021239, "MBMS power usage," Espoo, Finland, October 8-9, 2002.


A Simulator Description

A.1 OPNET Introduction

We have evaluated the proposed multicast resource management algorithms, by using OPNET [62]. OPNET provides a comprehensive development environment supporting the modelling of communication networks and distributed systems. Both behaviour and performance of modelled systems can be analyzed by performing discrete event simulations. The OPNET environment incorporates tools for all phases of a study, including model design, simulation, data collection, and data analysis.

OPNET are structured hierarchically, in a manner that parallels real network systems. Specialized editors address issues at different levels of the hierarchy. This provides an intuitive modelling environment and also permits re-use of lower level models. Three modelling domains essentially span all the hierarchical levels of a model, namely Network, Node and Process. These modelling domains are demonstrated in Figure A-1.

The network domain's role is to define the topology of a communication network, and describes in terms of subnetworks, nodes, links and geographical context. A network model may contain any number of communicating entities called nodes. Nodes are instances of node models.

The node domain provides for the modelling of communication devices that can be deployed and interconnected at the network level, and describes in terms of functional elements and data flow.
between them. Node models consist of modules and connections. Modules are information sources, sinks, and processors. Some modules have predefined behaviour; some modules are programmable via their “process model”. Connections allow information to flow between them.

Process models define behaviour for programmable modules. A process is an instance of a process model and operates within one module. A process can in many ways be thought of as similar to an executing software program. Behaviour of processes (protocols, algorithms, applications) is specified using finite state machines and extended high-level language.

An important feature of OPNET is that OPNET simulations are based on a discrete-event modelling approach, where the progress of the model over simulation time is decomposed into individual points where change can take place. The OPNET term for each such point in time is an event. When an event is executed, a corresponding action is taken accordingly and the state is updated. The following lists some common actions that are modelled as events:

- Receipt of a message (i.e., a packet) or a command by a process;
- Expiration of a timer (e.g., while waiting for an acknowledgement);
- Indication of availability of a resource;
- Indication of completion (or partial completion) of a task by a resource;
- Change in a statistic that is monitored by a process;
- Start or end of a packet’s transmission/reception on a link;
- Generation of a new message, command, or task by an application process;
- Failure (or recovery from failure) of a device.

Each time that a new event occurs, it is said to be executed by the simulation. As new events are executed, simulation time increases monotonically. Because OPNET supports modelling of distributed systems, it must allow multiple events to occur simultaneously (in terms of simulation time), and to affect different components of the system. Thus, it is possible for any number of distinct events to occur at the exact same simulation time.

OPNET simulations manage events with an event list. The purpose of maintaining an event list is to make sure that all events are executed in the proper time order. Each event has an associated time at which it is specified to occur. The requesting of new events, which requires specification of the simulation time at which they must execute, is called event scheduling. Events may be scheduled for any future time, or for the current time, but never for past times. The simulation event list maintains all of the events in time-order so that the next event may be readily determined when the current one has completed execution. The event list grows and shrinks as
new events are inserted and events are executed over the course of the simulation. At the start of the simulation certain pre-defined events are inserted in the event list to allow execution begin. If the event list becomes empty at any point, the simulation completes. Events are managed by the Simulation Kernel in the event list. New events are inserted into the list by the Kernel. The Kernel extracts the events from the event list and executes them in time-order.

A.2 Simulator Diagram

The simulator diagram used in our study is illustrated in Figure A-2.

![Simulator Diagram](image)

**Figure A-2 Simulation block diagram**

The simulator comprises the mobility module, the traffic module, the network module, the service scheduling module, the user assignment module and the performance calculation module that calculates the outputs. In the network module, we considered an interworked DVB and UMTS network infrastructure. They are overlapping in a common service area. Each of the networks is made up of a certain number of cells. In each cell, the available radio resources in the downlink are limited. The radio resource limitation in the uplink is not simulated. The resource cost of a network is randomly generated. In the traffic model, a set of multicast contents are modelled. Each of the contents is characterized by a set of discrete bandwidths, the service duration, and the arrival rate of interested users' requests. The content characteristic varies randomly according to a certain probability statistics. For instance, the content duration and the request arrival rate are determined by the Exponential and Poisson distribution respectively. Users' reneging behaviour is also modelled, i.e., after waiting a certain period of time, a user will quit the service request and leaves from the system at any time. In addition, we considered three types of user terminals, namely, multimode, single-mode with the cellular network interface and single-mode with the broadcasting network interface. In the mobility mode, we assume users are uniformly distributed.
in the service area that covers a certain number of cells. A user has a mean velocity and the movement direction is uniformly distributed over $[0, 2\pi]$. The residing time of a user within a cell is exponentially distributed [99].

The user profile (e.g., delay tolerance, content preference, terminal mode, and location) and the content profile (e.g., content bandwidth and duration) are generated from both the mobility module and the traffic module. The network status parameters (e.g., available radio bandwidth in each cell and the resource cost of the unit transmission data) are generated from the network module. The studied MDS algorithms are implemented in the service scheduling module by taking inputs from user and content profiles. In this module, users’ requests are batched in different queues and wait for the resource allocation that will be carried out by the MUA algorithms in the user assignment module. After the eligible contents for the resource allocation are determined, the service scheduling module triggers the user assignment operation by sending a MUA request message that contains the parameters relevant to the eligible contents (e.g., the content bandwidth) and remaining users interested in these contents (e.g., the content preference, terminal mode, and location). Taking into account these parameters as well as the network status, the MUA algorithms in the user assignment module are performed to select the best cells and the transmission bandwidths to serve users that have been picked up from the service scheduling module. Moreover, this module deals with the bandwidth adaptation for handed-over users. As a result of the selection, the remaining radio resources in the chosen cells will be updated accordingly to reflect the decision-making. In addition, after the service is completed, the radio resources in respective serving cells are also updated accordingly. After performing the MUA algorithms, the user assignment module sends back with a MUA reply message to the service scheduling module. This message will tell which user is able to be served with the eligible contents. Subsequently, these users’ requests are deleted from the queue.

**A.3 Simulator for Multicast User Assignment**

A Monte Carlo based simulator has been constructed using OPNET. Statistics are collected over multiple Monte Carlo runs. Each run represents a snapshot of the system status. During each run, algorithms are performed to assign users into suitable cells.

**A.3.1 Network Model**

The exact architecture of any future B3G is scalable but hard to predict because the B3G concept envisions a heterogeneous infrastructure comprising diverse wireless systems that could potentially include new systems in the future. Therefore, we propose a concept of Basic Cell Unit.
(BCU) on which the B3G system is constructed. Each BCU represents a certain geographic area. It is assumed that each cell in each of the different access networks has an area of an integral number of BCUs. This concept is illustrated by an example in Figure A-3.

Two types of operators, $n_0$ and $n_1$ are considered. Our study does not depend on the exact structure of the networks. For the sake of the results' explanation, it is assumed that $n_0$ operates a cellular network (e.g., UMTS) with seven cells and $n_1$ operates a broadcasting network (e.g., DVB-H) with one cell. These networks are overlapping in a common service area $A$. Each cell of $n_0$ has an area of one BCU and each cell of $n_1$ has an area of seven BCUs. Their capacity limits are uniformly generated between 0 and 1024 kbps per cell, and between 0 and 2048 kbps per cell respectively. Since there is no standardized method to calculate the resource cost $r_c$ in a certain network, $r_c$ is randomly generated from 0 to $c$ monetary cost$^3$ per Kbits of transferred data for each operator, where $c$ is the maximum resource cost and is determined by the operator. We assume that the broadcasting network has a higher resource cost than the cellular network. This is because that the broadcasting network is less spectrally efficient than the cellular network in terms of the amount of radio spectrum it requires to send different data across several cells [106]. Unlike cellular network that can re-use the same 5MHz frequency across several cells in an area, broadcast network however has to use another 8MHz frequency to transmit different data in different cells. The higher power consumption in broadcasting transmitter than that in cellular transmitter is another cause resulting into the higher broadcasting cost than cellular cost [105]. Therefore without the loss of generality, we assume that given the maximum resource cost of cellular network $c_c$, the maximum resource cost of broadcasting network $c_b$ is directly proportional with $c_c$: $c_b=ac_c$, where $a\geq 1$.

Table 4 presents two Network Deployment (ND) cases studied in terms of the number of operators.

---

$^3$ The monetary cost has no unit without loss of generality. Any unit of currency is applicable (e.g., cent, penny, pound or dollar).
Appendix A. Simulator Description

Table 4 Network deployment cases

<table>
<thead>
<tr>
<th>ND Case</th>
<th>Number of operators $n_0$</th>
<th>Number of operators $n_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ND_Case 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ND_Case 2</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

A.3.2 Traffic Model

The content bit rate is randomly generated from the set \{32, 64, 128, 256, 384\} kbps. The non-zero price of each content is set to 1 monetary price in unit transmitted data (e.g., penny per Kbits). That is a pay-per-view fee (PPV) charged for every served user. Note that different prices can be assumed for different contents based on the popularities. Users are uniformly distributed over the coverage of service area A. Each user requests for a particular content according to a probability given by the Zipf distribution [67]. The average number of requests for a content is 10 per BCU area. Three types of user terminals are considered in the simulation, namely multimode, single-mode with the cellular interface and single-model with the broadcasting interface. Furthermore, we study three Terminal Heterogeneity (TH) cases that characterize the probability distribution of user terminal type, as shown in Table 5.

Table 5 Terminal heterogeneity cases

<table>
<thead>
<tr>
<th>TH Case</th>
<th>Multimode</th>
<th>Broadcasting</th>
<th>Cellular</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>single-model</td>
<td>single-model</td>
</tr>
<tr>
<td>TH_Case 1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TH_Case 2</td>
<td>0.5</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>TH_Case 3</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

A.3.3 Performance Measures

The main measure of performance recorded in the simulations is the system profit, user loss probability, the percentage of served user in cellular and broadcasting networks over all the users in the system respectively. The user loss probability is the percentage of lost users over all users in the system. A lost user is defined as one that is blocked at service set up. Also, we measure the system profit improvement of a algorithm A over the comparative algorithm B as

$$\text{System Profit Improvement} = \frac{\text{profit}_A - \text{profit}_B}{\text{profit}_B}$$

(A-1)
We assume the utility function $u(r)=wr$ when measuring the system utility, where $r$ is the received transmission rate at the user's terminal; $w$ is a weight factor. As mentioned before we consider scenarios, with and without resource cost charged by operators. In the case of without considering the cost, the objective of the system optimization could be maximizing the system throughput, which is the objective of the traditional resource management. So, the weight factor is set to 1 (unitless). The objective of maximizing the system throughput can be easily modified to be the objective of maximizing the revenue of network operators, by considering the price of the desired contents. Therefore, the weight is the price of a content. When the resource cost is considered, the objective of the system optimization turns out to be maximizing the monetary profit, which is the function of the system throughput, the price, and the resource cost. It is derived as follows:

Suppose there are $I$ served users, and $J$ radio multicast channels used for serving those users. The $i$th user is served with the rate of $r_i$. The resource cost of the $j$th multicast channel is $b_j$ cost per unit of transmission data, and the rate is $r_j$. Given $p$ the price of each multicast content in unit transmission data, $\theta$ the profit earned in a unit time, $\theta$ is derived as follows:

\[
\theta = \sum_{i=1}^{I} p r_i - \sum_{j=1}^{J} b_j r_j
\]  

(A-2)

\[
\frac{\theta}{p} = \frac{\sum_{i=1}^{I} r_i - \sum_{j=1}^{J} b_j r_j}{p}
\]  

(A-3)

In the equation (A-3), $\theta/p$ is the normalized profit to the content price (in the unit of bps). $\sum_{i=1}^{I} r_i$ is the original system throughput without considering the resource cost. $\sum_{j=1}^{J} b_j r_j/p$ can be regarded as the reduced system throughput when the resource cost is considered. Therefore, the normalized profit can be used to represent the system throughput with and without cost. Clearly, a resource management scheme is profitable if $\sum_{i=1}^{I} p r_i > \sum_{j=1}^{J} b_j r_j$.

In our simulation results, the normalized profit is used to represent the system throughput with and without the cost, which we call the overall system profit. Note that the equation of the normalized profit assumes that the price is the same for every content. The concept of the normalized profit was proposed in [55], in order to minimize the impact of the content price, instead to focus on the impact of the resource cost and traffic load on the system performance, i.e., the system profit. In the case of considering particular prices for different contents, the effective MUA scheme should be aware of their prices when prioritizing these contents.
Appendix A. Simulator Description

A.4 Simulator for Combined Batching and Multicast User Assignment

To evaluate the performance of the proposed resource management algorithms for multicast service delivery, a dynamic simulator has been constructed with OPNET. A number of users' requests are generated randomly from the call generator. The period of the residing time of a user within a cell is also randomly generated. More details about the simulation models are detailed below.

A.4.1 Network Model

The overlapping networks consist of a MBMS-enabled UMTS network with seven cells and a DVB network with one cell. Traditionally DVB is designed to distribute high quality multimedia data to a large group of users, it therefore supports a high data rate. For instance, in a DVB multiplex, 15Mbps bit rate is possible for mobile environment at 8MHz bandwidth covering an area ranging from 1 km radius to 30 km radius, which can support roughly 120 streams of 128 Kbps at the same time [69]. In practice, the whole capacity of 15 Mbps in a multiplex is impossible to be available only for multicast-on-demand contents, considering the other traffic loads, such as TV program streams. To be more realistic, we assume the multicast contents use a part of DVB capacity, for instance, changing from 512 to 3072 Kbps. Therefore, a total of 4-24 simultaneous multicast streams of 128 Kbps could be allocated in a DVB cell.

For 3G/MBMS, it is worth mentioning that to support MBMS, UMTS does not use fast power control mechanisms to determine the transmission power for each active receiver as used for unicast services. Instead, the proportion of the available Node B power allocated to a particular session with a certain bandwidth requirement is essentially fixed to provide certain coverage. To support MBMS the transmitted power is set for each bit rate to cover a certain percentage of the geographical locations within a cell. These power settings have been obtained from link level simulations [66]. It shows that the power requirement of a PtM MBMS transmission as a fraction of the cell power for different data rates, geometries (i.e. cell coverage) and error rate targets and as a function of physical layer configurations (TTI value & transmit diversity mode) and propagation environment (i.e. radio channel models).

It has been investigated that based on the functionality currently available in 3GPP UTRAN Release-5, 28% of the total base station power would be required for providing 64kbps PtM MBMS in a macro cellular environment [66] [70]. This can be reduced to 4-12% through the introduction of new techniques in the Release-6 physical layer, such as macrodiversity combining that includes selective combining and soft combining. These two techniques can significantly reduce the transmit power requirement for one 64kbps MBMS channel to 9.3%-12% and 4.6%-
7.6% of the Node-B power respectively, in order to achieve 95% coverage at 1% BLER. Thus, assuming 80% of the cell power is being available for MBMS, the selective combining and soft combining allow for a total 3G/MBMS cell capacity of around 426-550 Kbps and 673-1113 Kbps respectively, while it only allows a total 3G/MBMS cell capacity of 183 Kbps without macrodiversity. Moreover, plus receiver diversity techniques, a total 3G/MBMS cell capacity of 1066-1764 Kbps is achievable. Also, simulation results have indicated that the provision of up to 256 Kbps PtM MBMS is likely to be viable with Release-6. More details can be found in [66].

It should be noted that, at the time of writing, the precise MBMS macrodiversity mechanisms to be included in UTRAN Release-6 is still under discussion. We do not intend to further investigate the technologies in physical layer, which improves the resource efficiency for MBMS services. Instead, we try to use the latest knowledge as we can obtain, to specify the 3G/MBMS capacity used in our simulator. To cover a wide range of practical achievable 3G/MBMS cell capacity, we study two cases of the capacity, such as, 1000Kbps and 1792 Kbps.

Two scenarios, with and without resource cost are studied respectively. Note that as discussed in the Appendix A.3, the resource cost represents the economy cost per unit of transmitted date, which differs in different types of systems depending on the access technology, spectrum band, deployed infrastructure etc.

In summary, it can be seen from the simulation parameters that the DVB generally has bigger capacity than 3G/MBMS, and thus potentially able to accommodate more simultaneous streams in one cell. Furthermore, considering that DVB consumes less radio channels to serve vast audiences in a certain area than 3G/MBMS does, because a DVB cell covers bigger geographical area than 3G/MBMS, it is expected the cooperation with DVB could bring significant benefits to 3G/MBMS, by sharing a portion of MBMS traffic loads with DVB. On the other hand, 3G/MBMS can offer interactive/on-demand services and effective location management within one radio technology and localised service area. Furthermore, given the higher resource cost of DVB due to the less efficiency of spectrum usage, and higher power consumption, it could be more profitable to use 3G/MBMS than DVB.

**A.4.2 Traffic Model**

There is a set of multicast contents that are offered to users in the system. Each content has a maximum rate of 256 Kbps. The non-zero price of each content is set to 1 monetary price per Kbits of transmitted data. The set of multicast contents might be delivered through any of the networks, determined by the resource management algorithms for the purpose of cost-efficient service provisioning. These contents could be owned by one network operator (3G/MBMS or
DVB), or could be offered by a trusted third party that has a business relationship with both 3G/MBMS and DVB operators.

Requests for contents are generated according to a Poisson distribution, randomly distributed in each cell of the network. The probability of choosing a specific content offered by a network is given by the Zipf distribution [67].

A deterministic reneging model is used in our study [57]. In this model, after waiting a certain period of time given by the waiting threshold, the user's remaining time until departure is determined by an Exponential distribution with a mean waiting time.

### A.4.3 Mobility Model

A fluid flow mobility model is adopted [99]. In this model, the mobile terminals are uniformly distributed over a given cell. Mobile terminal have a mean velocity of $V$ and their direction of movements are uniformly distributed over $[0, 2\pi]$. The residing time of a mobile within a cell is exponentially distributed and the mean dwell time can be estimated by $1/\eta$, where $\eta$ is the cell crossover rate. The crossover rate for the circular cell of the radius $r$ is given by:

$$\eta = \frac{2V}{\pi r}$$

(A-4)

The simulation parameters are listed in Table 6.

<table>
<thead>
<tr>
<th>Simulation Parameter</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multicast contents request arrival rate</td>
<td>3000 requests/hour</td>
</tr>
<tr>
<td>Number of contents</td>
<td>10</td>
</tr>
<tr>
<td>Contents QoS range</td>
<td>256, 128, 64, 32, 16 Kbps</td>
</tr>
<tr>
<td>Mean duration</td>
<td>5 minutes</td>
</tr>
<tr>
<td>3G/MBMS cell radius</td>
<td>1400m</td>
</tr>
<tr>
<td>3G/MBMS cell capacity</td>
<td>1000, 1792 Kbps</td>
</tr>
<tr>
<td>DVB-H MUX capacity</td>
<td>512–3072 Kbps</td>
</tr>
<tr>
<td>Resource cost of 3G/MBMS</td>
<td>0, 0.5 cost per Kbits</td>
</tr>
<tr>
<td>Resource cost of DVB</td>
<td>0–100 cost per Kbits</td>
</tr>
<tr>
<td>Mobility speed</td>
<td>3 km/hour</td>
</tr>
</tbody>
</table>
Appendix A. Simulator Description

<table>
<thead>
<tr>
<th>Mean waiting time</th>
<th>1 minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waiting threshold</td>
<td>1~11 minutes</td>
</tr>
</tbody>
</table>

**Table 6 Simulation parameters**

The default simulation parameters are given in Table 7, otherwise specific stated.

<table>
<thead>
<tr>
<th>Simulation Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3G/MBMS cell capacity</td>
<td>1000 Kbps</td>
</tr>
<tr>
<td>DVB-H MUX capacity</td>
<td>1000 Kbps</td>
</tr>
<tr>
<td>Resource cost of 3G/MBMS and DVB</td>
<td>(0, 0)</td>
</tr>
<tr>
<td>Waiting threshold</td>
<td>2 minutes</td>
</tr>
</tbody>
</table>

**Table 7 Default simulation parameters**

In line with the simulator in Appendix A.3, the normalized system profit and the user loss probability are adopted as the major performance metrics.
Appendix B. Aggregate Resource Consumption

B Aggregate Resource Consumption

A greedy approach has been suggested for approximate solution of the knapsack problems [64]. For the classical 0-1 KP, it involves two steps: (1) sort the items in descending order of value-resource ratio $v_i/r_i$ and (2) assign as many items as possible from the beginning of the list until the resource constraint is violated.

To apply the greedy method to MDKP, Toyoda proposed a new measurement called aggregate resource consumption [63], $a_i = \frac{r_i \cdot C}{|C|}$, where $C = (C_1, C_2, \ldots, C_m)$ denotes the current resource usage vector and $r_i = (r_{i1}, r_{i2}, \ldots, r_{im})$ denotes the resource requirements of item $i$. In our MP heuristic algorithm, the item maps to the content; $C_m$ to the current resource usage in the cell $m$ and $r_{im}$ to the required resource to service the content $i$ in the cell $m$. The solution of the MDKP involves iterative assignment of items until the resource constraint is violated. Initially no item is assigned, i.e., a new content will be assigned to cells where no contents have been assigned, $a_i = r_i$. In each iteration, the values of $v_i/a_i$ are calculated, and the item with the greatest value of $v_i/a_i$ is assigned. This selection criterion penalizes the items giving low profit with high resource consumption. It is highly probable that the item giving the highest profit per unit resource consumption will lead us to a near-optimal solution.
C Resource Management Implementation on Testbed Prototype

The viability of the resource management mechanism developed in this thesis has been evaluated in a real testbed prototype. Basically, two types of networks are emulated, namely a MBMS-enabled UMTS network with three cells and a DVB network with one cell. They are overlapping in a common service area. Two IoN entities were implemented in the prototype, namely RM (resource manager) and GM (group manager). These two entities interact with each other to offer cost-efficient interactive multimedia services to mobile users by using multicast transmission mode. More details regarding the GM implementation can be found in [68]. This section aims to explain the prototype implementation of resource management algorithms that consists of MDS (multicast-on-demand scheduling) algorithm and MUA (multicast user assignment) algorithm. More details regarding the resource management algorithm, resource management function architecture and multimedia-on-demand service provision procedure can be found in Section 6.2, Section 6.3 and Section 6.4 respectively.

C.1 Resource Manager

In Figure A-4, the RM implemented in the prototype and the interface with the GM are illustrated.

![Diagram of Prototype implementation of RM](image)

Figure A-4 Prototype implementation of RM

The multicast service is configured at the GM using a service management interface. The GM collects information of registered users who randomly send their requests via a portable website.
Then the GM triggers RM and transfers the information of every registered user to RM. This information includes the content identity desired by users, user terminal capability and user location. In addition to the user relevant parameters, the GM also provides RM with the service relevant parameters, such as a set of service transmission rates.

After being triggered by GM, in the RM, a particular batching queue is created for users who require the same content, regardless of the transmission rate. These users will be batched for a certain time in the batching queue. Two algorithms are implemented to identify the batching time, namely the timeout-based batching and size-based batching. In the case of timeout-based batching, a timer is set for every batching queue, when the first user request arrives at this batching queue. The timer is expired after a certain time period. Different values of the time period for timer expiry were evaluated. In the case of size-based batching, there is also a timer configured for each batching queue since the first user request arrives. Different from the timeout-based batching, in the size-based batching, the timer is expired when a certain threshold number of user requests are reached. Difference values of this threshold were evaluated in the prototype.

When the timer for a particular batching queue is expired, the RM executes MUA algorithm to assign specific transmission networks and data rates to users in this batching queue. The MP algorithm is implemented for MUA, taking into account the user terminal capability, user location, service transmission rate and resource cost in each cell of constituent networks. The user and service related context information are provided by GM. The cost quote is calculated in the resource cost calculation function. This function takes as inputs the current available resources in each network from a database that is located in the IoN gateway. This parameter is actually collected by the IoN local monitor that is not implemented for the moment. After obtaining all the parameters, the MP algorithm is carried out in network & QoS selection function. Basically, the MP algorithm assigns a user to the network where the highest transmission rate is achieved. If both 3G/MBMS and DVB networks offer the same transmission rate, the DVB is selected due to the more cost-effective transmission. More details regarding the MP algorithm can be found in Chapter 4.

As a result of the resource management performed by RM, each user who requires a particular content is assigned to a particular serving cell. Moreover, the transmission rate of the desired content is determined to serve these users in a cell-base. For those users who are not assigned to any cells due to the shortage of available resources, they are lost from the system and marked as blocked users. The results of chosen serving cells and transmission rates for each user are stored into the corresponding user data structure and transferred to GM to implement the decision. Meanwhile, information of the resource availability in the database is updated by RM to reflect the decisions.
C.2 Testbed Setup

The prototype implementation is built on a wireless network testbed that provides a flexible, reconfigurable wireless network platform capable to actually implement networking and service provision scenarios and situations rather than simulating them [100]. At the heart of the testbed is a SUN E10000 Server, which monitors and controls the main network capabilities such as the mobility management, multimedia service provision and routing, network management, etc. The network structure of the wireless network testbed is illustrated in Figure A-5.

![Network structure of wireless network testbed](image)

The network mainly consists of the Cisco networking equipment, connecting Sun and Windows driven workstations, a LAN based (IEEE 802.11b) wireless access part, Toshiba Laptops and Compaq iPacs.

Constructed on the wireless network testbed, the network structure for the RM and GM implementation is illustrated in Figure A-6 [68].
The setup consists of Cisco hardware routers and switching equipment. A WLAN access point is used to represent a DVB network cell, while three UMTS network cells are represented by different Ethernet subnets. Three laptops are used to represent mobile clients. Note that aside from these three laptops, a group of mobile clients are automatically generated when starting the test. Each of the laptops is equipped with a WLAN and Ethernet interface. The laptops are assumed to be located in separate UMTS cells, covered by a common DVB cell. As a consequence each laptop is connected to different Ethernet subnets, while at the same time, having connectivity to the same WLAN access point. The IoN gateway manager application that includes the RM and GM is set-up on a desktop server in the network. For the content provisioning, a video streaming server is setup on a desktop machine in the network.
C.3 Proof of Concept

The RM was tested in a fully operational model. The tests have been carried out for different resource conditions and user heterogeneity cases. To perform the resource management, a network operator or service provider needs to configure the features of multicast user services, as well as the resource management algorithms, such as the scheduling policies. GUIs were used for these configurations, as shown in Figure A-7 and Figure A-8.

![Figure A-7 GUI for multicast user service configuration](image)

A 'BBC Movie World' multicast user service is configured, including two service flows, each of which is configured with an IP multicast address, port number and a QoS parameter, e.g., data rate. The high QoS of the service represents a data rate of 256 Kbps and the low QoS of the service represents a data rate of 64 Kbps.
During the resource management configuration, two scheduling policies including the respective parameters can be chosen, namely the timeout-based batching and size-based batching. In addition, to evaluate the resource management in different resource availability scenarios, the GUI enables the configuration of the available resources in different cells. The Figure A-8 shows that the size-based batching policy is used with the batching size of 90. The resource conditions configured in our tests are presented in Figure A-9.

In order to verify the correctness of the resource management operation, display windows were adopted, and were implemented in the gateway manager to demonstrate the procedure of users’ requests aggregation, the status of each user after resource management and the amount of
available resources in each cell. Based on the configuration in Figure A-7, Figure A-8, the test results of the resource management functionalities are illustrated as follows.

<table>
<thead>
<tr>
<th>Operations</th>
<th>Resources</th>
<th>Tools</th>
<th>Help</th>
</tr>
</thead>
<tbody>
<tr>
<td>Received subscription request from user 81 for service group 1000</td>
<td>Added user 81 successfully to service group 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received subscription request from user 82 for service group 1000</td>
<td>Added user 82 successfully to service group 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received subscription request from user 83 for service group 1000</td>
<td>Added user 83 successfully to service group 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received subscription request from user 84 for service group 1000</td>
<td>Added user 84 successfully to service group 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received subscription request from user 85 for service group 1000</td>
<td>Added user 85 successfully to service group 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received subscription request from user 86 for service group 1000</td>
<td>Added user 86 successfully to service group 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received subscription request from user 87 for service group 1000</td>
<td>Added user 87 successfully to service group 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received subscription request from user 88 for service group 1000</td>
<td>Added user 88 successfully to service group 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received subscription request from user 89 for service group 1000</td>
<td>Added user 89 successfully to service group 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received subscription request from user 90 for service group 1000</td>
<td>Added user 90 successfully to service group 1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performing network selection!</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Starting video server for low quality on udp:239.0.0.2:2002
Starting video server for high quality on udp:239.0.0.1:2001

**Figure A-10 test result for the size-based batching in the resource management**

This result examined the correctness of the size-based batching policy for the scheduling function. It shows that the network selection function is triggered after 90 users’ requests are aggregated.
Appendix C. Resource Management Implementation on Testbed Prototype

<table>
<thead>
<tr>
<th>User ID</th>
<th>Terminal Type</th>
<th>UMTS Cell</th>
<th>DVB Cell</th>
<th>Selected Network</th>
<th>Selected QoS</th>
<th>Bitrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>UMTS</td>
<td>0</td>
<td>0</td>
<td>UMTS</td>
<td>High</td>
<td>256</td>
</tr>
<tr>
<td>4</td>
<td>UMTS</td>
<td>0</td>
<td>0</td>
<td>UMTS</td>
<td>High</td>
<td>256</td>
</tr>
<tr>
<td>5</td>
<td>UMTS</td>
<td>0</td>
<td>0</td>
<td>UMTS</td>
<td>High</td>
<td>256</td>
</tr>
<tr>
<td>6</td>
<td>UMTS</td>
<td>0</td>
<td>0</td>
<td>UMTS</td>
<td>High</td>
<td>256</td>
</tr>
<tr>
<td>7</td>
<td>UMTS</td>
<td>0</td>
<td>0</td>
<td>UMTS</td>
<td>High</td>
<td>256</td>
</tr>
<tr>
<td>8</td>
<td>UMTS</td>
<td>0</td>
<td>0</td>
<td>UMTS</td>
<td>High</td>
<td>256</td>
</tr>
<tr>
<td>9</td>
<td>UMTS</td>
<td>0</td>
<td>0</td>
<td>UMTS</td>
<td>High</td>
<td>256</td>
</tr>
<tr>
<td>10</td>
<td>UMTS</td>
<td>0</td>
<td>0</td>
<td>UMTS</td>
<td>High</td>
<td>256</td>
</tr>
<tr>
<td>11</td>
<td>UMTS</td>
<td>1</td>
<td>0</td>
<td>UMTS</td>
<td>Low</td>
<td>64</td>
</tr>
<tr>
<td>12</td>
<td>UMTS</td>
<td>1</td>
<td>0</td>
<td>UMTS</td>
<td>Low</td>
<td>64</td>
</tr>
<tr>
<td>13</td>
<td>UMTS</td>
<td>1</td>
<td>0</td>
<td>UMTS</td>
<td>Low</td>
<td>64</td>
</tr>
<tr>
<td>14</td>
<td>UMTS</td>
<td>1</td>
<td>0</td>
<td>UMTS</td>
<td>Low</td>
<td>64</td>
</tr>
</tbody>
</table>

Figure A-11 test results for UMTS terminals after network selection in the resource management
Figure A-12 test results for multimode terminals after network selection in the resource management

These results examined the correctness of the network selection function in the resource management. These results show that the UMTS terminals in the UMTS cell 0 are still served in the same cell with the high quality flow due to the sufficient resources. However, the UMTS terminals in the UMTS cell 1 are served with the low quality flow due to the insufficient resources. The multimode terminals in both UMTS cell 0 and cell 1 are served in DVB with the high quality flow due to the more efficient resource utilization in DVB. The remaining resources after the network selection are displayed as follows. We can see that the current available resources in both UMTS and DVB change accordingly based on the results of the network selection, as the chosen flows have been delivered through chosen networks.
In addition, during the test, we examined that the chosen service flows were properly delivered to users' laptops through the chosen network. In addition to the resource management, we also examined the correctness of the group management function in the testbed. That includes the service creation, service transmission, video display and vertical handover. More details about the evaluation of the group management in the testbed can be found in [68].