Flexible Radio Resource Management for Multicast Multimedia Service Provision : Modeling and Optimization

Qing Xu

▶ **To cite this version:**

HAL Id: tel-02084439

https://theses.hal.science/tel-02084439

Submitted on 29 Mar 2019

SPIM

Thèse de Doctorat

Flexible Radio Resource Management for Multicast Multimedia Service Provision: Modeling and Optimization

Qing XU

THÈSE présentée par

Qing XU

pour obtenir le

Grade de Docteur de
l'Université de Technologie de Belfort-Montbéliard

Spécialité : **Informatique**

Flexible Radio Resource Management for Multicast Multimedia Service Provision: Modeling and Optimization

Soutenue le 29 Août 2014 devant le Jury :

Aiqun HU	Rapporteur	Professeur à Southeast University
Stuart ALLEN	Rapporteur	Professeur à Cardiff University
Abdel LISSER	Examinateur	Professeur à Université Paris Sud – Orsay
Jin-Kao HAO	Examinateur	Professeur à Université d'Angers, LERIA - Faculté des Sciences
Alexandre CAMINADA	Directeur de thèse	Professeur à l'UTBM
Hakim MABED	Co-Directeur	Maître de Conférences à l'Université de Franche-Comté
Frédéric LASSABE	Co-Directeur	Enseignant-Chercheur à l'UTBM

Acknowledgments

Foremost, I would like to express my sincere gratitude to my thesis advisor Prof. Dr. Alexandre CAMINADA, for his continuous support to my PhD study and research, for his patience, motivation, enthusiasm, insightful comments and immense knowledge. His guidance helped during the time of my research and writing of this thesis. I would like to thank my co-advisor Assoc. Prof. Dr. Hakim Mabed and Dr. Frédéric Lassabe for their professional advice, patience, encouragement and friendship. I could not have imagined having a better group of advisors and mentors for my PhD study.

I would like to thank my friends who have supported me with their friendship and encouragement all these years of my studies.

Last but certainly not least, with my love and gratitude, I want to dedicate this PhD dissertation to my father and mother, to my husband and my son. They all supported me generously and unconditionally throughout my studies.

Résumé

Le conflit entre la demande de services multimédia en multidiffusion à haut débit (MBMS) et les limites en ressources radio demandent une gestion efficace de l'allocation des ressources radio (RRM) dans les réseaux 3G UMTS. À l'opposé des travaux existant dans ce domaine, cette thèse se propose de résoudre le problème de RRM dans les MBMS par une approche doptimisation combinatoire. Le travail commence par une modélisation formelle du problème cible, désigné comme Flexible Radio Resource Management Model (F2R2M). Une analyse de la complexité et du paysage de recherche est effectuée à partir de ce modèle. Tout dabord on montre qu'en assouplissant les contraintes de code OVSF, le problème de RRM pour les MBMS peut s'apparenter à un problème de sac à dos à choix multiples (MCKP). Une telle constatation permet de calculer les limites théoriques de la solution en résolvant le MCKP similaire. En outre, l'analyse du paysage montre que les espaces de recherche sont accidentés et constellés d'optima locaux. Sur la base de cette analyse, des algorithmes métaheuristiques sont étudiés pour résoudre le problème. Nous montrons tout d'abord que un Greedy Local Search (GLS) et un recuit simulé (SA) peuvent trouver de meilleures solutions que les approches existantes implémentées dans le système UMTS, mais la multiplicité des optima locaux rend les algorithmes très instables. Un algorithme de recherche tabou (TS) incluant une recherche à voisinage variable (VNS) est aussi développé et comparé aux autres algorithmes (GLS et SA) et aux approches actuelles du système UMTS; les résultats de la recherche tabou dépassent toutes les autres approches. Enfin les meilleures solutions trouvées par TS sont également comparées avec les solutions théoriques générées par le solveur MCKP. On constate que les meilleures solutions trouvées par TS sont égales ou très proches des solutions optimales théoriques.

Mots clés: gestion des ressources radio UMTS, service multimédia MBMS, transmission à échelle variable, problème de sac à dos, recherche tabou et recherche à voisinage variable.

Abstract

The high throughputs supported by the multimedia multicast services (MBMS) and the limited radio resources result in strong requirement for efficient radio resource management (RRM) in UMTS 3G networks. This PhD thesis proposes to solve the MBMS RRM problem as a combinatorial optimization problem. The work starts with a formal modeling of the problem, named as the Flexible Radio Resource Management Model (F2R2M). An in-depth analysis of the problem complexity and the search landscape is done based on this model. It is showed that, by relaxing the OVSF code constraints, the MBMS RRM problem can be approximated as a Multiple-Choice Knapsack Problem (MCKP). Such work allows us to compute the theoretical solution bounds by solving the approximated MCKP. Then the fitness landscape analysis shows that the search spaces are rough and reveal several local optimums. Based on the analysis, some metaheuristic algorithms are studied to solve the MBMS RRM problem. We first show that a Greedy Local Search (GLS) and a Simulated Annealing (SA) allow us to find better solutions than the existing approaches implemented in the UMTS system, however the results are instable due to the landscape roughness. Finally we have developed a Tabu Search (TS) mixed with a Variable Neighborhood Search (VNS) algorithm and we have compared it with GLS, SA and UMTS embedded algorithms. Not only the TS outperforms all the other approaches on several scenarios but also, by comparing it with the theoretical solution bounds generated by the MCKP solver, we observe that TS is equal or close to the theoretical optimal solutions.

Keyword: UMTS radio resource management, MBMS multimedia services, scalable transmission, knapsack problem, Tabu Search, Variable Neighborhood Search.

Contents

List of Tables

List of Figures

List of Algorithmes

Introduction

In the past decades, the rapid growth of mobile communication technology boosts the demand of wireless multimedia services. According to the Cisco mobile forecast highlights [20], the global consumer mobile data traffic grew of 74% in 2012, and grew more than 81% in 2013. From 2013 to 2018, the mobile data traffic is expected to grow at 61% Compound Annual Growth Rate (CAGE). By 2018, 69% of the world mobile data traffic will be video, up from 53% in 2013. The rapid growth of multimedia mobile data demands higher transmission rate with lower radio and network resource cost. Hence the content and service providers are increasingly interested in more efficient multicast communications over mobile networks.

The Universal Mobile Telecommunications System (UMTS) is the second world's most widely used wireless technology with 500 million customers [19]. In the radio network of UMTS where the radio resources (power and channelization codes) are limited, the sharing of resources among numerous users per cell is constrained with more services subscriptions and higher traffic bandwidth requirements. Hence UMTS is facing challenges with the rapid growth of multicast multimedia service requirement. In order to provide an efficient multicast and broadcast transmission platform in mobile networks, the 3GPP specified the Multimedia Broadcast/Multicast Service (MBMS) since Release 6 specifications [16].

In MBMS, the broadcast/multicast data are provided by a particular service center which performs control functions for all individual users in the same MBMS region. For each MBMS service, only one MBMS tunnel is established, through which the service data is sent to a class D IP multicast address (identifying a multicast group). Hence the network resource is saved by avoiding content duplication. In the radio air link, MBMS defines common logical channels for multicast data and signaling transmission. The same service is served to multiple users by a a common signal transmission facility and bearer (point-to-multipoint transmission mode), hence conserving radio resources.

MBMS aims to provide more efficient multimedia streaming within UMTS. Since the radio resource is limited in UMTS, to provide the multimedia service with satisfied quality requirement by using minimal transmission power and channel codes is the most critical topic in MBMS study. Besides, if the point-to-multipoint carrier is used for service that are not popular, the complicated MBMS signaling will actually lead to more overhead than with a simple unicast link. With these considerations, a wide range of work is investigated on the efficient Radio Resource Management (RRM) for MBMS. The literature in this field mainly focus on the static switching between conventional MBMS transmission modes. That is because different channels for carrying MBMS traffic have different characteristics in power consumption.

To bring additional gains, the enhancement in physical layer such as Macro Diversity and Spatio-Temporal Transmit Diversity (STTD) are also proposed. Few studies mention the selection of transmitted content for multicast service, which is based on the scalability encoding technology. Besides, stream schedule is another topic in the field of MBMS RRM, the base station needs to schedule streams by determining the target multicast group and transmit rate per time slot.

By emphasizing the existing literatures, some shortcomings are identified: first, the approaches are mono-objective, e.g. only consider the power consumption. In addition, almost all the existing allocation approaches study the selection among different transmission channels but not flexible combination of them. Such deterministic approaches are easy to implement but not optimum. Furthermore, none of the studies have ever propose a general model, which allows to evaluate all the existing approaches under the same criterion. To treat these problems, a general model is required to mathematically formulate the MBMS RRM problem and evaluate all existing allocation approaches. From the model, the problem complexity and characteristics should be analyzed. The analysis leads to design innovative algorithms which can conserve the transmission power and channel code utilization, and achieve the trade-off between the resource consumption and the service quality.

This thesis manuscript includes five chapters. Chapter 1 introduces the background information related to the research area of this thesis manuscript. The first section starts from the fundamentals in UMTS. The UMTS Terrestrial Radio Access Network (UTRAN) provides air interface for UE. In particular, the radio network controller (RNC) takes in charge the setup and the release of the radio bearers for data transmission. Then the Wideband Code Division Multiple Access (WCDMA), the radio access technology of UMTS, is explained. In which the notation of channelization codes is illustrated. In the second section, the development of multicast method before and with MBMS is introduced. The MBMS specific mechanisms help us to understand the advantages of MBMS as well as state the challenges and motivations in this thesis. The last section in background statement, briefly introduces the High Speed Downlink Packet Access (HSDPA) technology. HSDPA offers a new option carrying MBMS multicast service as it can use multiple codes to improve peak throughput for certain users. Based on its adaptive coding and modulation combination mechanism, the service quality can be guaranteed when link quality is very favorable. Finally the different RRM algorithms proposed in literature are detailed, their advantages and disadvantages are analyzed, following by the motivation of this work.

Chapter 2 presents the first contribution of this thesis: the mathematical modeling of the RRM Problem for MBMS system, named Flexible Radio Resource Management Model (F2R2M). This model maps the MBMS radio resource establishment procedures into a three-phase flow chart. Within this model, a dynamic radio resource allocation framework for MBMS is proposed. This allocation framework is

abstracted by seven functional modules. The framework explores the solution space by iteratively searching a new allocation solution by modifying the current solution. To evaluate the efficiency of the new solution, a two-dimensional cost function is proposed, such that the estimated throughput loss and the estimated power estimation are compared by a lexicographic-order evaluation criterion. These modules and the search procedure target at finding the best solution satisfying the QoS requirement of multicast service and minimizing the transmission power, with the feasibility control of channelization code availability and the power saturation. Besides, the proposed model could also be utilized as a general platform to abstract, implement and evaluate the other existing MBMS radio resource allocation approaches. Finally, the simulation parameters are described, several scenarios with different traffic loads and user distributions are designed for study in the following chapters.

Chapter 3 describes the second contribution: in-depth mathematical analysis on the proposed model. In the first section, it is shown that by reducing the channel code constraints, the MBMS RRM problem can be approximated as a Multi-Choice Knapsack Problem (MCKP). Based on this, the solution complexity can be analyzed and NP-Hard proof is provided. Also based on this, the solution bounds for MBMS RRM problem can be obtained by solving this MCKP problem. In the second section, the characteristics and the complexity of the problem based on fitness landscape analysis method are analyzed. First, two neighborhood functions are proposed, constructing two different fitness landscapes. Then the mathematical solution representations and solution distance measurement are proposed. The two fitness landscape constructed by these two operators are generated through Greedy Local Search (GLS) method. These two fitness spaces are studied in three aspects: the distribution of feasible solutions in the search space, the structure of fitness space, and the relationships between solution distance and fitness value. These analysis reveal that the studied problem is rugged, i.e. the search space is not flat, hence the search procedure is relevant and difficult. The comparative study of these two fitness landscapes helps us to select the appropriate neighborhood function. Based on the NP-Hard characteristics of the studied problem, it is reasonable to select metaheuristic approaches to solve it. Then in chapter 4 and chapter 5, the third contribution of our work is described: the optimization process.

Chapter 4 presents a Simulated Annealing (SA) based algorithm to solve the studied problem. SA is selected because it is easy to implement and can avoid the local optima by accepting the new solution with probability. Firstly, the general optimization procedure and parameter definitions in classical SA are introduced. Then the problem specific parameters are discussed and selected. In the construction of new solution, the selected neighborhood operator in chapter 3 is used. In the acceptance probability function, the Boltzmann function is modified to calculate the acceptance probability according to the change of the proposed fitness value. The SA results are compared with the results of the greedy local search. For small size scenarios, simulated annealing obtains equivalent solution quality as greedy local

search but with longer time cost. For larger size scenarios, simulated annealing only obtains worse solution than that of greedy local search, which is because of the randomness characteristics of SA while the studied problem is rugged.

Chapter 5 presents a Tabu Search (TS) based algorithm to further increase the efficiency of the search procedure in the proposed model. Three tabu memory structures are defined and their search performances are compared. Then the classical TS is extended by proposing a problem specific method named tabu repair mechanism, which helps to explore candidate solutions. Simulation results show that TS outperforms the deterministic algorithms. For most scenarios, tabu search can obtain feasible solutions with full utilization of power and channel codes. While existing approaches obtains either feasible solutions but with unnecessary throughput scarifies or unfeasible solutions but higher power consumption than tabu search results. Besides, tabu search results are also better than two other metaheuristic approaches: the SA and GLS. For small size scenarios, TS can find solutions with less power consumption than GLS and SA and equivalent QoS; for large size scenarios, TS obtains solution not only with less power consumption than GLS and SA but also with fully satisfied bandwidth hence if it decreases the possibility of channel code saturation.

Finally, the contributions of this thesis manuscript are concluded, the simulation results are compared, and the opportunities for future work are identified.

Background Knowledge

This chapter provides the background knowledge and the existing literature related to the research area of this thesis manuscript. The section 1.1 describes the Universal Mobile Telecommunications System (UMTS) architecture, including the functionalities of the Core Network (CN) and the UMTS Terrestrial Radio Access Network (UTRAN). Then the basic principles of Wideband Code Division Multiple Access (WCDMA) are described, with the fundamentals of channelization code allocation in 3G network. In section 1.2, the history of multicast in cellular networks and the Multimedia Broadcast Multicast Service (MBMS) systems are introduced, leading to the studied issues in existing literature and in this thesis. In section 1.3, the High Speed Downlink Packet Access (HSDPA) is briefly introduced. HSDPA defines the HS-DSCH channel, which is potentially used to carry multicast data. Finally, in section 1.4, the state of the art in the MBMS RRM is analyzed; the advantages and drawbacks of the existing radio resource allocation approaches are illustrated, then the motivation of this work is stated.

Contents

1.1 Universal Mobile Telecommunications System

The throughput limitations of the Global System for Mobile communications (GSM) led the International Telecommunications Union (ITU) to initiate work on a new worldwide standard, called 3G for the third generation network. The 3G Partnership Project (3GPP) develops the Universal Mobile Telecommunications System (UMTS) [53] that delivers high-bandwidth data and voice services to mobile users and mobile web data. UMTS is based on Wideband-Code Division Multiple Access (W-CDMA) for the radio part and inherits of the GSM/General Packet Radio Service (GPRS) topology for the network backbone.

1.1.1 UMTS network architecture

Figure 1.1 shows the architecture of a UMTS network. It consists of three parts: the Core Network (CN), the UMTS Terrestrial Radio Access Network (UTRAN) and the User Equipment (UE). The CN is responsible for switching/routing voice, inter-system handover, gateway to other networks (fixed or wireless), and perform location management when there is no dedicated links between the UE and the UTRAN. The UTRAN handles all radio-related functionalities, and operates in Frequency Division Duplex (FDD) or Time Division Duplex (TDD) modes using WCDMA protocol. The UE is the equipment used by the user to access the UMTS services.
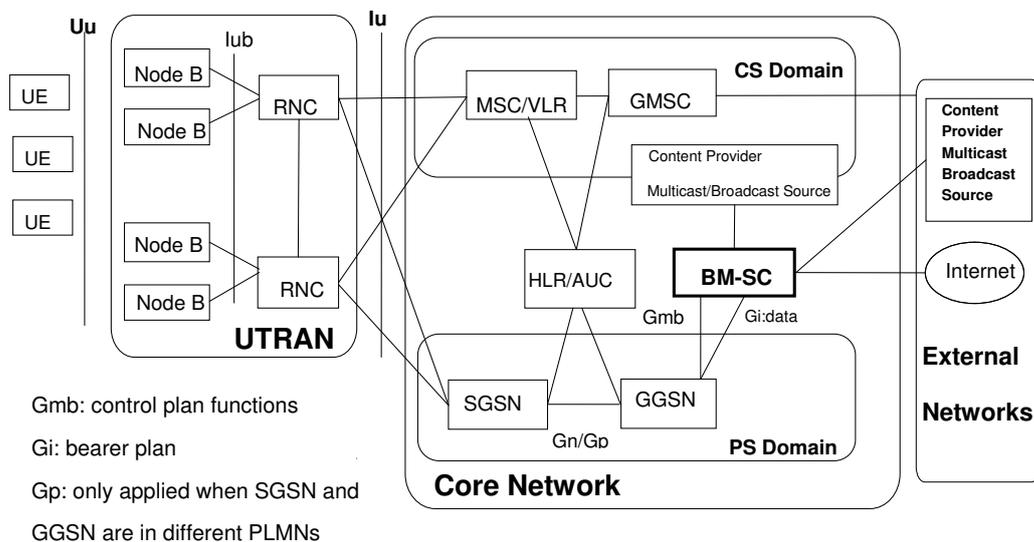
Figure 1.1: UMTS network architecture

1.1.1.1 Core Network

The CN is logically divided into two service domains: the Circuit-Switched (CS) service domain and the Packet-Switched (PS) service domain. The CS domain handles the voice-related traffic, while the PS domain handles the packet transfer. In the CS domain, the network includes the Mobile Switching Center (MSC), the Visitor Location Register (VLR) and the Gateway Mobile Switching Center (GMSC). The PS domain consists of the Gateway GPRS Support Node (GGSN) and the Serving GPRS Support Node (SGSN). Other network elements, such as the Home Location Register (HLR) and the Authentication Center (AUC) are shared by both domains.

1.1.1.2 UTRAN

The UTRAN provides the air interface for UE. As shown in Figure 1.1, the UTRAN consists of several Radio Network Subsystems (RNS). Each RNS is controlled by a Radio Network Controller (RNC) and has several Node B. Each Node B can control several antennas and each antenna covers an area called a radio cell. The UE can directly communicate with one or more antennas (when the UE is in handover procedure). The radio resources control is implemented through a distributed architecture; each RNC is connected with the CN over the interface named Iu and with a Node B over the interface named Iub.

Radio Network Controller The RNC is a key element within the UTRAN as it controls all the radio resources. A RNC is responsible for a wide range of tasks [72].

- **Admission control:** in the CDMA system, it is very important to keep the interference below a certain level. The RNC calculates the traffic within each cell, and then it decides to accept or reject the new coming calls.
- **Power control:** the RNC only performs the outer loop power control. This means that the RNC controls the transmission power in one cell on the basis of the interference received from the other neighbor cells. While the fast power control is performed by a Node B 1500 times per second, the outer loop power control helps the RNC to minimize the interference between the neighbor cells.
- **Radio bearer setup and release:** the RNC has to set up, maintain and release a logical data connection to a UE. This connection is called UMTS radio bearer.
- **Radio resource control:** the RNC controls all radio resources of the cells connected to it via a Node B. This task includes the interference control and load measurements.
- **Handover control:** based on the downlink/uplink signal strength and the signal-to-interference ratio received by the UE and the Node B, the RNC can

decide if the current cell is suitable for a given connection. When the RNC decides to handover, it informs the new cell and the UE.

To achieve above tasks, one physical RNC contains three logical functionalities [54]:

– **CRNC:** the Controlling RNC (CRNC) controls the resources of one Node B. It performs the load and congestion control within the cells of the Node B. Besides, a CRNC executes the admission control and the code allocation to establish new radio links in these cells.
– **SRNC:** the Serving RNC serves a particular UE and manages the connections (to/from the CN) with that UE.
– **DRNC:** the Drift RNC fulfills a similar role to the SRNC except that it is involved only in the case of soft handover.

The difference between the CRNC, SRNC and DRNC is that the CRNC is logically tied to the Node B, not to the connections. On the contrary, the SRNC and the DRNC are tied to the connections with the UE, which implies that the CRNC manages the common and the shared resource while the SRNC and the DRNC manage the dedicated resources.

Node B   The Node B is the base station and provides the radio coverage to one or more cells. It is connected directly with the UE via the WCDMA air accessing technology. An important task of the Node B is the inner loop power control. The Node B measures the link quality and the signal strength, it manages the air interface transmission and reception, the modulation and the demodulation, the physical channel coding, etc. With the emergence of High Speed Downlink Packet Access (HSDPA), the Node B even handles some logic functionalities (e.g. retransmission) for lower response times.

1.1.1.3   NAS Stratum and AS Stratum

Vertically, there are two strata in the UMTS signaling protocol stack: the Non-Access Stratum (NAS) and the Access Stratum (AS). The NAS protocols are applied between the UE and the core network, for which the access stratum acts as a relay. The UMTS non-access stratum consists in the Connection Management (CM), the Session Management (SM), the Mobility Management (MM), and the GPRS Mobility Management (GMM).

The access stratum consists of three layers. The layer 1 is the Physical Layer (PHY), the layer 2 consists of the Radio Link Control (RLC) and the Medium Access Control (MAC). The layer 3 is the Radio Resource Control (RRC).

The layer 1 service is the physical layer. It is responsible for transporting the

data received from the higher layers over the physical channels. It hides all details of the underlying physical media, and provides the transport channels to the MAC layer. The PHY layer provides transport channels to the L2/MAC layer. The concept of channels will be introduced later.

The layer 2 service consists of four sub-layers:

– **RLC:** The RLC layer provides service to the higher layers in both control plane and user plane. It provides Service Access Points (SAPs) to the higher layers, to invoke some service of the RLC layer.

– **MAC:** The MAC sub-layer provides services to the RLC layer by means of logical channels. This layer internally maps the data received over the logical channels to the transport channels.

– **PDCP:** The Packet Data Convergence Protocol (PDCP) sub-layer provides PDCP Service Data Unit (SDU) delivery service through the SAP for user plane packet data.

– **BMC:** The BMC (Broadcast and Multicast Control) sub-layer provides a broadcast/multicast transmission service in the user plane. It has functions for storing cell broadcast messages and transmitting BMC messages to the UE.
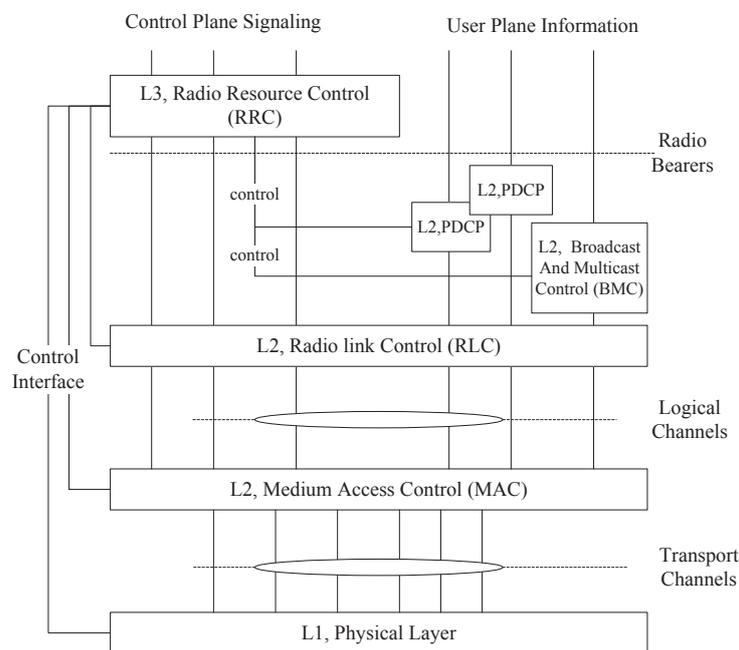
Figure 1.2: Radio interface protocol stacks in access stratum

As shown in Figure 1.2, the protocol stack in layer 3 can be divided into the user plane and the control plane. The control plane protocol stack deals with signaling

protocol. For example, the Radio Resource Control (RRC) protocol is part of the
control plane, which carries the network signaling messages. The user plane protocol
stack deals with user protocols, it carries the data streams from/to the user.

1.1.2   UMTS channels

The UMTS channels can be classified in terms of functionalities, data flow di-
rection and sharing. In terms of data flow direction, the downlink channels are
transmitted by the UTRAN and received by the UE, while the uplink channels are
transmitted by the UE and received by the UTRAN. In terms of sharing mechanism
among UE, the common channels send information toward and from multiple UE,
while the dedicated channels send information to and from a single UE. In terms
of functions, there are the logical channels, the transport channels and the physical
channels (see Figure 1.2).

Logical channels   The logical channels provide the data transfer service of the
MAC layer. The logical channel type is defined by its content and the kind of offered
data service. A general classification of logical channel is into two groups: the control
channels and the traffic channels. The control channels are used to transfer control
plane information, and the traffic channels for the user plan information.

Transport channels   The MAC layer provides the logical channel to transport
channel conversion. The connections between the logical channels and the transport
channels are shown in Figure 1.3.
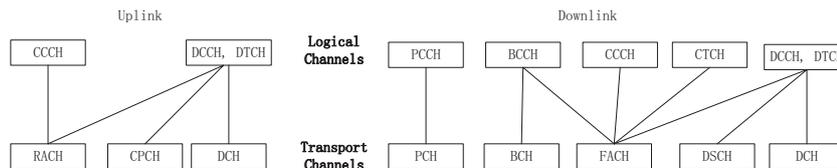
Figure 1.3: Logical channel to transport channel mapping

Different transport channels are defined from the type of information transferred
by that channel. The transport channels can be subdivided into the common trans-
port channels, the dedicated transport channels and the shared transport channels.
The common transport channel is a resource divided between all or a group of users
in a cell, whereas a dedicated transport channel resource, identified by a certain
code on a certain frequency, is reserved for a single user only. The common channels
are the Random Access Channel (RACH) in the uplink and the Forward Access
Channel (FACH) in the downlink. The common channels do not have a feedback

channel, and cannot use the fast closed loop power control, but only the open loop power control or fixed power. Therefore the link level performance of the common channels is not as good as the dedicated channels, and the common channels generate more interference than the dedicated channels. The Dedicated Channel (DCH) is a bi-direction channel with both uplink and downlink connections. Because of the feedback channel, the fast power control and the soft handover can be used. These features improve their radio performance and consequently less interference is generated than with common channels.

Physical channels   The transport channels are mapped in the physical layer to different physical channels. The physical channel is required to support variable bit rate transport channels to offer bandwidth-on-demand services, and to be able to multiplex several services to one connection.

Each transport channel is accompanied by the Transport Format Indicator (TFI) at each time event at which data is expected to arrive for the specific transport channel from the higher layer. The physical layer combines the TFI information from different transport channels to the Transport Format Combination Indicator (TFCI). The TFCI is transmitted in the physical control channel to inform the receiver which transport channels are active for the current frame. The transport channel to physical channel mapping is illustrated in Figure 1.4.
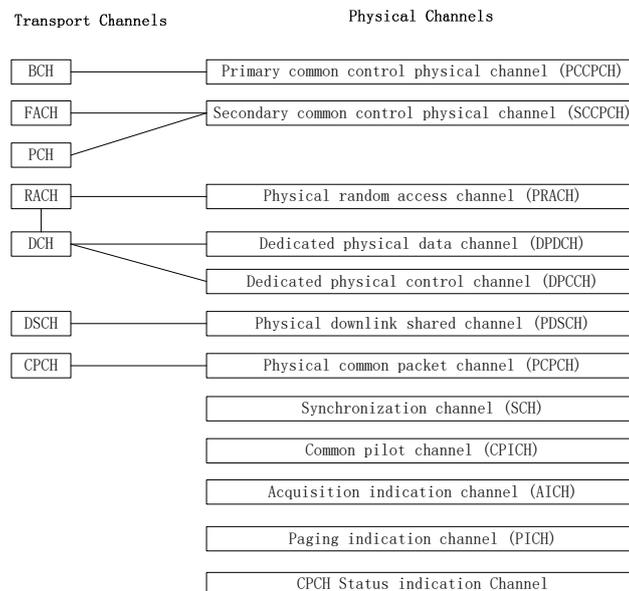
Figure 1.4: Transport channel to physical channel mapping

1.1.3   Wideband Code Division Multiple Access (WCDMA)

In the radio accessing technologies, one of the basic concepts is to allow several transmitters to send information simultaneously over the radio link. It means to share a band of frequencies, i.e. bandwidth between several users. The three most important families of radio access schemes are: Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA). These three mechanisms subdivide radio resources in the frequency, time and code domains, respectively.

WCDMA is a Wideband Direct-Sequence Code Division Multiple Access (DS-CDMA) system, i.e. user information bits are spread over a wide bandwidth by multiplying the user data with quasi-random bits (called chips) derived from CDMA spreading codes. In WCDMA, all users use the same frequency band; to separate different users, the codes used for spreading should be (quasi) orthogonal, i.e. their cross-correlation should be (almost) zero. The chip rate of 3.84 Mchip/s leads to a carrier bandwidth of approximately 5 Mhz. DS-CDMA systems with a bandwidth of about 1 MHz are commonly referred to as narrowband CDMA systems. The inherently wide carrier bandwidth of WCDMA supports high user data rates and also has certain performance benefits, such as increased multipath diversity.

WCDMA supports highly variable user data rates. The user data rate is kept constant during 10 ms frame, however, the data capacity among the users can changed from frame to frame. The different user data rates can be supported using different spreading factor (i.e., the number of chips per bit). This step is spreading operation, which multiplies different data streams with orthogonal spreading codes. UMTS uses Orthogonal Variable Spreading Factor (OVSF) codes to spread data symbol to chips on both the uplink and downlink. OVSF codes are also known as channelization codes or Walsh codes. In its general form, an OVSF code can be written as a sequence of CSF (k); where C stands for channelization code, SF stands for spreading factor, and k stands for code number with $0 < k \leq SF - 1$.
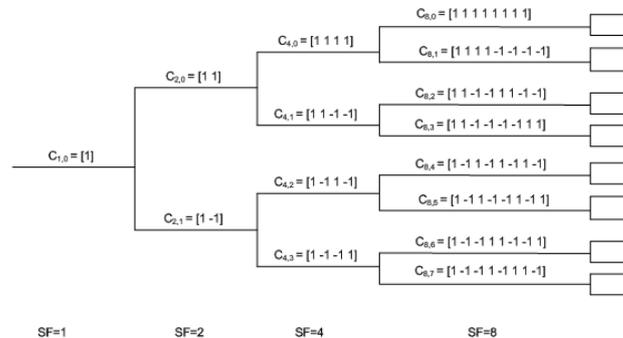
Figure 1.5: OVSF code tree used for orthogonal spreading

The OVSF codes are generated from a code tree as shown in Figure 1.5. It begins with the first generation one-bit code C1(0) = 1; where the subscript 1 stands for spreading factor 1, and (0) stands for code number 0. The second generation consists of two codes: C2(0) and C2(1). They are two-bit codes with a spreading factor of 2. The third and fourth generations consist of four-bit and eight-bit codes numbering of four and eight respectively. The code tree can go up to 10 generations with the 10th generation having 512 codes. For a given code tree generation, the spreading factor is equal to the number of codes. The following functions illustrate how the OVSF codes are generated:

$$C_1(0) = 1 \tag{1.1}$$

$$\begin{bmatrix} C_2(0) \\ C_2(1) \end{bmatrix} = \begin{bmatrix} C_1(0) & C_1(0) \\ C_1(0) & -C_1(0) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{1.2}$$

$$\begin{bmatrix} C_4(0) \\ C_4(1) \\ C_4(2) \\ C_4(3) \end{bmatrix} = \begin{bmatrix} C_2(0) & C_2(0) \\ C_2(0) & -C_2(0) \\ C_2(1) & C_2(1) \\ C_2(1) & -C_2(1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \tag{1.3}$$

Two sequences are said to be orthogonal to each other if they have zero cross-correlation. In the OVSF code tree, two codes are orthogonal if and only if neither code lies on the same path from the other code to the tree root. Graphically, two codes are non orthogonal if they belong to the same branch of the tree. For example, if the code $C_4(0)$ is assigned to a user, the codes $C_1(0)$, $C_2(0)$, $C_8(0)$, $C_8(1)$ and so on, cannot be assigned to any other user in the same cell. In this way, OVSF codes for different channels in the same cell are carefully chosen in order to be mutually orthogonal to each other, this restricts the number of available codes for a given cell.

With a given OVSF code, WCDMA performs the orthogonal spreading by multiplying each encoded symbol with a code, meaning that one symbol is represented by multiple chips. e.g. a Spreading Factor (SF) of 4 means 4 chips per symbol, while a SF of 256 means 256 chips per symbol. The chip rate is kept in constant at 3.84 Mcps(chips per second), and the ratio of chip rate and symbol rate is known as the spreading factor.

$$SF = \frac{Chip\,Rate}{Symbol\,Rate} = \frac{3.84Mcps}{Symbol\,Rate} \tag{1.4}$$

Therefore, high-rate transmissions use low spreading factor while low-rate transmissions use high spreading factors. For example, for voice with a symbol rate of 60 ksps, the 64-bit OVSF code used for spreading runs 64 times faster than the symbol, making the spread symbol run at 3.84 Mcps. For data with a symbol rate

of 960 ksps, the 4-bit OVSF code used for spreading runs four times faster than the symbol, resulting in again a spread symbol chip rate of 3.84 Mcps. Spreading factors can be range from 4 to 512 for the downlink and 4 to 256 for the uplink.

1.2   Multicast in 3G and 3G+

Traditionally, data communications concerns two entities: a transmitter and a receiver. Nowadays, with the introduction of mobile streaming, video conferencing etc, there is an increasing demand of traffic between one transmitter and many receivers, or even many transmitters and many receivers. Hence, efficient broadcast and multicast communications are required.

In broadcast, a message is sent to every possible destination. It is unknown that if only a few receivers are interested in the message. In a wireless mobile network, broadcast transmission not only wastes the network resources but also the receiver resources, since the receiver, whom are not interested in the broadcast data, must consume energy in order to process useless data.

Multicast is more efficient in terms of network and receiver resources than broadcast. Multicast data delivery increases the network efficiency and decreases the server load by sending one data stream to several particular destinations. When the network is aware of the fact that multiple receivers are targeted, it creates a distribution tree from the transmitter towards all receivers overlaying the network topology. The network will duplicate the data only at branching points of the tree towards the receivers. Thus, instead of sending many streams from the transmitter, one to each receiver, multicast lets all receivers listen to the same stream and avoids processing overheads replication at the source on the same link. Multicast requires additional mechanisms for group maintenance while broadcast does not.

1.2.1   Multicast in UMTS network prior to MBMS

Two services for transmitting data from a single source to several destinations were defined prior to MBMS: the Cell Broadcast Service (CBS) and the IP multicast service.

Since Release 4, the CBS service (CBS: Cell Broadcast Service) [14, 10] allows low bit-rate unacknowledged messages to be transmitted to all receivers in a particular area. The CBS broadcasts each message periodically, at a frequency and duration arranged with the information provider. The CBS, however, is targeted to text messaging and without any QoS, therefore, it is unsuitable for high bandwidth multimedia services.

The IP Multicast service [3, 4] is defined since Release 99. It allows IP applications to send data to a set of recipients (a multicast group) specified by an IP address. Any UE may join or leave a multicast group without restrictions. Release 99 IP multicast is implemented by separately sending each packet from the GGSN to each UE, therefore, no sharing gains are achieved, and high bandwidth multimedia services remain expensive. The IP multicast traffic can be received by mobile subscribers already. However, the IP multicast does not allow to share radio or core network resource hence no optimized transport solution exists.

In the initial UMTS multicast design [51], 3GPP decided to terminate the IP multicast routing protocol in the GGSN. With this design, GGSN serves as a rendezvous point routers. Also, GGSN serves as a router Internet Group Management Protocol (IGMP) designated and performs IGMP signaling on point-to-point packet-data channels. IGMP signaling is performed in the network user plane, that means it is seen as data traffic for the UMTS network. Multicast data is forwarded to the UMTS terminal on point-to-point packet-data channels, i.e. unicast distribution. The GGSN manufacturer can choose which IP multicast routing protocol to support. Only the UE and the GGSN are multicast compatible in this design. As shown in Figure 1.6, the distribution tree by the IP multicast service allows the network to treat multicast traffic in the same manner as unicast traffic.
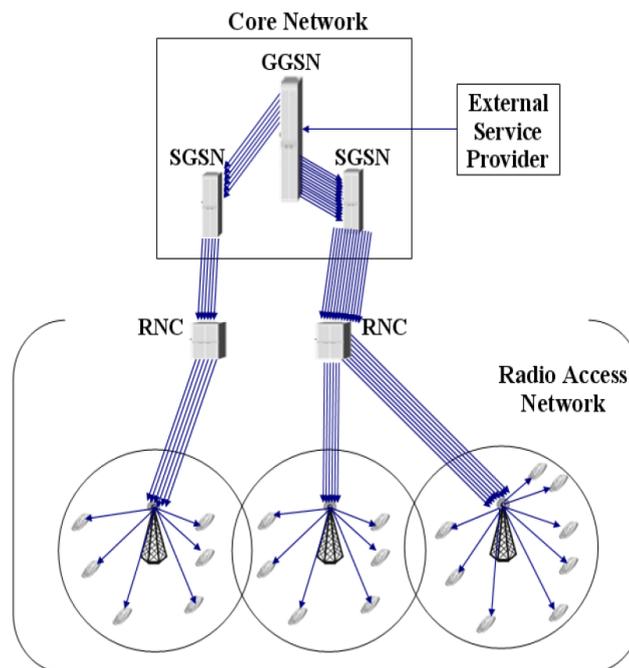
Figure 1.6: Multicast duplication in UMTS network prior to MBMS [51]

To send and receive multicast data, the terminal firstly needs to perform a GPRS

attach, then the terminal must establish a packet data channel with the GGSN. The UMTS terminal is now part of the IGMP environment, and can join and leave the multicast groups using normal IGMP signaling. Finally the terminal must establish one or more packet data channels (Packet Data Protocol (PDP) context activation) for the multicast data flows.

As shown in Figure 1.6, this multicast architecture reduces the load on a wireless source. The source only needs to send one copy of the multicast data to the GGSN. Then the GGSN replicates and forwards the packet on to the multicast distribution tree. However, the UMTS multicast source does not receive any information from multicast members. Thus even if the multicast group does not have any members, the source will continue transmitting its multicast data to the GGSN. The source is not aware of the empty state of the multicast group. A modified signaling connection between the GGSN and the source can avoid this situation. This architecture also imposes a high strain on the GGSN. The GGSN already has the responsibility for many complex mechanisms, thus it is important to avoid turning the GGSN into the UMTS networks bottleneck.

The drawback of this multicast architecture is that it requires more network resources than unicast distribution of the same data. Moreover, due to the fact that GGSN serves as a designated router, detailed membership information must be stored in the GGSN for the UMTS Terminals. This might work efficiently when the number of the users requesting to join the multicast group is low. But when a great number of users request the same MBMS service, the network may then collapse with huge capacity and processing requirements in the core and the radio network.

With these shortcomings in mind, 3GPP defined MBMS to decrease the amount of data within the network. It aims at offering an efficient way to transmit data from a single source to multiple destinations over the radio network. MBMS is transparent to end users (they have the same experience as with Point-to-Point connections) while saving resources on the UE side.

1.2.2   Multimedia Broadcast Multicast Service (MBMS)

To support efficient distribution of multicast multimedia services over mobile networking, the 3GPP specified the Multimedia Broadcast/Multicast Service (MBMS) since Release 6 specifications [18, 16, 17]. The main advantage of MBMS is that it allows many receivers in the same radio cell to be served by a common signal transmission facility, or bearer, thus saving radio resources.

1.2.2.1 MBMS system architecture

MBMS integrates broadcast/multicast transmission capabilities into 3G service and network infrastructures. In UMTS network, bandwidth is a limited resource. MBMS supports the network to transmit the data only once over a particular route. All users that belong to the same multicast group can receive service simultaneously on the same frequency and time slot. As shown in Figure 1.7, the existing Packet-Switched (PS) domain functional entities (UE, GGSN, SGSN, UTRAN and GERAN) and MBMS-specific interface functions (Gmb) are enhanced to support the MBMS bearer service.
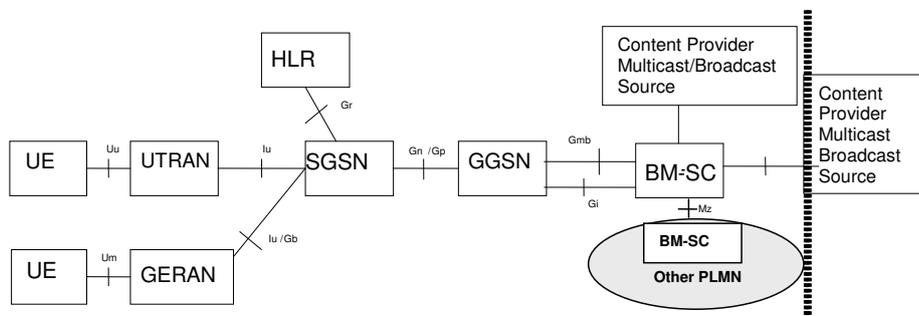
Figure 1.7: Enhancement of MBMS in UMTS [18]

Besides, MBMS system is realized by one additional component named Broadcast Multicast Service Center (BM-SC). The BM-SC is a MBMS specific functional entity supporting various MBMS user service such as schedule and deliver. It performs following functions:

  – Membership function.
  – Session and Transmission (in radio access network) function.
  – Proxy and Transport (in CN) function.
  – Service Announcement function.
  – Security function.

With these enhancement, the MBMS feature is split into the MBMS Bearer Service and the MBMS User Service.

The MBMS Bearer Service includes a Multicast and a Broadcast Mode. The MBMS Bearer Service uses IP multicast addresses for the IP flows. The advantage of the MBMS Bearer Service compared to unicast bearer services (interactive, streaming, etc.) is, that the transmission resources in the core and radio network are shared. One MBMS packet flow is replicated by GGSN, SGSN and RNC (see

Figure 1.7). MBMS may use an advanced counting scheme to decide the most efficient system usage between using (zero, one or more) dedicated (i.e. unicast) radio channels and using one common (i.e. broadcast) radio channel.

The MBMS User Service is basically the MBMS Service Layer and offers a Streaming and a Download Delivery Method. The Streaming Delivery method can be used for continuous transmissions like Mobile TV services. The Download Method is intended for Download and Play services. To increase the transmission reliability, an application layer FEC code may be used. Furthermore, a file-repair service may be offered to complement the download delivery method.

1.2.2.2   MBMS services

According to the different quality of service (QoS) types, the services targeted by MBMS are classified into three categories.

- **Streaming service:** Continuous media such as audio and video, plus supplementary text and images, similar to TV channels but enhanced with multimedia content. Images may also be used for banner images that advertise some product or service. These static media need to be synchronized and displayed with audio/video streams.
- **File Download services:** Reliable binary data transfers without strict delay constraints over an MBMS bearer, similar to conventional file transfers but with multiple receivers. It is necessary that the user receives all the data sent in order to experience the service.
- **Carousel service:** Combination of streaming service and file download service, similar to stock quote ticker tapes. The target media of this service is only static media (e.g. text and/or still images). Time synchronization with other media is also required, e.g. text objects are delivered and updated from time to time. Images are also collated to display low frame-rate video. The benefit of this service is that it is possible over a low bit-rate bearer. An example of the carousel service is a 'ticker-tape' type service in which the data is provided to the user repetitively and updated at certain times to reflect changing circumstances.

Therefore, potential MBMS applications include not only cellular band broadcast mobile television; but also cellular band broadcast mobile radio and area-specific target mobile advertising, etc. These potential applications in MBMS are shown in Table 1.1.

Service type	Service Content	Potential Application
News clip	Text distribution	News/sport highlights, movie trailers, economics, etc
Localized service	Text, video	Tourist information, restaurant, etc
Audio stream	Audio, timed text	Music, live traffic information, voice notification
Content distribution	Downloading, video, audio	Software updates, etc
Video clip	Video and audio streams, timed text	Live events, interactive television voting

Table 1.1: MBMS services and potential applications

1.2.2.3 MBMS service distribution procedures

MBMS supports broadcast and multicast services in mobile environment. It defines the transmission of service through Point-to-Multipoint method. With this manner, the same data is transmitted from a single source entity to multiple recipients while sharing the core and radio network resources.

The broadcast service is a unidirectional point-to-multipoint transmission service, from a single source entity to all users in a broadcast service area. It does not need subscription. The broadcast service is free of charge and does not need specific activation requirements. Broadcast supports streaming service while the Cell Broadcast Service (CBS) is intended for messages only.

The multicast service allows the unidirectional Point-to-Multipoint transmission of multimedia data from a single source point to a multicast group in a particular service area. Unlike the broadcast service, the multicast service requires a subscription to the multicast group. Users need to be notified of the service availability by service announcements. Then they need to join the corresponding multicast group by sending joining messages. From the network point of view, the same content can be provided in a Point-to-Point fashion if there are not enough users to justify the high power transmission of the Point-to-Multipoint channel. Unlike the broadcast, users are expected to be charged for multicast service.

As shown in Figure 1.8, there are several procedures to enable MBMS multicast service. The subscription, joining and leaving phase are performed individually for each single user. The other procedures are performed for all users interested in a multicast service. The sequence of procedures may repeat if it is necessary.
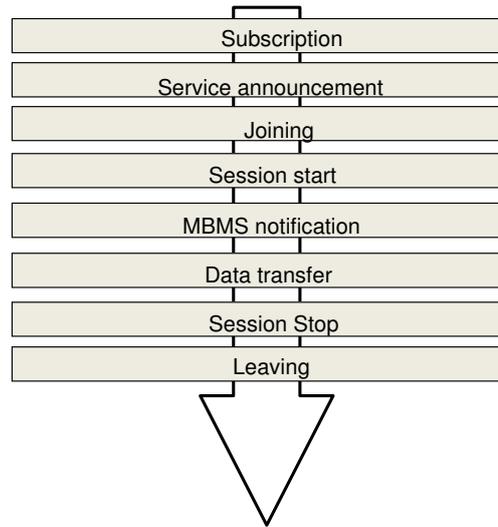
Figure 1.8: Multicast procedure [16]

– **Subscription:**   The subscription is the agreement of a user to receive service(s) offered by the operator, which allows the user to receive the related MBMS multicast service in further transmission. The subscription information is recorded in the BM-SC. This information allows to establish the relationship between users and the service provider.

– **Service announcement:**   By receiving MBMS service announcement, users shall discover the range of MBMS services and service availability. This mechanism distributes to users information about the service content, parameters required for service activation (e.g. IP multicast address(es)) and possibly other service related parameters (e.g. service start time).

– **Joining:** The joining is an activation message sent by the UE. In this procedure, the UE indicates to the network that it wants to receive multicast mode data of a specific MBMS service. This activation procedure allows a subscriber to become a member of a multicast group, and to be recorded by the network.

– **Session Start:** This step is the point at which the BM-SC is ready to send data. A session start message is sent for each MBMS bearer service. This can be identified with the start of a multicast session. Session start occurs independently of activation of the service by the user, which means, a given user may join the service before or after session start. Session start is the trigger for radio/network resource establishment for MBMS data transfer.

– **MBMS notification:**   It informs the UEs about forthcoming (and potentially about ongoing) MBMS multicast data transfer.

– **Data transfer:** In this procedure, MBMS data are transferred to UEs. The user data and control messages are transmitted, as well as the error recovery packets if it is necessary (e.g. fading conditions or handover).

  – **Session Stop:** This step is the point at which there will be no more data to
    send for a period of time. The bearer resources are released at session stop.
  – **Leaving:** In this process a subscriber leaves, i.e. stops being a member. Then
    MBMS multicast is deactivated by the user.

Figure 1.9 shows a timeline example of the multicast service procedures, in which
two users join and receive data flow from the same service sequentially. It is il-
lustrated that the subscription, joining, leaving, service announcements as well as
MBMS notification can run in parallel to other phases. Moreover, there are three
important periods related with the management of radio resource for MBMS service.
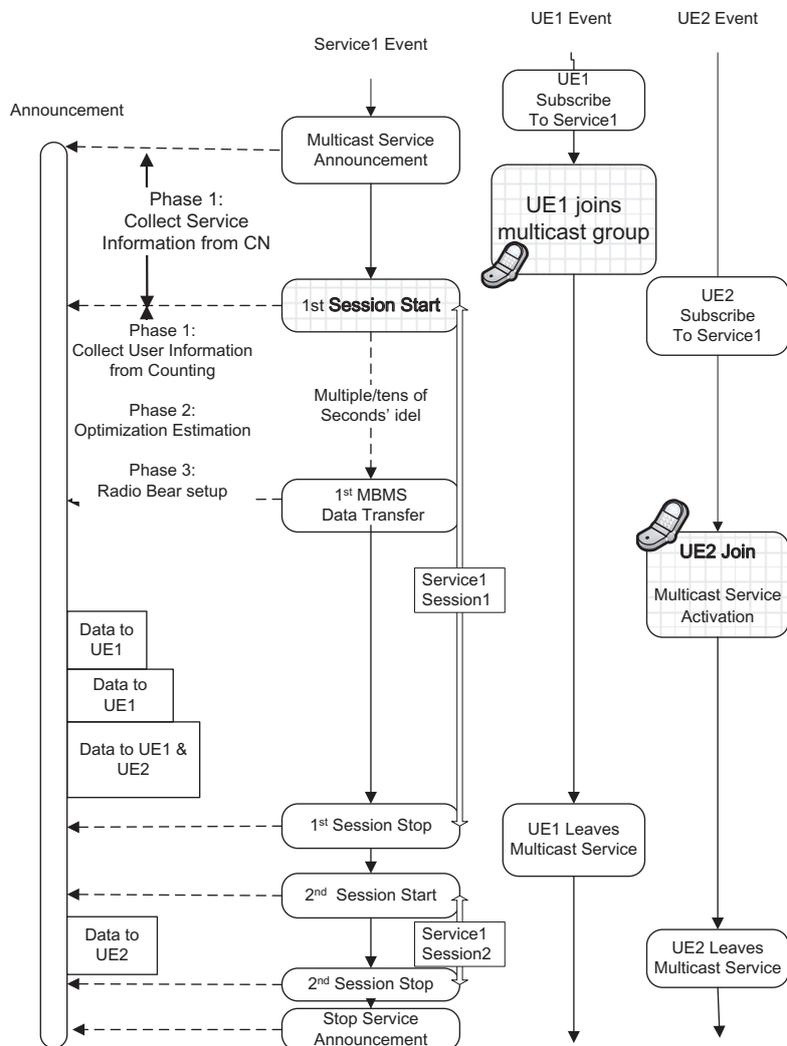
Figure 1.9: An example of two users receiving MBMS multicast service

– **Period between Service Announcement and Joining.**   The Joining
time depends on the user's choice and the possible user is in response to a
service announcement. Users will typically join at a chosen time so that the
period between announcement and joining may be very long or very short. In
order to avoid overload situations being caused by many users attempting to
join in a short period of time, the UE shall be able to use parameters sent by
the BM-SC in the service announcement to randomize the joining time.

– **Period between Joining and Session Start.**   Some MBMS multicast
services may be 'always on'. In this case, Joining can take place immediately
after Service Announcement and possibly many hours before, or after, the
session start. In other cases, if a Session Start time is known, Joining may
take place immediately before Session Start or after Session Start. For these
services, the announcement may contain some indication of a time period at
which users and UEs should use to choose a time to Join the MBMS bearer
service.

– **Period between Session Start and First Data Arrival.**   Session Start
indicates that the transmission is about to start, then the network actions
will take place for the arrival of first data. Therefore, the time delay between
a Session Start indication and actual data arrival should be long enough for
these network actions, e.g. provision of service information to the UTRAN,
and establishment of the network and radio resources. Session Start may be
triggered by an explicit notification from the BM-SC.

1.2.3   MBMS transmission modes

MBMS defines two transmission modes in UTRAN to provide MBMS multicast
service: point-to-point transmission (PTP) mode and point-to-multipoint trans-
mission (PTM) mode. To carry the relevant MBMS data and signaling through
PTM mode, three new logical channels are added to Release 6. They are MBMS
Control Channel (MCCH) carrying MBMS control signaling; MBMS Traffic Chan-
nel (MTCH) carrying MBMS application data; and MBMS Scheduling Channel
(MSCH) carrying MBMS scheduling information to support discontinuous recep-
tion in the UE. MTCH is used to carry MBMS multicast data through PTM mode.

MTCH is a logical channel specifically used for a PTM downlink transmission
of user plane information between network and UEs. The user plane information
on MTCH is MBMS service specific and is sent to UEs in a cell with an activated
MBMS service [18]. MTCH maps to one transport channel named Forward Access
Channel (FACH) and then to physical channel named Secondary Common Control
Physical Channel (SCCPCH) in the downlink direction. Initially, FACH is a down-
link transport channel that carries control information to terminals known to be
located in the given cell. This is used, for example, after a random access message
has been received by the base station through a Random Access Channel (RACH).
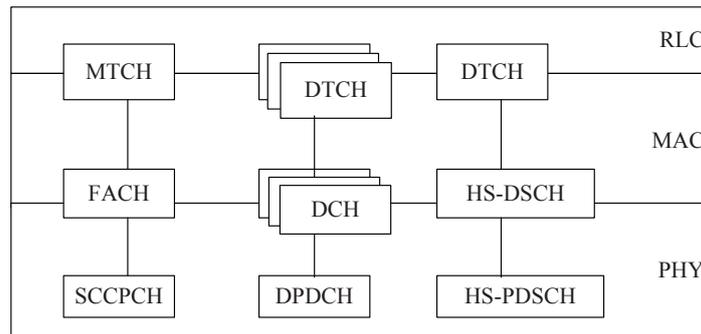
Figure 1.10: MBMS transmission modes and channel mapping

Currently, MBMS defines the PTM transmission mode which supports to transmit data packets on FACH. It aims at overcoming network congestion when a larger number of users request the same service. For each service, only one FACH is required for the transmission of service stream, while no traffic load on the uplink connections is required. FACH does not use fast power control (see section 1.1.1.2), and the transmitted messages should include band identification information to ensure their correct reception. On one hand, the reliability of FACH transmission is less than channels using a feedback channel for receiving quality. On the other hand, PTM mode needs rather high-power level to reach all users in the cell since it lacks of physical layer feedback in uplink. One or several FACH(s) are carried on the secondary common control physical channel (S-CCPCH); each FACH is sent with a fixed data rate (depending on the traffic bandwidth).

The PTP mode uses the logical channel named Dedicated Traffic Channel (DTCH). DTCH is defined in the user plane and transfers the information of a given service dedicated to a single user. It exists both in the uplink and downlink direction. Different DTCH may coexist for a given UE whenever several services are provided simultaneously (e.g. data and voice connections). Each DTCH serves one dedicated UE, and maps to transport channel Dedicated Channel (DCH) then over the downlink physical channel Downlink Physical Dedicated Channel (DPDCH). DCH is bi-directional with inner and outer loop power control. Because DCH is a Point-to-Point channel, multiple DCHs are required for transferring common data to a multicast group. Since DCH can employ fast closed-loop power control, it can achieve a highly reliable transmission quality. DCH also consists of an uplink channel, which is used to feedback power control information. RNC could control the state transitions between PTP and PTM modes which allows the radio network to keep the efficient power utilization state. Hence it needs to transfer the traffic volume between Cell_FACH and Cell_DCH and vice versa.

1.3   High Speed Downlink Packet Access (HSDPA)

High Speed Downlink Packet Access (HSDPA) [54] and MBMS are two important aspects of the UMTS network evolution. HSDPA is introduced in 3GPP Release 5, it defines a new channel named High-Speed Downlink Shared CHannel (HS-DSCH). The HS-DSCH is a downlink transport channel shared by a number of UEs, it could be transmitted over only a portion of the cell using smart antennas. Figure 1.10 shows the channel mapping of HS-DSCH. HS-DSCH maps to DTCH logical channels above it and to a HS-PDSCH physical channel below it.

Feature	FACH	DCH	HS-DSCH
Spreading factor	4 to 256	4 to 512	16
Modulation	QPSK	QPSK	QPSK/16QAM
Soft handover	No	Yes	No
Interleaving	10-80 ms	10-80 ms	2 ms
Multi-code operation	No	Yes	Yes, extended
Hybrid ARQ	RLC level	RLC level	L1 packet combining
Fast power control	No	Yes	Only for associated DCH
Adaptive Modulation and Coding (AMC)	No	No	Yes
Base Transceiver Station (BTS) scheduling	No	No	Yes
Specification	Release 99	Release 99	Release 5

Table 1.2: Comparison of properties of transport channels

Table 1.2 illustrates the technologies used in HS-DSCH and compares it with R99 channels [54]. The upgrade of HSDPA to the existing WCDMA network is a software upgrade, which helps the Node B to obtain new functionalities supporting HSDPA: the fast scheduling based on the quality feedback, UE capability, buffer status, etc. HSDPA disables two features of WCDMA: the variable SF (SF=16 for HSDPA) and fast power control. Instead, the utilization of Adaptive Modulation and Coding (AMC), extensive multi-code operation and a fast efficient retransmission strategy (BTS scheduling), helps HSDPA to select a coding and modulation combination for the users. To benefit from the dynamic range of HSDPA link adaptation, a user may simultaneously utilize up to 15 multi-codes in parallel. When link conditions

are very favorable, based on the scheduling decisions done in the Node B, most of the cell capacity may be allocated to one user for a very short time. In this way, additional user throughout could be achieved, in general for free. The peak rate of HS-DSCH is up to 10 Mbps with 16 quadrature amplitude modulation (QAM). Therefore, HS-DSCH provides a more flexible and efficient method for utilizing radio resource management to achieve significant improvements on the downlink capacity, reduced network latency and higher data rates for packet data services.

Although Release 99 transport channels involving FACH and DCH, have already been standardized for MBMS multicast transmission, MBMS over HS-DSCH is still an interesting research topic [73, 80, 25, 68, 31].

1.4 State of the art for MBMS RRM

This section introduces the related work in the study of radio resource allocation for MBMS. In UTRAN, the sharing of limited radio resources among numerous users per cell is constrained with more services subscriptions and higher requested traffic bandwidth. The RNC is responsible for the efficiency of transmission scheme selection. Before MBMS data transfer (see Figure 1.8), in the period between Joining and Session Start, the RNC needs to choose the appropriate transmission schemes and relevant radio resource allocation. In the period between Session Start and First Data Arrival, the radio bearers are established for data transmission, aiming to achieve an efficient overall utilization of radio and network resources. Then during the MBMS session, the RNC should adapt to continuous changes in the dynamic wireless environments, optimally allocate radio resources and satisfy the service requirement in real-time, during which, the switch of MBMS transmission modes is crucial to the allocation efficiency.

In particular, the Node B transmission power is a limited resource and should be shared among all MBMS multicast users in a cell. Hence the power control is essential in order to minimize the power consumption, thus to eliminate intercell interference and reserve cell capacity. When the number of subscribers for multicast services and traffic requirement increases, the main concern in the development of radio resource management for MBMS session is to serve the purpose of power saving (with or without lower QoS). The existing MBMS RRM techniques are presented in following paragraphs.

1.4.1 MBMS UE counting

The 3GPP designed the MBMS UE Counting mechanism in TS 25.346 [18]. The UE counting supports to determine the switching threshold between PTP and PTM

modes based on the number of MBMS subscribers. This function is performed by the RNC before MBMS data transfer. First, the RNC sends counting messages to users in a given cell, then identifies the numbers of users by received counting response messages. When the number of users that wish to receive a multicast session for a particular service is below an operator-defined threshold, the RNC will establish PTP connections through the DCH channel(s). During MBMS service transmission, the switch from PTP to PTM resources should occur, when the user numbers exceeds the predefined threshold, and vice versa. The study in [37] claimed that the threshold is 7 UEs per cell, with the assumption that the FACH transmission power is set to 4 W. While in [56] the threshold is 5 UEs.

However, since the PTP transmission power would be different for different geographic distribution of users, to determine the appropriate radio bearer only based on the number of users is simply to implement but not sufficient.

1.4.2   MBMS power counting

The inefficiency of the UE counting and the power limitation motivated the 3GPP to define the MBMS Power Counting (MPC) mechanism in TR 25.922 [15]. This function aims to minimize the Node B power requirements during the transmission. The study under these assumptions is presented in [55], where the authors first consider that the PTM transmission power remains as the same level to cover the whole cell. Then the switching point between PTP and PTM modes is based on the power consumption of PTP mode. When the estimated power consumption of PTP mode in a cell is less than that for PTM, the network establishes PTP connections. The switch from PTP to PTM occurs when the power exceeds the threshold, and vice versa. Furthermore, the study in [24] proposes a power control scheme for the MBMS transmission channel selection among PTM and PTP modes. Later this work is analyzed in micro and macrocell environments [22].

MPC has limited flexibility because it only considers delivering service for all users with full service quality; it does not support PTP and PTM for one service concurrently. Therefore, when MBMS transmission power in one cell is near saturation, MPC does not provide alternative allocation scheme to save resource for new service or new users to access into the network. These alternative schemes involve reducing power consumption by decreasing service's quality, or applying flexible power allocation for PTM mode.

1.4.3   MBMS FACH enhancements

FACH is the only PTM channel that carries MTCH traffic. In conventional PTM downlink transmissions, FACH does not support the fast power control. Therefore,

in order to be received by all UEs throughout the cell, each FACH should be established at a fixed power level high enough to ensure the requested QoS in the whole cell. The following studies aim to increase the FACH efficiency.

1.4.3.1   Physical layer enhancement

As FACH power is limited, FACH transmission efficiency strongly depends on the maximizing diversity. In [67], the authors propose a longer Transmission Time Interval (TTI), using 80 ms instead of 20 ms as the FACH TTI for MBMS transmission, thus to provide time diversity against fast fading. While longer TTI introduces more complexity and larger memory space requirement in user side, to obtain macro diversity in which a user receives the same signal from multiple transmitters located in different cells, the combing transmissions (soft and selection combing) is proposed from multiple cells [67]. Besides, the Spatio-Temporal Transmit Diversity (STTD) processing techniques exploit diversity in both spatial and temporal domains. It assumes two transmitting antennas and a single data stream in order to improve the signal quality and reduce the power requirement. These statements are confirmed through analytical investigations in [12].

1.4.3.2   Dynamic power setting

Dynamic Power Setting (DPS) for PTM mode was initially proposed in [78]. Instead of fixing the FACH power to cover the whole cell, the RNC dynamically adjusts the FACH power to just achieve the worst users of multicast group. This technology aims to save the power efficiently in PTM mode. To support DPS, the MBMS users need to turn on measurement report mechanism while they are on the Cell_FACH state. Based on such dynamic and periodic measurement reports, the Node B adjusts the transmission power for FACH [28].

1.4.3.3   Scalable FACH transmission

The multicast service delivery quality can be improved by adapting the scalability ratio (bit rate and frame rate) of multimedia streams through different coding structures. This technology is also named as rate splitting because the service can be split into several streams. The transmission schemes of multicast service are divided into single layer (SL) and multilayer (ML) transmission schemes [52]. For the ML scheme, the service is scalable encoded by several flows; each flow has lower bitrate and QoS requirements comparing to a nonscalable stream.

The scalable technology, integrated with FACH transmission (named S-FACH in this work), is initially proposed in [79]. Then it is further studied in [31]. In

the scalable FACH transmission technology, to provide basic service quality, the most important stream (basic flow) is sent to all the users in the cell; then the less important streams (advanced flows) are only sent to users who have better link qualities. For these users (e.g. the users near the Node B), the reception of advanced flows can enhance the service quality on top of the basic flow. The authors in [23] studied the performance of scalable FACH transmission in cells with site-to-site distance of 1 km. The simulation shows that a single 64 kbps stream carried by one FACH channel requires the power of 7.6 W to cover the whole cell (95%). While for a double stream transmission, only 5.8 W is required for transmitting two data streams of 32 kbps through two FACHs, in which the basic flow with 95 % coverage requests the power of 4.0 W, and 1.8 W for the advanced flow covering 50 % of the cell. Thus, 1.8 W can be saved through the Scalable FACH transmission.

However, the scalable FACH transmission involves certain negative results. Some users which are near the cell edge will not be fully satisfied, as they only receive the basic service quality. Besides, the utilization of the scalable transmission for MBMS multicast is less flexible as it only supports PTM mode and with fixed coverage under static assumption [79, 31]. Therefore, the gain of S-FACH is rather limited as the Node B could not able to weigh the transmission power consumption and the users' satisfaction.

1.4.4   Dual transmission mode

Dual Transmission Mode (DTM) allows the co-existing usage of PTP and PTM mode for one MBMS service [28]. It adapts the FACH coverage for users with better link quality, while the users near the cell edge are served by DCHs. DTM is active both before and during MBMS transmission. The FACH coverage is dynamically adapted by changing its transmission power (i.e. Dynamic Power Setting for FACH), meanwhile the DCH connections are released or established for the rest of users. DTM enriches the candidate transmission modes for MBMS. The advantage of DTM is obvious during handover for single user. However, DTM does not consider the integration with scalable transmission, and the simulation in [9] concluded that DTM is only beneficial for up to 5 users with PTP connections. Therefore, only applying FACH and DCH co-existing for transmission modes is limited not only in power efficiency, but also in terms of channelization codes usage [24].

1.4.5   MBMS over HSDPA

The HSDPA as a means to deliver MBMS streaming is investigated by many researchers. The studies mainly focus on the scheduling and multi-resolution aspects (modulation and coding) [73, 31]. It is proved that with suitable packet scheduler algorithm and hierarchical QAM constellations, MBMS over HSDPA can achieve a

good fairness and capacity. Besides, some literature have studied the power utilization of MBMS transmission through HS-DSCH [68, 80, 25]. There are two methods to allocate HS-DSCH power for carrying MBMS multicast data. The first method considers the HS-DSCH transmission with fixed amount of power (e.g. 7 W [23, 31]). Then when the estimated power of FACH or DCH transmission will be more than 7 W, the multicast service transmission will be carried by HS-DSCH. Otherwise, when the estimated FACH or DCH transmission power are less than 7 W, the transmission mode switch will be triggered, to transfer the multicast data by FACH or DCH, depending on which channel consumes less power. In the second method, after the power for common control channel and signaling channel are allocated, the Node B will allocate the rest of available power to HS-DSCH . In this method, the HS-DSCH power will dynamically change based on the available transmission power and the serving service and users. In the study of HS-DSCH for MBMS [25, 68], the second method is used to estimate the required HS-DSCH power, then to compare it with the power of FACH and DCH, and to select the proper transmission mode by using the mode which is less power consuming. But none of these literature has considered the co-existing usage with the UMTS R99 channels.

1.4.6   Literature analysis and motivation

Although MBMS RRM in 3G network and HSDPA has been well studied, there are three main drawbacks in the existing allocation approaches. Firstly, the MBMS in 3G network has not been systematically analyzed, neither the problem complexity nor the problem characteristics. The existing approaches try to improve the RRM efficiency, rather than following systematic guidances. Therefore, almost all algorithms can only be compared with the standardized allocation approaches, while rare study compares their performance with competitive algorithms.

Secondly, all existing approaches only consider the allocation scheme for mono-objective, i.e. to minimize the consumption of transmission power. Therefore, only a one dimensional evaluation criterion is proposed in existing studies. However, the channelization code is also a crucial limitation in UTRAN, which impacts the cell capacity and data rate. While considering both these two radio resources will bring significant problem complexity in algorithm design as well as implementation.

Last but not least, the transmission modes in the existing approaches offer limited flexibility regarding the dynamic mobile environments and radio resources consumption. For example, when the transmission power is saturated, the transmission mode should be determined between transmission service i) through basic quality link with full coverage and ii) through advanced quality link with smaller coverage. When the transmission power is sufficient, meanwhile, the channel codes is also an important but limited radio resource. Therefore, in case of channel codes saturation, the transmission mode should be selected among i) less power consumption scheme

and ii) less occupation of channel codes.

To overcome these limitations mentioned above, we propose a Flexible Radio Resource Management Model (F2R2M) which combines the flexible transmission mode selection and the multimedia scalability. The contribution of this work is three-fold: Firstly, a general model of the MBMS RRM problem is provided, which abstracts the input parameter, decision parameter and optimization objectives. In this model, the mathematical formulation based on classical combinatorial optimization problem is developed. Then the fitness landscape analysis is conducted, which proves the problem difficulties and guides the optimization strategies.

Secondly, based on the proposed model and the problem properties, we propose an allocation procedure with two-dimensional objective along with a lexicographic-order evaluation criteria to evaluate the quality of resources allocation in terms of service satisfaction and resource consumption. Hence the allocation approach is able to minimize the power consumption, and to save the channel code occupation. Finding the optimal solution which efficiently allocates the radio resource for MBMS multicast service, while guarantees the Quality of Service (QoS), maintains the service coverage and offers the cell capacity as high as possible.

Thirdly, our allocation approach offers higher flexibility in terms of allocation efficiency. It can self adapt to the dynamic user scenario in the mobile environment: the user number, the user distributions and the service requests which are different from time to time. The proposed allocation approach is implemented, and the other approaches are reproduced, then they are applied to the same scenarios in the same platform; in this way, we offer a general test-bed to compare these algorithms under common situation.

1.5 Synthesis

This chapter provides the background knowledge and the existing literature study related to the research area in this thesis manuscript.

The first section presented the fundamentals of UMTS. UMTS uses WCDMA as the radio access technology, which subdivides the radio resource in code domain for simultaneous transmissions over the radio link. The high level architecture of UMTS includes UE, UTRAN and CN. UTRAN takes in charge the radio resource establishment, reconfiguration and release. In these procedures, the channelization codes and the transmission power of channels are limited resource in UTRAN.

The second section introduced the MBMS framework and its features since the Release 6. MBMS provides an efficient distribution platform for the multicast multimedia services transmission over mobile networks. MBMS defines the logical channel

MTCH to support the PTM transmission mode for multicast service, which saves the unnecessary radio resource in UTRAN. MBMS also supports the PTP mode for multicast, which is carried by DCH channel. Following by which the HS-DSCH transport channel defined in HSDPA is introduced. Then HS-DSCH is compared with the traditional MBMS transmission modes. The HS-DSCH feature is analyzed as an open issue for MBMS multicast carrier.

Based on the background knowledge, the related literature work in the field of MBMS RRM study has been introduced. The advantages and the disadvantages of these approaches are analyzed. It is found that the existing studies have the following shortcomings:

1. Lack of systematic modeling of the targeting problem.

2. Lack of systematic analysis of the solution space for the targeting problem, therefore they can not prove whether the achieved radio resource assignment is optimal or not.

3. Lack of enough flexibility and therefore they can only be beneficial in some scenarios but they have limitations in other scenarios.

With these in mind, the motivation of this work is declared:

1. To propose a general mathematical model to allow in-depth analysis for the MBMS RRM problem. The model should also be able to provide enough flexibility to enable efficient radio resource allocation schemes.

2. To systematically analyze the problem complexity and the structure of the solution space, so that the algorithm performance can be better evaluated.

3. To propose novel and practical algorithms, which aim to find the optimal radio resource assignments for arbitrary user scenarios in dynamic mobile environment, considering both the throughput requirement and the power consumption limitation.

The proposed flexible radio resource management model for MBMS multicast service will be detailed in chapter 2.

CHAPTER 2

F2R2M: A Flexible Model for RRM of MBMS

This chapter will describe the first contribution: a general modeling of MBMS RRM problem named the Flexible Radio Resource Management Model (F2R2M). Based on this model, the target is to perform dynamic radio resource allocation for multicast service by using optimization approach. In terms of allocation scheme, this model provides the abstraction for the mixed usage of three transport channels, as well as the abstraction for the scalable encoded multimedia stream, hence it offers the flexibility and possibility for high efficient radio resource utilization, and it satisfies the search complexity. In this model, a two-dimensional cost function is proposed which reflects the optimization for both the power consumption and the quality of service. A lexicographic evaluation criterion is also proposed to allow us to meet the throughput requirement while at the same time to minimize the transmission power. Finally we obtain solution with proper service quality, low power and channelization code consumption, hence minimizing the cell interference and maximizing the cell capacity.

Contents

2.1 Model framework

The flow chart in Figure 2.1 shows the establishment of network and radio resource during the MBMS service provision procedure. The procedures are mapped into three phases in our model: the parameter collection phase, the estimation phase, and the resource allocation phase. The radio resource management and the allocation process are supposed to be operated in RNC for each Node B.
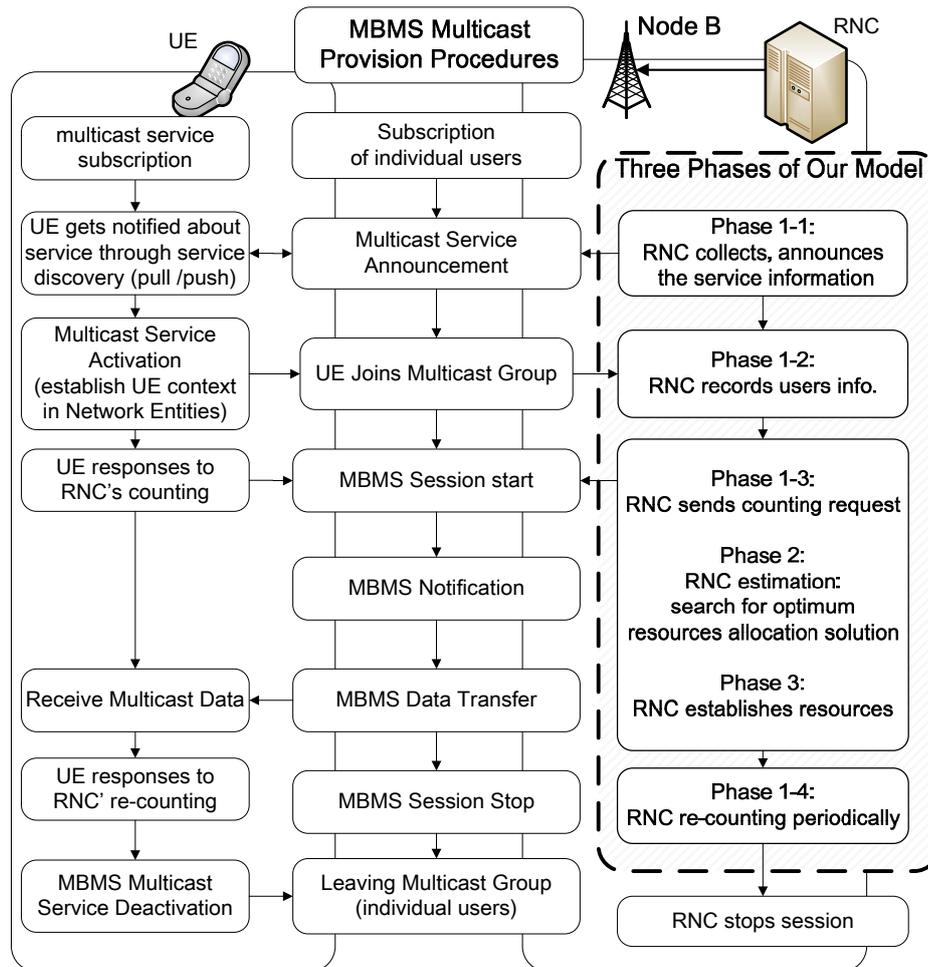
Figure 2.1: A three phase framework

2.1.1 Phase 1: Parameter collection phase

Before the MBMS session starts, the RNC collects the MBMS service and the UE information in the first phase: the parameter collection phase. More particularly, this phase consists of four different sub steps: phase 1-1, phase 1-2, phase 1-3 and

phase 1-4.

The phase 1-1 is active when the network distributes the multicast service announcement to users. The RNC announces to users the service availability and collects the service content in the service announcement distribution procedure. The announcement messages include the parameters required for service activation, e.g. IP multicast address(es), and other parameters which are service related, e.g. service bandwidth, encoded technologies and service start time. Then the phase 1-2 is performed when UE are joining the session, such that the RNC records the user membership from the UE activation message. In the subsequent phase 1-3, the user information is complemented in the MBMS counting procedure, by which the network sends the MBMS counting request and collects counting response from the UE. In this step, the context reporting process is required [21], which reports the current location of UE. The collection phase is repeated periodically if necessary. The phase 1-4, named periodic re-counting process will be triggered by any change of the MBMS session state, e.g. user mobility or service announcement.

In this first phase, the following parameters are collected as the input variables of model.

- UE information
  - $T(c) = \{t_1, \ldots, t_k\}$. Set of mobile terminals in cell c. The terminal indexes are ordered from the closest to the farthest distance from the Node B at any time slot (after each TTI).
  - $D(c) = \{d_1, \ldots, d_k\}$, $t_k \in T(c)$. Set of instantaneous distances from the Node B to the terminal, this value can be obtained through the channel quality measurement report from UE side [78].
- Service information
  - $S(c) = \{s_1, \ldots, s_{N_s}\}$. Set of services to be transmitted to multicast groups located in cell c. The total number of service is N_s.
  - $F(s_i)$. The flow set of service s_i and flow bandwidth, $s_i \in S(c)$. $F(s_i) = \{f_{s_i,0}\}$ if s_i is a single layer (SL) transmission scheme service. Or $F(s_i) = \{f_{s_i,1}, f_{s_i,2}, [f_{s_i,3}]\}$ if s_i is a multilayer (ML) transmission scheme service, where $f_{s_i,1}$ is the basic flow and $f_{s_i,j}$ with $j > 1$ are advanced flows. Let's assume that each service has $N_f(s_i)$ flows, and N_f is the total number of flows, $N_f = \sum_{i=1}^{N_s} N_f(s_i)$.
  - $M(s_i) \subseteq T(c)$. Multicast group of service s_i. One group includes all the terminals $\{t_1, \ldots, t_{k'}\}$ requiring the same service content with $k' \le k$. The number of terminals in each group is $N_t(s_i) = N_t(f_{s_i,j})$, $\forall f_{s_i,j} \in F(s_i)$, and N_t is the total number of terminals in the cell, $N_t = \sum_{i=1}^{N_s} N_t(s_i)$.

2.1.2 Phase 2: Estimation phase

A MBMS multicast procedure named session start is the trigger of radio and network resource establishment for data transfer. During the interval between session start and user joining, the RNC conducts the estimation phase, which adopts the obtained UE and service information in phase 1 and estimates the best assignment of radio resources to start the MBMS session. The estimated radio resource assignment is called a solution. Therefore, the estimation phase plays a key role for radio resource allocation performance for the total framework. By modeling this phase, the MBMS RRM problem can now be abstracted as an optimization problem, which tries to find the best radio resource assignment in the estimation phase, based on a given search space. The search space is determined by the UE and service information obtained in the collection phase. The way to obtain the estimated best found solution can be deterministically computed like previously mentioned in the state of art. However, it can be obtained by adaptive and iterative approaches which are proposed in our model. The detailed description of our optimization process in the estimation phase will be presented in section 2.2. In order to better help to explore the search space, two additional flexibilities for the radio resource in F2R2M will be explained below: the flow-based channel assignment and the flexible transmission mode selection.

2.1.2.1 Flow-based channel assignment

To fully utilize the allocation efficiency for scalable encoded multicast service, the transport channels are allocated for each flow transmission, unlike the service-based channel allocation in existing approaches. For single layer transmission scheme service, which has only one flow f_0, the channel allocation is determined for f_0. For multilayer transmission stream service, the multimedia stream is divided into two or more flows, then the transport channels are allocated for each flow f_j, $j \geq 1$.

2.1.2.2 Flexible transmission mode selection

F2R2M supports the flexible transmission mode selection, meaning different users can be assigned with different channels based on their service requests and their link quality.

The possible assignments of transport channels are summarized in Table 2.1. The proposed channel types mainly follow UMTS R99 and HSDPA but also contain the innovation to allow the coexisting of PTM and PTP transmission modes for each flow.

Table 2.1: Candidate transport channel assignment

	Transmission mode	Candidate transport channels
Existing transmission modes	pure PTM	FACH
	pure PTP	DCH
		HS-DSCH
Proposed (new) transmission modes	mix of PTP	one HS-DSCH and m × DCH
	mix of PTP and PTM	FACH and m × DCH
		FACH and n × HS-DSCH
		FACH, m × DCH and n × HS-DSCH

2.1.3   Phase 3: Resource allocation phase

Finally, according to the best found solution in the estimation phase, the RNC establishes the transport channels for selected UEs, then allocates the planned power and the spreading codes for the transmission channels, which have been determined. After that, the multicast session will start based on the established radio bearers.

2.2   Model abstraction

This section explicates the details in the estimation phase. Regarding the flexible channel assignment, the RNC needs to determine the transmission mode for each flow, and which users are served through which channel. The final solution should satisfy the throughput requirement, while it is constrained with the power and the channel codes limitation. Therefore, in the estimation phase, which selects the transmission mode for each user and each flow, there are three types of decision variable to consider: the UE partition, which selects the transmission mode for each user and each flow, the channel code allocation, and the power allocation. Also in the studied model, an iterative architecture for the best solution search is proposed. As shown in Figure 2.2, these decision variables are initialized as input solution to the search model, then the search procedure is conducted iteratively through seven modules until the stop criteria is satisfied. Then the final solution is output to the allocation phase.

2.2.1   UE partition search engine

The UE partition search engine is responsible for generating a neighborhood solution from the current found solution. This part contains the modeling of the
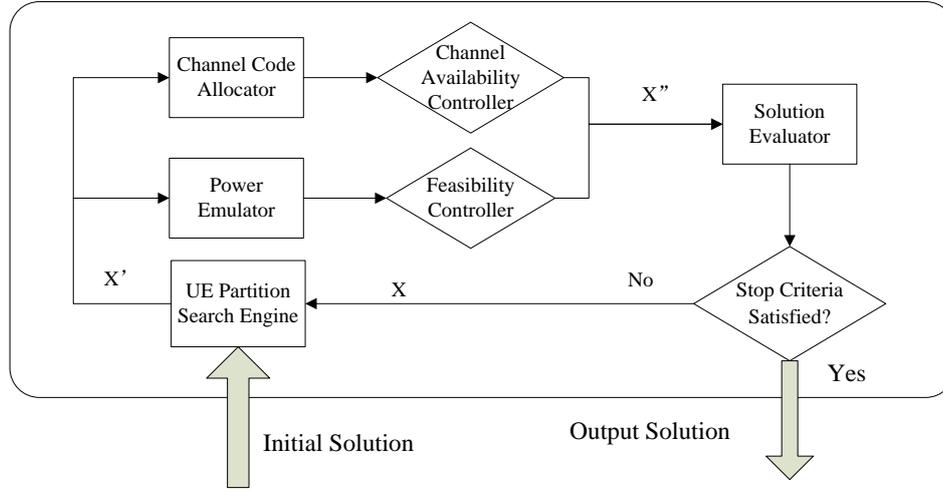
Figure 2.2: Model abstraction in the estimation phase

transports channels as well as the algorithm to generate the neighborhood solutions. It introduces the solution variations and drives the search procedure. The way to generate the neighborhood solution is called neighborhood operator.

2.2.1.1 Transport channels

The selected transmission mode for each flow should indicate not only the selected transport channel(s) through which the users receive the flow content, but also the transmission destination (i.e. selected users) for each channel. To represent such allocation scheme, the concept of user channel set $UE_{ch}(f_{s,j})$ is adopted. Each user set includes the served users through the channel defined by ch. Therefore, the task of the UE partition search engine, is to partition multicast group into several user sets, where each set reflects the served users through the corresponding transport channel(s). More particular, for each flow $f_{s,j}$ of service s with flow index j, the UE search engine partitions the service's multicast group $M(s)$ into four disjointed sets:

  – $UE_{fach}(f_{s,j})$: users served through a FACH.
  – $UE_{dch}(f_{s,j})$: users served through DCHs.
  – $UE_{hs}(f_{s,j})$: users sharing a HS-DSCH.
  – $UE_{noch}(f_{s,j})$: non-served users.

$UE_{noch}(f_{s,j})$ is modeled for two purposes. On one hand, it is to distinguish between the users that belong to multicast group $M(s)$ but not selected for flow reception, and the users in cell but not belong to $M(s)$. On the other hand, it is convenient for the mathematical expression of search operation, as the user not

selected for flow reception is also a part of the radio resource assignment solution. For $UE_{ch}(f_{s,j})$ with $ch \in \{fach, dch, hs\}$, an empty user set means this channel is not selected for flow transmission. If $f_{s,j}$ is not selected for transmission, then its subscribed users are all moved into the set of $UE_{noch}(f_{s,j})$. This operation generally only happens for advanced flow solution operation.

To guarantee the service coverage, two basic principles are defined. First, all users in multicast group should be selected to receive $f_{s,0}$ or $f_{s,1}$. This principle is defined as a constraint in Equation 2.1a. In Equation 2.1b, $R(f_{s,j})$ represents the users who are receiving $f_{s,j}$. The principle defined in Equation 2.1c restricts that the advanced flow is only sent to users who also receive the lower level flow, that is to avoid the redundant content transfer to the same user.

$$\forall s, j \in \{0, 1\}, R(f_{s,j}) = M(s) \tag{2.1a}$$

$$R(f_{s,j}) = UE_{fach}(f_{s,j}) \cup UE_{dch}(f_{s,j}) \cup UE_{hs}(f_{s,j}) \tag{2.1b}$$

$$\forall s, j \geq 2, R(f_{s,j}) \subseteq R(f_{s,j-1}) \tag{2.1c}$$

Moreover, the user partition for $f_{s,j}$ should be in accord with the channel characteristics. If d_i is the distance of the user t_i from the base station, and d_{thr} is the distance threshold found during the optimization procedure, other constraints are defined in Equation 2.2.

$$\forall t_i \in UE_{fach}, d_i \leq d_{thr} \tag{2.2a}$$

$$\forall t_j \in UE_{dch} \cup UE_{hs} \cup UE_{noch}, d_j > d_{thr} \tag{2.2b}$$

$$\forall ch_m, ch_n \in \{fach, dch, hs, noch\}, UE_{ch_m} \cap UE_{ch_n} = \phi \tag{2.2c}$$

The constraint in Equation 2.2a is to guarantee that $UE_{fach}(f_{s,j})$ includes the closest users in the multicast group. That is because FACH is a common channel and can be listened by all users within its coverage. The constraint in Equation 2.2b means that the users in multicast group, farther than the FACH coverage, are assigned to HS-DSCH or DCH. When there is no available channel code for a given user, this user is switched to $UE_{noch}(f_{s,j})$. The constraint in Equation 2.2c guarantees that the user sets for each flow do not overlap as sending the same flow to the same user through more than one channel will waste resource.

In the proposed model, the user partition is initialized randomly or by predefined method, and then is modified during each iteration of the optimization procedure. The modification strategies will be described in the next section.

2.2.1.2 Neighborhood Operators

As shown in Figure 2.2, at each iteration, the UE partition search engine is applied to a new search partition. This search is done through a neighborhood

operation function which moves the current solution x into a neighbor solution x'. The neighborhood operator is the key element that leads to different optimization performance. In this section, two neighborhood operators are proposed, which define two associated neighborhood structures.

Single insert operator. The single insert operator δ_{SI} moves only one user from an initial channel ch_o to a target channel ch_i for each operator application. It randomly selects a user t_j and moves it from UE_{ch_o} to UE_{ch_i}. If FACH is one of the channel ch_o or ch_i, then t_j is determined by the following principles:

– From FACH: t_j is the farthest user within UE_{fach}, i.e. it reduces the FACH coverage as well as its power consumption.
– To FACH: t_j is the nearest user within UE_{hs} or UE_{dch} or UE_{noch}, i.e. it minimizes the additional power consumption for FACH.

The δ_{SI} operation is implemented through the following three steps:

1. Choose randomly one channel set ch_o, $UE_{ch_o} \neq \phi$, UE_{ch_o} is the "output" user set.

2. Select randomly another channel set UE_{ch_i} as the "input" user set, $ch_i \neq ch_o$.

3. Select randomly user t_j from UE_{ch_o}, δ_{SI} moves t_j from UE_{ch_o} to UE_{ch_i}.

Table 2.2: Single insert operation example

Solution	UE_{fach}	UE_{dch}	UE_{hs}	UE_{noch}
x_1	$t_1, t_2, \underline{t_3}$	t_4, t_5, t_7, t_8	t_6, t_9	t_{10}, t_{11}
x_2	$t_1, \underline{\underline{t_2}}$	$\underline{t_3}, t_4, t_5, t_7, t_8$	t_6, t_9	t_{10}, t_{11}
x_3	t_1	$\underline{\underline{t_2}}, t_3, t_4, t_5, t_7, t_8$	$\underset{\sim}{t_6}, t_9$	t_{10}, t_{11}
x_4	t_1	$t_2, t_3, t_4, t_5, \underset{\sim}{t_6}, t_7, t_8$	t_9	t_{10}, t_{11}

Table 2.2 shows the example of δ_{SI} operations, which transfer solution x_1 to x_4. The first column lists the solution name, each row illustrates the UE partitions for one solution. The four channel user sets in the corresponding solution are listed from the second to the fifth columns. Therefore each table cell shows the terminals in one user channel set. The terminal index is arranged in ascending order of the distance from the Node B, e.g. the smaller index represents the closer user distance. To move the solution x_1 to x_4, three steps of single insert operator are performed:

– Move 1: $x_1 \rightarrow x_2$, ch_o = FACH, ch_i = DCH. The farthest user t_3 in UE_{fach} is moved to UE_{dch}.
– Move 2: $x_2 \rightarrow x_3$, ch_o = FACH, ch_i = DCH. The farthest user t_2 in UE_{fach} is moved to UE_{dch}.

– Move 3: $x_3 \rightarrow x_4$, ch_o = HS-DSCH, ch_i = DCH. t_6 is randomly selected.

Multiple insert operator.   The multiple insert operator δ_{MI} moves several users from an initial channel ch_o to a target channel ch_i. It randomly selects a user t_j and move a set of users from UE_{ch_o} to UE_{ch_i}. If FACH is one of the channel ch_o or ch_i then:

– If t_j moves from FACH, all users farther than t_j within UE_{fach} are moved to the chosen ch_i to reduce FACH power.
– If t_j moves to FACH, all users nearer than t_j can now hear from FACH, thus, they are inserted in UE_{fach} to reduce power consumption in ch_o.

δ_{MI} is implemented in three steps:

1. Choose randomly one channel ch_o, $UE_{ch_o} \neq \phi$, UE_{ch_o} is the "output" user set.

2. Select randomly another set UE_{ch_i} as the "input" user set, $ch_i \neq ch_o$.

3. Select randomly user t_j from UE_{ch_o}, δ_{MI} moves this single t_j or a block of users including t_j from UE_{ch_o} to UE_{ch_i}.

In the third step, the moved users depends on the chosen ch_i and ch_o. For example, once the algorithm decides to move t_j from UE_{hs} to UE_{fach}, the FACH coverage will be enlarged to t_j. In that case, all the users of the cell that are nearer than t_j can now hear from FACH, thus, no matter what user sets they are currently allocated at, they need to stay or be inserted in FACH user set. Therefore, once a user t_j is chosen to be moved to FACH set, firstly the user distributions served by the other channels are checked, then the users within the enlarged FACH coverage will be picked out to UE_{fach}. Finally, in that case we will have several users assigned to different channels all moving to the FACH all together. By contrast, once we decide to move one user t_j out of UE_{fach}, i.e. reducing the FACH coverage, then all users farther than t_j within UE_{fach} should be picked and moved to the chosen ch_i. In that case, one destination channel is considered for all moved users.

Table 2.3: Multiple insert operation example

Solution	UE_{fach}	UE_{dch}	UE_{hs}	UE_{noch}
x_1	$t_1, \underline{t_2, t_3}$	t_4, t_5, t_7, t_8	t_6, t_9	t_{10}, t_{11}
x_2	t_1	$\underline{t_2, t_3}, t_4, t_5, t_7, t_8$	$\underline{\underline{t_6}}, t_9$	t_{10}, t_{11}
x_3	t_1	$t_2, t_3, t_4, t_5, \underline{\underline{t_6}}, t_7, t_8$	t_9	t_{10}, t_{11}

Table 2.3 shows the example of δ_{MI} move operations. Two steps of multiple insert operator are conducted from solution x_1 to x_3:

– Move 1: $x_1 \rightarrow x_2$, ch_o = FACH, ch_i = DCH, $t_j = t_2$. Supposing that FACH coverage is reduced, t_3 is farther than t_2 from the base station, then both t_2 and t_3 are moved to DCH user set.

– Move 2: $x_2 \rightarrow x_3$, ch_o = HS-DSCH, ch_i = DCH, $t_j = t_6$. t_6 is ordered before t_7 as it is closer to the Node B than t_7.

By comparing with the example in Table 2.2, it is shown that δ_{MI} operator needs less operations than δ_{SI} to move the users, then there are two neighborhood structures as a Variable Neighborhood Search (VNS) algorithm.

2.2.2 Channel code allocator and availability control

When the UE partition search engine determines the user selection and the channel assignment, the channel code allocator then assigns available downlink orthogonal codes to nonempty user sets. The channelization codes are picked from the Orthogonal Variable Spreading Factor (OVSF) code tree (see Figure 1.5). The allocation procedure corresponds to the orthogonal principle of OVSF codes [6]: if one code on the OVSF tree is used, all codes underneath it are no longer usable. Hence the channel code with proper spreading factor should be selected according to the users in $UE_{ch}(f_{s,j})$, the requested flows bandwidth, and the occupation situation of OVSF code tree.

The use of OVSF codes allows the orthogonality between different spreading codes. We define $CH(f_{s,j}, t)$ to represent the channelization code allocated to user t for the receiving flow $f_{s,j}$. The selection of $CH(f_{s,j}, t)$ is related with the data rate of $f_{s,j}$. On one hand, the allocated channel bandwidth should be not higher than the flow bandwidth; this is to avoid unnecessary radio resource waste. On the other hand, the channel code allocator uses channelization codes from the branch with the smallest spreading factor satisfying the required bandwidth.

As shown in Table 2.4, the lowest used spreading factor in UMTS FDD downlink is SF=4, which corresponds to the raw symbol rate of 960 kbps, the channel bit rate of 1920 kbps (QPSK modulation) and the user bit rate of 768 kbps with 1/2.5 rate coding. Full rate voice is on SF=128, meaning a maximum of 128 simultaneous voice calls on a 5 Mhz carrier, which corresponds to the raw symbol rate of 30 kbps, the channel bit rate of 60 kbps and the user bit rate of 12.2 kbps with 1/5 rate coding.

The channel availability controller will check if all the channel codes are available in the OVSF code tree. If any channel code is already occupied for a given user, through a given channel, the related user will be moved to the non-served user set $UE_{noch}(f_{s,j})$.

Let's take the solution x_1 in Table 2.3 as an example. The spreading codes will

Table 2.4: Spreading factor and downlink user bit rate in UMTS FDD [5]

Spreading factor	Channel symbol rate (kbps)	QPSK channel bit rate (kbps)	Rate coding $1/k$ (k)	User bit rate (kbps)
4	960	1920	2.5	768
8	480	960	2.5	384
16	240	480	3.75	128
32	120	240	3.75	64
64	60	120	3.75	32
128	30	60	5	12.2 full rate voice
256	15	30	4	7.5 half rate voice
512	7.5	15	2	7.5 half rate voice

be assigned to users in x_1 to receive a flow with bandwidth of 64 kbps, with $1/3.75$ rate coding. For channel bandwidth of 64 kbps, the channel spreading factor is 32 (see Table 2.4). In the OVSF code tree, the root depth is 1, the channel code with SF=32 is at depth $\log_2 32 = 5$.

$$x_1 = \underbrace{t_1, t_2, t_3,}_{UE(fach)} \underbrace{t_4, t_5, t_7, t_8,}_{UE(dch)} \underbrace{t_6, t_9,}_{UE(hs)} \underbrace{t_{10}, t_{11}}_{UE(noch)} \tag{2.3}$$

As shown in Figure 2.3, there are 32 codes in depth 5, each code is represented as $C_{32,n}$, $n \in \{0, \ldots, 31\}$, where n is the code index.

The channel code allocator assigns the available channel codes to users in x_1 from the OVSF tree through the following procedure: firstly $C_{32,0}$ is assigned to the FACH because the UE_{fach} in x_1 is non-empty. Note that two OVSF codes are orthogonal if and only if neither code lies on the path from the other code to the root. Then the codes $C_{4,0}$, $C_{8,0}$, $C_{16,0}$, $C_{64,0}$, $C_{64,1}$, $C_{128,0}$, $C_{128,1}$ and so on, cannot be assigned to any other user in the same cell. Secondly, four channel codes are assigned for users in UE_{dch}. The codes $C_{32,1}$, $C_{32,2}$, $C_{32,3}$ and $C_{32,4}$ are then selected, each code for one dedicated channel serving one dedicated user. In consequence, the codes $C_{16,1}$, $C_{16,2}$ in lower depth and so on, and the codes $C_{64,2}, \ldots, C_{64,9}$ in higher depth and so on, are all blocked. Thirdly, there are two users in UE_{hs}; the total bandwidth requirement for HS-DSCH is $64 \times 2 = 128$ kbps. Because HS-DSCH only allocates channel codes with SF=16, $C_{16,3}$ is selected and assigned to HS-DSCH, which will be shared by users t_6 and t_9. Finally, the total allocated codes for x_1 are:

– $C_{32,0}$ for all mobiles in UE_{fach}
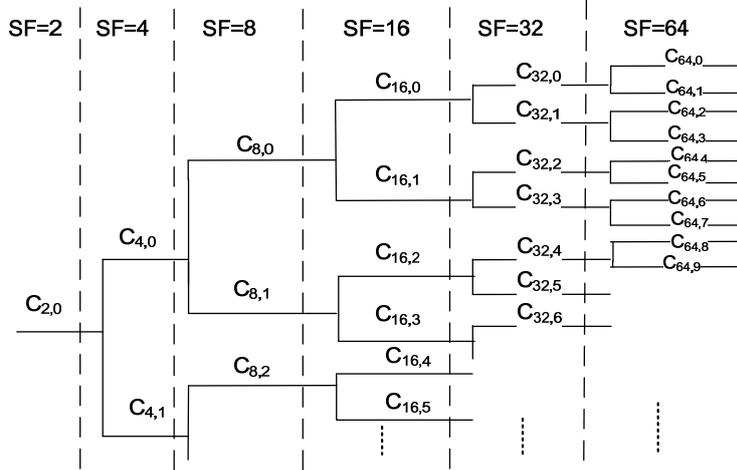
Figure 2.3: Allocation of spreading codes: an example

– $C_{32,1}$, $C_{32,2}$, $C_{32,3}$, $C_{32,4}$ to each mobile of UE_{dch} respectively.
– $C_{16,3}$ for all mobiles in UE_{hs}

2.2.3   Power emulator and feasibility control

Once the user partition and the channel codes allocation are determined, the power emulator will estimate the consumed transmission power for the current allocation scheme. The transmission power is implicitly determined from the user distribution and the allocated channel bandwidth.

2.2.3.1   FACH Transmission Power

Unlike the traditional FACH power allocation which sets the FACH power such that it covers the whole cell, our model supports the dynamic power setting (DPS) for FACH. This idea is initially proposed in [78]: the power of FACH could be adjusted to achieve the worst users in its target transmission coverage, hence the FACH/S-CCPCH power can be efficiently save. With this idea, the S-CCPCH power requirements are different for different transmission coverages. DPS technology

requests to enable channel quality measurement report from UE in CELL_FACH state, e.g. the received Common Pilot Channel (CPICH) Ec/No signal quality on UE side [78]. Based on the periodic report from UE, typically every 600 ms for CELL_FACH, the FACH power will be different depending on the various user distributions in $UE_{fach}(f_{s,j})$.

Table 2.5: Estimated S-CCPCH power vs. cell coverage (PedestrianB 3km/h) [5]

Estimated cell coverage, %	Power for 64 kbps service, W
30	0.64
40	0.8
50	1.0
60	1.2
70	1.8
80	2.8
90	4
100	10

Table 2.5 shows the simulation results in the 3GPP technical report [12]. In this report, the authors study the S-CCPCH power level to capture the simulation results in MBMS standardization [1] in which the simulated FACH power are conducted by SIEMENS private simulator presented in [32, 31]. As the dynamic power setting for FACH is not standardized in MBMS, the simulation results in [12] do not involve all geometry coverages of the cell for various service bandwidth. For example, for a 64 kbps service, 80ms TTI and 1% BLER target, the FACH transmit power is simulated with a cell coverage from 30% to 100% (see Table 2.5). While for 16, 32, 128 and 256 kbps services, the results are only for the cell coverage from 90% to 95%.

Besides, in our simulation setup, the cell layout is 3-cell sectorization with site-to-site distance of 3 km, i.e. the distance between two Node B. Where the Node B is located in the corner of the three cells, thus the cell radius is 1 km and the farthest user distance from Node B could be 2 km. While in [1] the site-to-site distance is 1 km, the Node B sits in the cell center and the farthest distance between the UE and the Node B is 0.5 km. Regarding the parameter setting difference and lacking of various geometry coverages, the FACH estimated power is simulated by OPNET Modeller 15.0.A [8], with the parameter setup in Table 2.6.

The obtained FACH transmission power is shown in Figure 2.4. The simulation parameter settings are Pedestrian B 3 km/h with geometry factor = - 3dB, Block Error Rate (BLER) = 1% and Transmission Time Interval (TTI) = 80ms. The
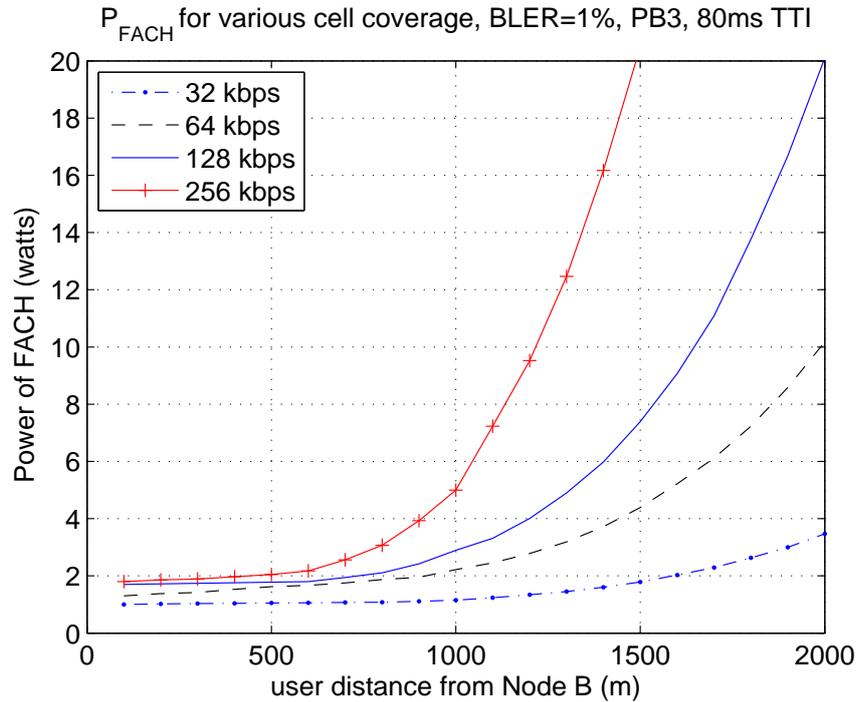
Figure 2.4: FACH transmission power obtained by OPNET 15.0.A

BLER is the error rate of the transport data passed by the physical layer to the MAC layer for a given transport channel. The TTI is the interval of time over which a transport data is transmitted on the radio link.

FACH is the common channel that many users can access at the same time with the same resource (one channel code and one portion of power). Although FACH is the favorite choice for MBMS transmission, other crucial factors such as the number of users within the multicast group, the user distances from the serving Node B, the service bandwidth, the throughput requirement and the E_b/N_o target affect the choice of the most efficient transport channel.

2.2.3.2 DCH Transmission Power

On the PTP downlink transmissions, where the multiple DCHs are used, fast power control is used to maintain the quality of each link (set target $E_b/N_o = 3$ dB in our simulation [53]) and to provide a reliable connection for the receiver to obtain data with acceptable error rate. Besides, DCH power control supports to maintain required link quality with just enough power thus to ensure the minimum interference.

Equation 2.4 (see section 4.4.4 in [70]) is used to calculate the base station transmission power for DCH covering n users in a cell. Based on this equation, the DCH power is variable depending on the number of users, the geographic distributions of users, the service bandwidth and the target signal quality for each user. In Equation 2.4, the length unit is meter and the power unit is W.

$$P_{DCH} = \frac{P_{ccc} + \sum_{i=1}^{n} L_i \cdot \frac{\frac{P_{noise}+I_i}{W}}{(\frac{E_b}{N_o})R_i} + \rho}{1 - \sum_{i=1}^{n} \frac{\rho}{(\frac{E_b}{N_o})R_i} + \rho} \tag{2.4}$$

In Equation 2.4, P_{ccc} is the power devoted to common control channels, P_{noise} is the background noise, W is the bandwidth in UMTS environment, E_b/N_o is the target experienced signal quality of user. ρ is the orthogonality factor since orthogonal codes are used in the downlink but some orthogonality is lost due to multi-path. ρ is between 0 and 1, 0 represents perfect orthogonality. R_i is the ith user transmit rate; in this model it is the allocated channel bit rate for this user. L_i is the user path loss from its attached Node B; it is calculated with Okumura Hata's model [65]: $L_i = 128.1 + 37.6\text{LOG}(d_i)$, where d_i is the user distance to the attached Node B. I_i is the intercell interference observed by the ith user given by a function of the transmitted power from neighboring stations P_j and path loss $L_{i,j}$ from the jth station to the ith user [70].

$$I_i = \sum_{j=1}^{N} \frac{P_j}{L_{i,j}} \tag{2.5}$$

where P_j is the transmission power from neighboring station j and N is the number of neighboring stations. $L_{i,j}$ is the path loss from the jth station to the ith user.

An example of P_{DCH} calculation.   Given an example in one cell C_1, two terminals t_1 and t_2 are receiving one flow through two DCH at 64 kbps. The user distances to the Node B in C_1 are $d_1 = 0.614$ km and $d_2 = 0.949$ km. Firstly the path loss from the Node B in C_1 to terminals t_1 and t_2 are:

$$L_{1,0} = 128.1 + 37.6\text{LOG}(0.614) = 120.14 \tag{2.6}$$
$$L_{2,0} = 128.1 + 37.6\text{LOG}(0.949) = 127.245 \tag{2.7}$$

There are 14 neighboring Node B around the C_1 with transmitted power $P = 5W$. The coordinates of these Node B are: $(2600, 4500)$, $(0, 3000)$, $(5200, 3000)$, $(-2600, 1500)$, $(2600, -4500)$, $(0, -3000)$, $(5200, -3000)$, $(2600, 1500)$, $(-2600, -4500)$, $(5200, 0)$, $(-3000, -5200)$, $(-5200, 0)$, $(-2600, -1500)$ and $(2600, -1500)$. From the link quality report received at these Node B sides, the terminal distances to Node

B are obtained, then the neighborhood cell interference to each terminal could be calculated:

$$i \in \{1,2\}, I_i = \sum_{j=1}^{14} \frac{5}{L_{i,j}} \qquad (2.8)$$

where $L_{1,j}$ is the path loss from jth neighbor cell to terminal t_1. According to the parameter setting in Table 2.6, the value of P_{DCH} for transmission to t_1 and t_2 is calculated:

$$P_{DCH} = \frac{1 + \sum\limits_{i=1}^{2} L_i \cdot \frac{10^{-10}+I_i}{\frac{5\times 10^6}{(10^{3/10})64000}+0.5}}{1 - \sum\limits_{i=1}^{2} \frac{0.5}{(10^{3/10})64000} + 0.5} = 0.0787W \qquad (2.9)$$

2.2.3.3 HS-DSCH Transmission Power

HS-DSCH is a rate controlled rather than a power controlled transport channel [49]. In HSDPA, the fast power control and variable spreading factor principles (characterizing Release 99 channels) are replaced by the link adaptation functionality, including techniques such as dynamic Adaptive Modulation and Coding (AMC), multicode operation, fast scheduling, Hybrid ARQ (HARQ) and short TTI of 2ms.

There are two modes to allocate the transmission power for HS-DSCH. In the first mode, a fixed amount of power is explicitly allocated per cell, and might be updated any time later. In the second mode, the base station is allowed to use any unused power remaining after serving other power controlled channels (such those for voice and non HSDPA UE) for HS-DSCH transmission [49]. In the first mode, setting the power too high might result in too much interference in cell, while setting it too low could not achieve the highest data rate. Therefore, in our work, like in most of the real networks, the second mode is considered to provide only the required amount of power so as to satisfy the MBMS multicast users. Equation 2.10 is used to calculate the minimum required HS-DSCH transmission power to achieve a minimum required HS-DSCH data rate at the edge of channel coverage [49].

$$P_{HS-DSCH} = SINR_{hs} \times [\rho - G^{-1}]\frac{P_{intra}}{SF_{16}} \qquad (2.10)$$

Where ρ is the orthogonality factor. SF_{16} is the spreading factor equal to 16, as the allocation of HS-DSCH is in units of channelization codes of length 16. P_{intra} is the total power transmitted in the serving cell. G is the geometry factor or the Carrier-to-Noise-and-Interference Ratio ($C/(N + I)$) defined according to the

following equation [53]:

$$G = \frac{P_{intra}}{I_{inter} + P_{noise}} \tag{2.11}$$

Where I_{inter} is the power spectral density of a band limited white noise source (simulating the interference from neighboring cells) and P_{noise} is the effective power spectral density of the Additive Gaussian White Noise (AGWN).

G is related with the user position. For a user at the cell edge, the interfering power from the neighboring cells is higher than the transmitted power from the serving cell, thus G is expressed by a lower value. In the macrocell (hexagonal layout with 1000 m base station spacing), the user within 80% coverage experiences a geometry factor of -2.5dB or better, that is I_{inter} is the double of P_{intra}, within 95% the geometry factor is at least -5.2dB [32], that is I_{inter} is the triple of P_{intra}.

$SINR_{hs}$ in Equation 2.10 is the target user experienced Signal-to-Interference-plus-Noise Ratio (SINR) for any t_k assigned to UE_{hs}. Based on the analytic formulation driven by link-level simulation results in [11], the value of $SINR_{hs}$ could be obtained according to the Channel Quality Information (CQI) and target BLER (set 1% in this work). The CQI is obtained through the target HS-DSCH total bandwidth and mapping table of MAC-hs bit rates versus CQI (Appendix B) [13]. Then $P_{HS-DSCH}$ is calculated by applying $SINR_{hs}$ and G into Equation 2.10.

An example of $P_{HS-DSCH}$ calculation. Recall the example presented in section 2.2.3.2, we assume that the terminals t_1 and t_2 are receiving the flow 64 kbps through HS-DSCH. According to the total bandwidth of HS-DSCH (128 kbps), from the mapping table in [13], the target CQI value is 5. Equation 2.12 shows the theoretical transfer formulation.

$$BLER = (1 + 10^{2\frac{SINR_{hs}-1.03CQI+5.26}{\sqrt{3}-\text{LOG}(CQI)}})^{-\frac{1}{0.7}} \tag{2.12}$$

Then the practical formula used in enhanced simulator NS2 [11] is shown in Equation 2.13. The estimated $SINR_{hs}$ by Equation 2.13 is 2 dB with BLER fixed at 1%.

$$SINR_{hs}(dB) = 0.5(1.73205 - \text{LOG}(CQI))(\text{LOG}(BLER^{-1.43} - 1.03))$$
$$+ 1.03CQI - 5.36 \tag{2.13}$$

According to the simulation results in [32] and the higher distance of the two terminals (949 meters) in a 3 km diameter cell, the geometry factor is 2.5 dB, hence the transmission power of HS-DSCH for t_1 and t_2 is:

$$P_{HS-DSCH} = 10^{SINR_{hs}(dB)/10} \times [0.5 - \frac{1}{10^{G/10}}]\frac{20}{16} = 0.1288W \tag{2.14}$$

2.2.4 Solution evaluator

The optimization target of our model is first to guarantee the throughput request, then to minimize the transmission power while avoiding power saturation. A two-dimensional fitness or objective function is defined to reflect these aspects and are computed for any solution x. The first objective is to minimize the loss of throughput in one cell c:

$$\text{Minimize } Th(x) \tag{2.15a}$$

$$\text{with } Th(x) = \sum_{s_i \in S(c)} \sum_{f_j \in F(s_i)} \sum_{t_k \in M(f_{s_i,j})} \max\{-\Delta_{j,k}, 0\} \tag{2.15b}$$

Where $\Delta_{j,k}$ is the difference between the allocated channel bandwidth (determined by its OVSF code(s) [6]) and the required flow bandwidth. For example, the user t_k requires $f_{s,j}$ (64 kbps) and receive 32 kbps through a DCH channel (SF = 64), then $-\Delta_{j,k}$ is: $-(32 - 64) = 32$ kbps.

As shown in Equation 2.16, the second optimization objective is to minimize the power consumption on one cell c.

$$\text{Minimize } Po(x) \tag{2.16a}$$

$$\text{with } Po(x) = \sum_{s_i \in S(c)} \sum_{f_j \in F(s_i)} \sum_{ch \in \{fach,dch,hs\}} P(f_j, ch) \tag{2.16b}$$

$$Po(x) \leq P_{mbms\_budget}(c) \tag{2.16c}$$

Equation 2.16c enforces the total power consumption on one cell c to simultaneous MBMS services to be lower than the maximum power budget.

With the two-dimensional fitness value, the comparison of a new solution x' and current solution x is conducted in lexicographic order:

Algorithm 1 Lexicographic evaluation.

Require: $Th(x), Po(x); Th(x'); Po(x')$
Ensure: x' is better or worse than x
  if $Po(x') \leq P_{mbms\_budget}(c)$ **then**
    if $Th(x') < Th(x)$ **or** $(Th(x') = Th(x)$ **and** $Po(x') < Po(x))$ **then**
      x' is better than x
    end if
  end if

2.2.5 Model complexity

Let's assume that the service s_i has $N_f(s_i)$ flow and $N_t(s_i)$ users. When the user t is receiving the flows F_t, $F_t \subseteq F(s_i)$, the number of possible channel assignment of user t is:

$$
\begin{array}{ll}
4 & \text{for } F_t = \{f_{s_i,0}\} \text{ or } F_t = \{f_{s_i,1}\} \\
3 \times 4 & \text{for } F_t = \{f_{s_i,1}, f_{s_i,2}\} \\
3 \times 3 \ldots 4 & \text{for } F_t = \{f_{s_i,1}, f_{s_i,2} \ldots f_{s_i,N_f(s_i)}\}
\end{array}
\tag{2.17}
$$

Hence, the number of allocation schemes for one user receiving service s_i is:

$$
4 + \sum_{i=0}^{N_f(s_i)-1} 4 \cdot 3^i = 2\left(3^{N_f(s_i)} + 1\right), s_i \in S(c)
\tag{2.18}
$$

The total number of possible channel assignment for all users and all services is:

$$
\prod_{i=1}^{N_s} \prod_{1}^{N_t(s_i)} 2\left(3^{N_f(s_i)} + 1\right)
\tag{2.19}
$$

In Equation 2.19, N_s is the number of services, $N_t(s_i)$ is the number of users receiving the s_i, $N_f(s_i)$ is the number of flows in s_i. The number of candidate solutions exponentially increases with the number of users, flows, and services. By looking at this, the exact algorithm by exhaustive search is not practical to solve this problem, and metaheuristic approach is a reasonable choice for solving such combinatorial optimization problem. In chapter 3, the MBMS RRM problem will be formally proved as NP-Hard problem.

2.2.6 Model synthesis

From subsection 2.2.1 to subsection 2.2.4, a mathematical model for the flexibility radio resource management for MBMS multicast service is built. Briefly, this model can be summarized as follows.

2.2.6.1 Input variables

The input variables of the proposed model are:

1. UE information

   (a) $T(c) = \{t_1, \ldots, t_k\}$. Set of mobile terminals in cell c. The terminal indexes are ordered from the closest to the farthest distance from the Node B at any time slot (after each TTI).

(b) $D(c) = \{d_1, \ldots, d_k\}$, $t_k \in T(c)$. Set of instantaneous distances from the Node B to the terminal, this value can be obtained through the channel quality measurement report from UE side [78].

2. Service information

   (a) $S(c) = \{s_1, ..., s_{N_s}\}$. Set of services to be transmitted to multicast groups located in cell c. The total number of service is N_s.

   (b) $F(s_i)$. The flow set of service s_i and its bandwidth, $s_i \in S(c)$. $F(s_i) = \{f_{s_i,0}\}$ if s_i is a single layer (SL) transmission scheme service. Or $F(s_i) = \{f_{s_i,1}, f_{s_i,2}, [f_{s_i,3}]\}$ if s_i is a multilayer (ML) transmission scheme service, where $f_{s_i,1}$ is the basic flow and $f_{s_i,j}$ with $(j > 1)$ are advanced flows. We assume that each service has $N_f(s_i)$ flows.

   (c) $M(s_i) \subseteq T(c)$. Multicast group of service s_i. One group includes all the terminals $\{t_1, t_2, \ldots, t_{k'}\}$ requesting the same service content with $k' \leq k$. The number of terminals in each group is $N_t(s_i) = N_t(f_{s_i,j})$, $\forall f_{s_i,j} \in F(s_i)$.

2.2.6.2   Decision variables

To support the combinational allocation modes for each flow (see Table 2.1), the decision variables include the UE partition and the channel code assignment. The UE partition indicates the terminals allocated in four disjointed channel user sets.

  – $UE_{fach}(f_{s,j})$, users served through a FACH.
  – $UE_{dch}(f_{s,j})$, users served through DCHs.
  – $UE_{hs}(f_{s,j})$, users sharing a HS-DSCH.
  – $UE_{noch}(f_{s,j})$, non-served users.

Then the channel code assignment is performed by the channel code allocator which assigns the available channel code to a given channel.

2.2.6.3   Decision constraints

The UE partition for flow $f_{s,j}$ should be in accord with the channel characteristics, as shown in following decision constraints:

$$\forall s, j \in \{0, 1\}, R(f_{s,j}) = M(s) \tag{2.20}$$

$$R(f_{s,j}) = UE_{fach}(f_{s,j}) \cup UE_{dch}(f_{s,j}) \cup UE_{hs}(f_{s,j}) \tag{2.21}$$

$$\forall s, j \geq 2, R(f_{s,j}) \subseteq R(f_{s,j-1}) \tag{2.22}$$

$$\forall t_i \in UE_{fach}, d_i \leq d_{thr} \tag{2.23}$$

$$\forall t_j \in UE_{dch} \cup UE_{hs} \cup UE_{noch}, d_j > d_{thr} \tag{2.24}$$

$$ch_m, ch_n \in \{fach, dch, hs, noch\}, UE_{ch_m} \cap UE_{ch_n} = \phi \tag{2.25}$$

Where $R(f_{s,j})$ represents the users whom are receiving $f_{s,j}$. d_i is the distance of the user t_i from the base station, and d_{thr} is the distance threshold found during the optimization procedure.

2.2.6.4   Optimization objective

A two-dimensional optimization objective is defined in the proposed model. The first optimization objective is to minimize the throughput loss, which is the accumulated bandwidth difference between the allocated channel bandwidth and the required flow bandwidth, in one cell c.

$$\text{Minimize } Th(x) \tag{2.26a}$$

$$\text{with } Th(x) = \sum_{s_i \in S(c)} \sum_{f_j \in F(s_i)} \sum_{t_k \in M(f_{s_i,j})} \max\{-\Delta_{j,k}, 0\} \tag{2.26b}$$

The second optimization objective is to minimize the consuming power in cell c.

$$\text{Minimize } Po(x) \tag{2.27a}$$

$$\text{with } Po(x) = \sum_{s_i \in S(c)} \sum_{f_j \in F(s_i)} \sum_{ch \in \{fach,dch,hs\}} P(f_j, ch) \tag{2.27b}$$

$$Po(x) \leq P_{mbms\_budget}(c) \tag{2.27c}$$

2.3   Simulation setup

The proposed F2R2M is implemented as a simulator with the MBMS RRM core model, together with its simulation environment. This section describes the simulation parameter settings and the designed experiment scenarios which are used to evaluate the proposed model.

2.3.1 Simulation parameters

Figure 2.5 depicts the macrocell layout which consists of 19 hexagonal cells. The red rings indicate the location of the Node B. The simulation focus on MBMS transmission in cell $C1$, assuming that only multicast services are transmitted in this cell.
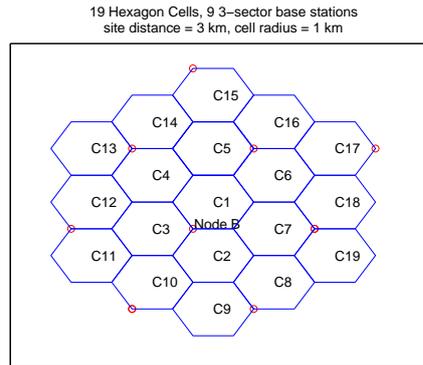
Figure 2.5: Cellular layout in simulation setup

Table 2.6 lists the simulation parameters. The maximum power for MBMS in one cell is 19 W, which is the total transmission power (20 W) minus the power for common channel (1 W); it is the conventional power setup for MBMS study [24, 28, 12]. In macrocell environment, the Okumura Hata path loss model is applied to calculate the path loss L_i of user t_i, with a carrier frequency of 5 MHz and a base station antenna hight of 15 meters, that is: $L_i = 128.1 + 37.6\text{LOG}(d_i)$, where d_i represents the distance from the Node B to the UE in km [2].

2.3.2 Simulation scenarios

As shown in Table 2.7, a number of scenarios are designed with different service requests and user distributions. These scenarios will be used in the following chapters for model analysis and algorithm evaluation.

The scenario name indicates the number of service s, the total number of users u and the service transmission scheme, where S stands for scalable transmission scheme and N stands for non-scalable transmission scheme. For example, the scenario *2s20uSN* has two multicast services s_1 and s_2, both having 10 users. s_1 is a scalable transmission scheme service. It consists of three flows: the basic flow $f_{s_1,1}$ with the bandwidth 32 kbps, and the two advanced flows $f_{s_1,2}$ and $f_{s_1,3}$ with 32 kbps and 64 kbps, respectively. s_2 is a non-scalable transmission scheme service, i.e.

Table 2.6: System level parameter setting

Parameters	Value
Cellular layout	19 hexagonal cells, 3-sector sites
Number of neighboring cells	18
Orthogonally factor (ρ)	0.5
UMTS bandwidth	5 MHz
Site to site distance	3 km
Cell radius	1 km
Base station transmit power	43 dBm (20 Watts[49])
Background noise	-100 dBm
Power of neighbor cell	37 dBm (5 Watts)
Propagation models	Okumura Hata
IUT path loss environment	Pedestrian B, 3 km/h
Common channel power	30 dBm (1 Watt)
Target E_b/N_o	3 dB
Block Error Ratio (BLER)	1 %

one flow $f_{s_2,0}$ with the bandwidth 128 kbps. The total bandwidth requirement of *2s20uSN* is 2560 kbps, obtained by $10 \times (32 + 32 + 64) + 10 \times 128$.

2.3.3   Power simulation of three transport channels

In this section, to validate the power emulator in our model, simulation is conducted to synthesize the power consumption of three transport channels, for various service bandwidths. For each service bandwidth, three types of user distributions (depending on their distance to the Node B) are defined. They are named as "close", "median" and "far", in which, the users are randomly generated in cell $C1$, located with random distance in given ranges: $]0, 700]$, $]700, 1400]$ and $]1400, 2000]$ meters from the Node B.

Figure 2.6 shows an example of users located in $C1$. There are 12 users, marked as black points, close to the Node B, the blue plus signs represent 12 users in median locations, and the 12 red asterisks are the users which are far from the Node B, some of them are in the cell edge.

The power simulation is conducted as follows. Assume that all users belong to

Table 2.7: Experimental scenarios

Scenario name	Flows of service	User number	Total traffic requirement (kbps)
1s30uS	s_1 (f_1: 64 kbps, f_2: 64 kbps)	30	3840
1s60uS	s_1 (f_1: 64 kbps, f_2: 64 kbps)	60	7680
1s150uS	s_1 (f_1: 64 kbps, f_2: 64 kbps)	150	19200
2s20uSN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_0: 128 kbps)	10 10	2560
2s20uSS	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_1: 64 kbps, f_2: 64 kbps)	10 10	2560
2s50uSN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_0: 128 kbps)	30 20	6400
2s50uSS	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_1: 64 kbps, f_2: 64 kbps)	30 20	6400
3s30uSNN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_0: 128 kbps) s_3 (f_0: 64 kbps)	10 10 10	3200
3s30uSSN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_1: 64 kbps, f_2: 64 kbps) s_3 (f_0: 64 kbps)	10 10 10	3200
3s50uSNN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_1: 128 kbps) s_3 (f_0: 64 kbps)	10 20 10	4480
3s50uSSN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_1: 64 kbps, f_2: 64kbps) s_3 (f_0: 64 kbps)	10 20 10	4480
3s80uSNN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_0: 128 kbps) s_3 (f_0: 64 kbps)	30 30 20	8960
3s80uSSN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_1: 64 kbps, f_2: 64kbps) s_3 (f_0: 64 kbps)	30 30 20	8960
3s100uSNN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_0: 128 kbps) s_3 (f_0: 64 kbps)	30 40 30	10240
3s100uSSN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_1: 64 kbps, f_2: 64kbps) s_3 (f_0: 64 kbps)	30 40 30	10240
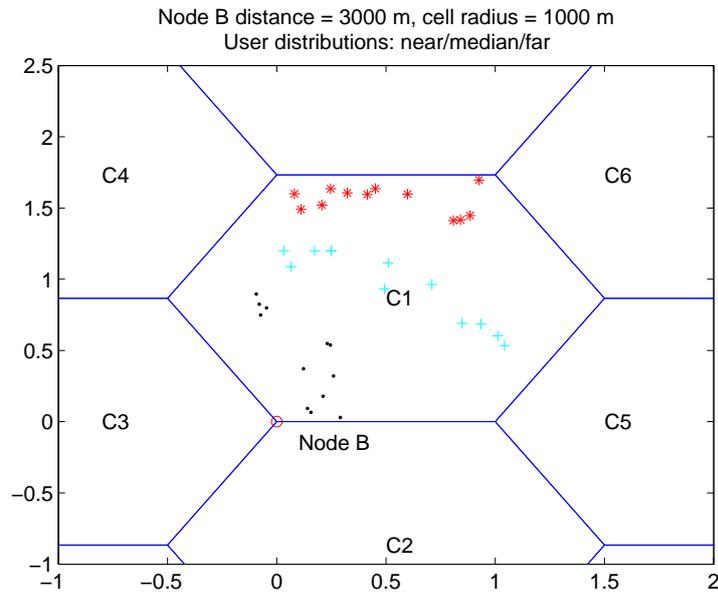
Figure 2.6: Three user distributions for power comparison

the same multicast group, for a given service bandwidth and a given user distribution, three transport channels (i.e. FACH, HS-DSCH and DCH) are separately applied in three different runs, hence there is no channel co-existing for the same service. The service bandwidth could be 32, 64, 128 and 256 kbps, respectively. The user distribution could be "close", "median" or "far"; each distribution has 1 to 18 users. There is no optimization in these evaluations; we run the different scenarios separately to calculate the impact of the different parameters on MBMS performance: user location, service bandwidth and transport channels.

Figure 2.7 shows the transmission power of the three transport channels for sending 32 kbps and 64 kbps services to users located close, median and far from the Node B. The x-axis shows the number of users from 1 to 18, while the y-axis plots the power level of three channels separately. When the users are close to the Node B, for 32 kbps service, DCH consumes the lowest power, while FACH power is fixed around 1 W; it is sufficient to cover the farthest user in the multicast group. When users are in the middle of the cell, HS-DSCH offers a better power adaptation than the other two channels. When the users are far from Node B, the transmission power of these channels fluctuates. For 32 kbps service, HS-DSCH consumes the lowest power when the number of users is smaller than 11 (Figure 2.7(e)), while FACH consumes the lowest power for user number larger than 11. For 64 kbps service, HS-DSCH always consumes lower power than FACH and DCH whatever the number of users is (Figure 2.7(f)). In these six runs, all required bandwidths are satisfied.
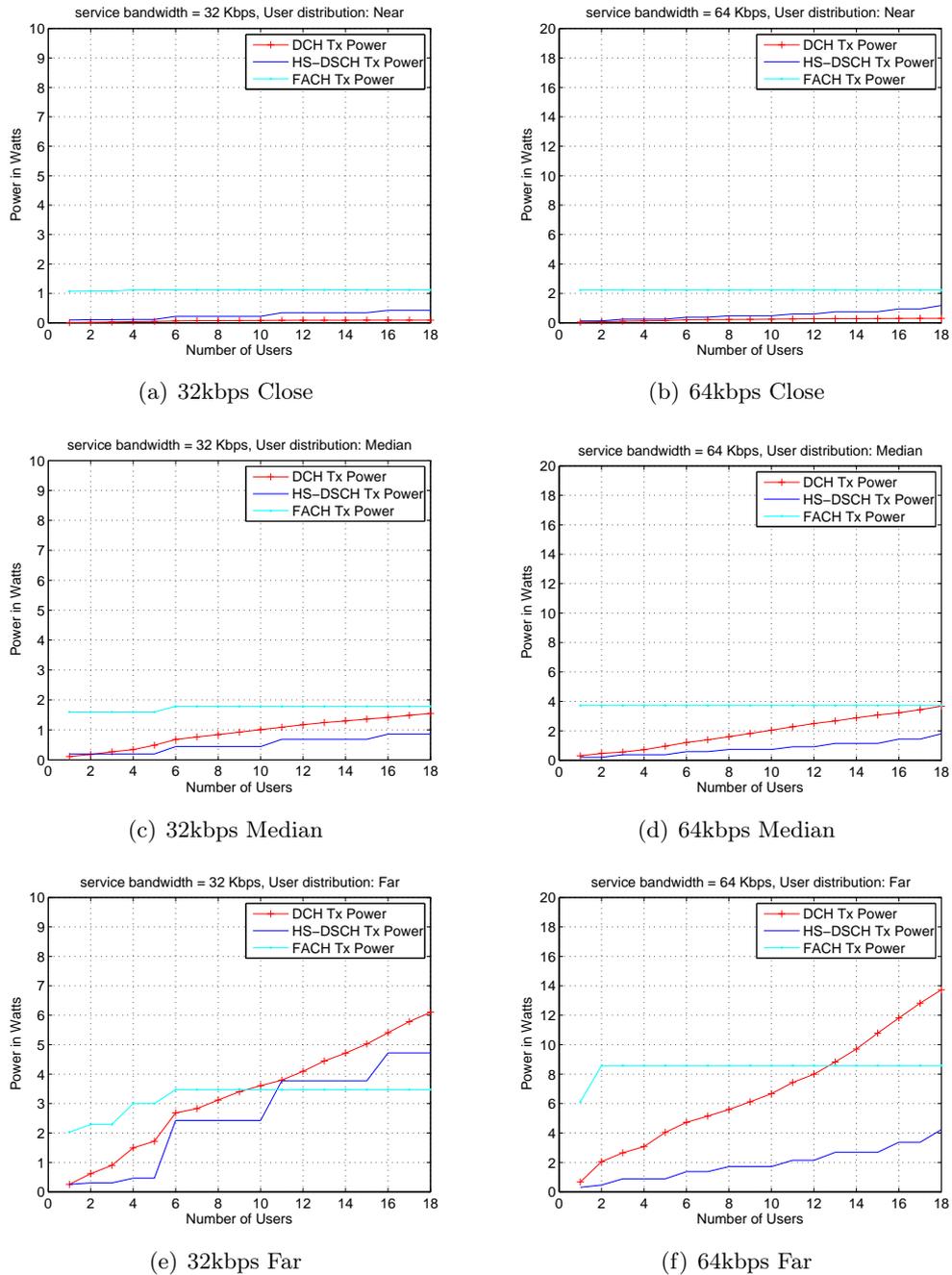
Figure 2.7: Separate power consumption of the three channels for 32 kbps and 64 kbps services
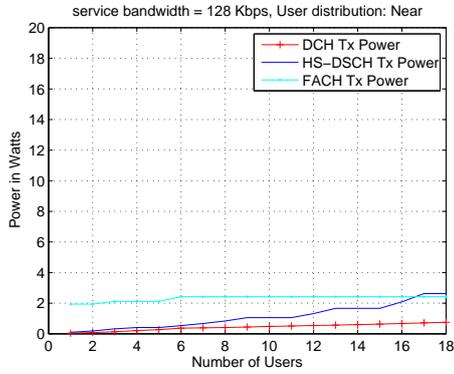
(a) 128kbps Close
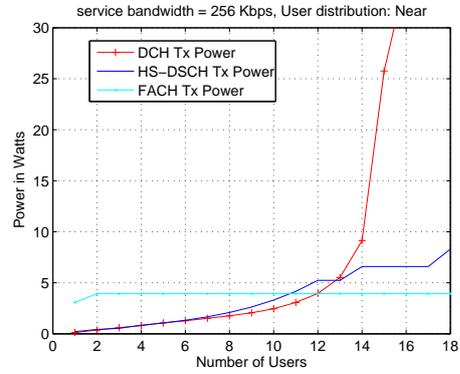
(b) 256kbps Close
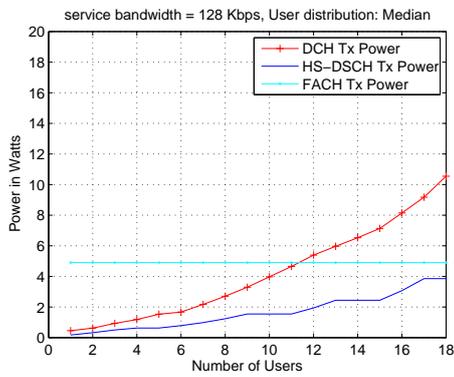
(c) 128kbps Median
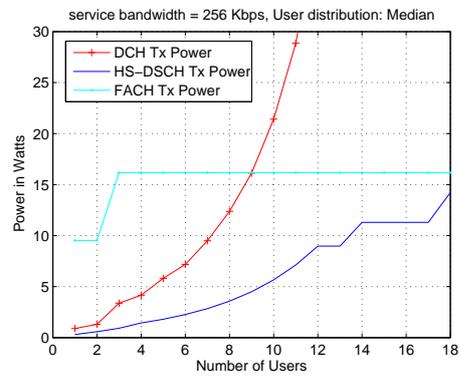
(d) 256kbps Median
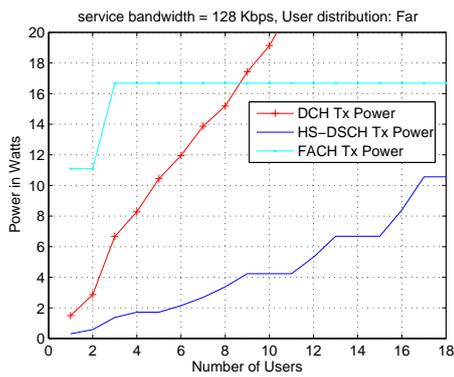
(e) 128kbps Far
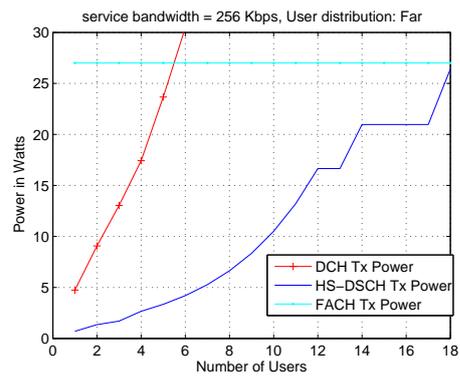
(f) 256kbps Far

Figure 2.8: Separate power consumption of three channels for 128 kbps and 256 kbps services

Figure 2.8 shows the power emulation of the three transport channels for the service 128 kbps and 256 kbps. It can be observed that the power consumption increase with the number of users and the user distance from the Node B and sometimes the total power consumption overpasses the maximum power set to 19 W (Figure 2.8(b), 2.8(d), 2.8(e) and 2.8(f)).

Then two conclusions can be obtained. First, none of these three channel gains absolute transmission efficiency in terms of power consumption when the users numbers and users positions are different. In this study, the user positions are randomly changed, hence the same user number and similar geographic scenario may still have different power levels. Therefore, the MBMS UE counting mechanism or any transmission decision algorithm based on the user numbers cannot be efficient whatever the input scenarios. Second, HS-DSCH can truly consume the lowest power for certain scenarios in particular when the terminal distance is not close to the Node B.

Besides, the advantages of scalable transmission can be observed. In Figure 2.8(f) at 256 kbps, when the user number is over 14, the three channels consume more than 19 W, hence achieve the power saturation within one cell. While in Figure 2.8(e), for one 128 kbps stream transmission to 14 users far from Node B, the lowest power is 6.4 W with HS-DSCH and is also acceptable for FACH. Therefore two 128 kbps streams will need 12.8 W, much less than the power requirement of one 256 kbps stream transmission with the same channel (22 W in Figure 2.8(f)). To fully take advantages of the transmission characteristics of the three transport channels, we propose to integrate the scalable transmission with the combined channel selection and transmission.

2.4 Synthesis

This chapter presents a model for the MBMS RRM problem. This model maps the radio resource establishment procedure for MBMS into three phases: i) collect the input variables; ii) determine the decision variables; and iii) establish the radio bearers according to the best found solution. The second phase plays the most important role for the system performance and is the focus of our model. In this phase, the MBMS RRM is modeled as an optimization problem which tries to find the optimum radio resource assignment for a given scenario. To solve this problem, an iterative search architecture is designed to explore new solutions. In each iteration, seven modules are invoked consecutively:

1. **UE partition search engine** generates a candidate solution x' by modifying a current solution x.

2. **Channel code allocator** associates available channelization code to each

assigned transport channel.

3. **Channel availability controller** will shift user(s) to UE_{noch} when there are no more available channel code for these users, thus the candidate solution is modified regarding the limited channel code resource.

4. **Power emulator** estimates the transmission power of the candidate solution after the channel code verification.

5. **Feasibility controller** will pass the candidate solution if its transmission power is below the cell power budget, otherwise this solution will be rejected as unfeasible solution.

6. **Solution evaluator** compares the candidate solution and the current solution in lexicographic order, it accepts the new solution if it is better than the current solution.

7. **Stop criteria** determines if the search procedure should stop, then output the best found solution.

To better explore the solution space, two additional flexibilities are proposed. First, instead of service based channel allocation, the flow based allocation is proposed, which supports the scalable streaming data. Second, this model supports the flexible transmission mode selection. For solution representation, a solution contains the UE partitioning of four different channel types, as well as the channel code allocation and the transmission power allocation. For solution evaluation, a two-dimensional fitness function is proposed, which reflects both the allocated bandwidth and the consumed power of the allocation solution. Therefore, the allocation solution is found regarding a two-dimensional objective function, while the other MBMS radio resource allocation approaches consider the MBMS RRM as a mono-objective optimization problem. For solution updating, two neighborhood operators are proposed: the single insert operator δ_{SI} and the multiple insert operator δ_{MI}. The current flow solution is transfered to a new flow solution by one application of the operator.

The proposed F2R2M is implemented as a simulator which reflects the MBMS RRM core model, together with its simulation environment which contains the OVSF code allocator and the power emulator. A number of scenarios are also designed, which will be used for algorithm performance evaluation in the next chapters. Besides, to verify the implemented power emulator, the channel code allocator and the model validity, the transmission power of the three transport channels are simulated for various service bandwidths. The simulation results show that the transmission power fluctuates with different users number and users distributions. Hence three observations were concluded: firstly, none of these three channel gains absolute transmission efficiency in terms of power consumption for arbitrary scenarios; secondly, HS-DSCH can consume the lowest power for certain scenarios, hence the co-existing of R5 channel and R99 channels could improve the efficiency of radio resource allocation; thirdly, the adopted scalable transmission technology can

reduce the power consumption for several scenarios, and therefore proves its value in our model.

The proposed F2R2M enables systematic analysis for MBMS RRM problem. Also it makes it possible that optimization algorithms (e.g. metaheuristic algorithms), can be used to solve this MBMS RRM problem. The detailed analysis and algorithm design will be presented in the following chapters.

Model Analysis

In this chapter, we conduct an in-depth analysis on the studied problem. To understand the problem complexity and the solution boundaries, the work shows that, by relaxing the OVSF code constraints, the MBMS RRM problem can be approximated to be a multiple-choice knapsack problem (MCKP). It gives two outcomes: first, the NP-hard proof for MBMS RRM is self-contained because MCKP is NP-hard. Second, solving MCKP will give theoretical solution bounds for MBMS RRM solver. To understand the structure of the solution space, the landscape analysis technique is conducted on F2R2M. In this analysis, the landscapes differentiated by two neighborhood operators, single insert operator and multiple insert operator, are generated. Then the characteristics of the studied problem are analyzed regarding the search space and the solution distribution. Simulations show that the studied problem is rugged in both search spaces. Simulations also show that multiple insert operator is better than the single insert operator for efficient search.

Contents

3.1   Knapsack problem approximation

In this section, the MBMS RRM problem is formulated as a variant of the knapsack problem (KP), which represents a set of integer combinatorial optimization problems known as NP-Hard [30, 60].

3.1.1   Knapsack problem and its variants

The knapsack problem and its variants appear in real-world applications in a wide variety of fields, such as cargo systems [47] and telecommunications [81]. Many different kinds of knapsack problems are found in the literature, including multi-dimensional [40], multi-objective [26], multiple-choice knapsack problems [74] and their combinations [64].

Classic 0-1 knapsack problem (KP).   The 0-1 knapsack problem is a classical problem in combinatorial optimization. Suppose a hiker needs to fill a knapsack for a trip. He has a set of items, each with a value (benefit) of v_i and a weight of w_i. It is common to assume that all values and weights are positive. The hiker needs to package selected items to his knapsack to maximize the overall value in the knapsack, while guaranteeing that the aggregated size of all selected items do not exceed the knapsack capacity constraint W. A binary decision variable x_i is introduced to model the decision process. If the i-th item is selected then $x_i = 1$ while $x_i = 0$ otherwise. The standard knapsack problem is given by Equation 3.1.

$$\text{Maximize } z = \sum_{i=1}^{n} v_i x_i \tag{3.1a}$$

$$\text{Subject to } \forall i, x_i \in \{0, 1\}, \sum_{i=1}^{n} w_i x_i \leq W \tag{3.1b}$$

The objective function in the standard knapsack is to maximize the sum of the values of the items in the knapsack, so that the sum of the weights must be less than the knapsack capacity.

Multi-dimensional knapsack problem (MKP).   A problem may have different criteria to measure the solution quality, and it is not possible to select a most important criterion or to combine the criteria into a single objective function. The goal of a Multi-dimensional Knapsack Problem (MKP), also called m-dimensional knapsack problem, is to maximize the profit of selected items. The selected items should be partitioned over a set of knapsacks contrary to the classical KP where only one knapsack is available. Each knapsack is characterized by a capacity representing

the maximal total weight it can contain. Given a set of n items and a set of m knapsacks ($m \leq n$), with $v_i = value$ of item i, $w_i = weight$ of item i, $W_j = capacity$ of knapsack j. The decision here is not only whether to select a single item but also in which knapsack it is packed. Therefore, MKP consists in selecting m disjointed subsets of items so that the total value of the selected items be maximal, and each subset can be assigned to a different knapsack whose capacity is no less than the total weight of items in the subset. Formally, this problem is modeled in Equation 3.2:

$$\text{Maximize } z = \sum_{j=1}^{m} \sum_{i=1}^{n} v_i x_{ij} \tag{3.2a}$$

$$\text{Subject to } \forall j \in \{1, \ldots, m\}, \sum_{i=1}^{n} w_i x_{ij} \leq W_j \tag{3.2b}$$

$$\forall i \in \{1, \ldots, n\}, \sum_{j=1}^{m} x_{ij} \leq 1 \tag{3.2c}$$

$$\forall j \in \{1, \ldots, m\}, i = \{1, \ldots, n\}, x_{ij} \in \{0, 1\} \tag{3.2d}$$

$$w_i \geq 0, v_i > 0, W_j \geq 0 \tag{3.2e}$$

A binary variable $x_{ij} = 1$ means that the item i is selected and packed into the knapsack j, otherwise $x_{ij} = 0$. The constraints of the problem ensure that each item is selected at most once in these knapsacks. The MKP is a well-studied and strongly NP-Hard combinatorial optimization problem occurring in many different application [29].

Multi-objective m-dimensional knapsack problem (MOKP). The multi-objective optimization problem is frequently encountered in practice. This problem can be formulated as in Equation 3.3.

$$\text{Maximize } \mathcal{Z} = \{\sum_{i=1}^{n} v_i^1 x_i, \ldots, \sum_{i=1}^{n} v_i^k x_i, \ldots\}, k = \{1, \ldots, t\} \tag{3.3a}$$

$$\text{Subject to } \forall j \in \{1, \ldots, m\}, \sum_{i=1}^{n} w_i^j x_{ij} \leq W^j \tag{3.3b}$$

$$\forall i, x_i \in \{0, 1\} \tag{3.3c}$$

$$v_i^k > 0, w_i^j \geq 0, W^j \geq 0 \tag{3.3d}$$

In the multi-objective m-dimensional knapsack problem, each item i has t different values ($\{v_i^1, \ldots, v_i^t\}$), which are corresponding to its benefit according to different criteria (e.g. price, reliability). Instead of trying to compute a single optimal solution with maximum value, this problem targets at computing a set of feasible solutions covering all possible trade-offs between the different profit values. In addition, each item i has m different weights, ($\{w_i^1, \ldots, w_i^m\}$) corresponding to the

resources consumption (e.g. weight, volume, ...), while the knapsack is defined by m different capacities ($\{W^1, \ldots, W^m\}$) representing the corresponding resources availability. Therefore, the knapsack capacity must be satisfied for each of the m resource categories.

Multiple-choice knapsack problem (MCKP). The Multiple-Choice Knapsack Problem is a variant of the 0-1 knapsack problem, also known as a NP-Hard problem [74]. Given a set of n items partitioned into k sets (classes) $K_1, \ldots, K_k$ that should be packed into a knapsack of capacity W. Each item i is associated with a value, v_i, representing its benefit and a weight w_i. The problem goal is to choose one and only one item from each class in such manner that the sum of the values of selected items is maximized and the knapsack capacity W is not exceeded. The MCKP is formulated in Equation 3.4.

$$\text{Maximize } z = \sum_{i=1}^{n} v_i x_i \tag{3.4a}$$

$$\text{Subject to } \sum_{i=1}^{n} w_i x_i \leq W \tag{3.4b}$$

$$\forall j \in \{1, \ldots, k\}, \sum_{i \in K_j} x_i = 1 \tag{3.4c}$$

$$\forall i = \{1, ..., n\}, x_i \in \{0, 1\} \tag{3.4d}$$

All parameters v_i, w_i and W are positive integers, the classes $K_1, \ldots, K_k$ are mutually disjoint, n_j is the size of the class K_j. The total number of items is $n = \sum_{j=1}^{k} n_j$.

3.1.2   Approximating MBMS RRM as a knapsack problem

For MBMS RRM, the main objective is to maximize the accumulation of allocated bandwidth over all users and all flows, then to minimize the allocated transmission power. And the total transmission power should be less that a predefined value P_{mbms_budget}.

In F2R2M, when there is no channel code satisfying the required bandwidth for a given set of users, a channelization code offering lower bandwidth will be selected. In this way, the allocated bandwidth will be less than the required throughput, and is named as bandwidth fraction. Besides, the solution of a given flow, i.e. the allocation of the OVSF code, should be considered regarding the OVSF code occupation by the other flows.

If the studied MBMS RRM problem is approximated as KP, the channel code orthogonality can not be simultaneously considered in the candidate solution se-

lection for flows, because the item selection for one flow (i.e. allocation schemes) is independent to the item selection for the other flows. Moreover, the bandwidth fraction will be meaningless without considering the OVSF code occupation and the code orthogonality.

Therefore, to map the studied problem into a KP, it is assumed that there is no channel code limitation. Once a user is selected, the required flow bandwidth is assumed fully available. The power limitation remains as the only reason for not serving a given user for a given flow. With the OVSF code constraint in mind, the primary KP based formulation is proposed in Equation 3.5 as a MCKP.

$$\text{Maximize} \sum_{j=1}^{N_f} \sum_{i=1}^{n_j} v_{ij} x_{ij} \tag{3.5a}$$

$$\text{Subject to} \sum_{j=1}^{N_f} \sum_{i=1}^{n_j} w_{ij} x_{ij} \leq W \tag{3.5b}$$

$$\forall j \in \{1, \ldots, N_f\}, \sum_{i=1}^{n_j} x_{ij} = 1 \tag{3.5c}$$

$$\forall j \in \{1, \ldots, N_f\}, \forall i \in \{1, \ldots, n_j\}, x_{ij} \in \{0,1\} \tag{3.5d}$$

Where each flow f_j is associated with n_j items, corresponding to a class K_j. The decision variable $x_{ij} = 1$ means that the ith item is selected as the allocation scheme for the flow f_j. In Equation 3.5a, v_{ij} is the allocated (required) channel bandwidth of the item i concerning the flow f_j (while in F2R2M, the bandwidth of allocation solution is determined by the spreading factor of allocated channel code). In Equation 3.5b, the weight of the i-th item in the j-th class, w_{ij}, is the power consumption of the i-th item for flow f_j. Besides, the total power consumption should be less than W, which is the value of P_{mbms_budget}. x_{ij} is a binary decision variable, $x_{ij} = 1$ when the i-th item in the j-th class is packed into the knapsack, otherwise $x_{ij} = 0$. The objective function in Equation 3.5 is to maximize the sum of the values selected, such that the sum of the weights must be less than W.

Equation 3.5 gives the hints that, by relaxing the OVSF code constraints, the studied problem may possibly be formulated as a MCKP, where each class is a set of candidate items for a given flow, each item represents a possible partition of its users among DCH served users, HS-DCH served users, FACH served users and not served users. The optimization target is to choose one and only one item from each class (Equation 3.5c), to maximize the allocated bandwidth (Equation 3.5a) and satisfy the power limitation (Equation 3.5b). However, more formulation details need to be specified:

– The representation of the knapsack: the knapsack could represent a transport

channel, a type of transmission mode or a whole cell.

– The item representation: an item could represent an allocation scheme for the services, or for the flows, or for a given user.

– The calculations of v_{ij} and w_{ij} which depend on the item structure.

To address these issues, three candidate propositions are discussed later.

3.1.2.1 Proposition 1: multi-dimensional MCKP

In Equation 3.5, in a given class associated with a given flow, an item could represent a user for the flow reception. In order to distinguish which channel serves the user (i.e. which channel set the user belongs to), instead of one knapsack with capacity of P_{mbms_budget} (Equation 3.5b), three knapsacks are defined, each represents a transport channel. Therefore, in this proposition, the flexible MBMS RRM problem is proposed to be formulated as the Multi-dimensional Multiple-Choice Knapsack Problem (MMKP).

As defined in the multi-dimensional KP, each knapsack is characterized by a capacity representing the maximal weight it can contain. Each item has m types of weight, and each weight measures the solution quality knowing that each weight is a constant value in the corresponding knapsack.

Based on the problem characteristic, in this proposition, the knapsack capacity could represent either the maximal bandwidth for each channel, or the maximal power for each channel.

1. MMKP with three knapsacks. Each knapsack represents one of the three transport channels: FACH, DCH and HS-DSCH. The knapsack capacity is the power limitation for each of these channels.

2. MMKP with four knapsacks. One knapsack represents the total power limitation in the whole cell. The capacity of the other three knapsacks is the maximal bandwidth offered by the channel category.

In the first assumption, the weights (power) of item for different knapsacks (channel) are different. Equation 3.6 shows the calculation of the total weight of the selected items packed in each knapsack (channel), which is the transmission power

for all flows served by the channel ch, $ch \in \{fach, dch, hs\}$.

$$w_{dch} = \sum_{i=1}^{N_s} \sum_{j=1}^{N_f(s_i)} w_{dch,j} = \sum_{i=1}^{N_s} \sum_{j=1}^{N_f(s_i)} \sum_{k=1}^{N_t(f_{s_i,j})} P_{dch}(t_k, bw_{f_j}) \qquad (3.6a)$$

$$w_{fach} = \sum_{i=1}^{N_s} \sum_{j=1}^{N_f(s_i)} w_{fach,j} = \sum_{i=1}^{N_s} \sum_{j=1}^{N_f(s_i)} P_{fach}(t_{d_{max}}, bw_{f_j}) \qquad (3.6b)$$

$$w_{hs} = \sum_{i=1}^{N_s} \sum_{j=1}^{N_f(s_i)} w_{hs,j} = \sum_{i=1}^{N_s} \sum_{j=1}^{N_f(s_i)} P_{hs}(t_{d_{max}}, N_t(f_{s_i,j}) \times bw_{f_j}) \qquad (3.6c)$$

N_f denotes the total number of flows in the cell. Equation 3.6a shows the accumulated weight of items in the knapsack $ch = dch$ (served by DCH). $w_{dch,j}$ is the DCH power for the flow f_j, $P_{dch}(t_k, bw_{f_j})$ is the DCH power transmission for f_j to user t_k, which is a constant value. Equation 3.6b is the accumulated weight in the knapsack $ch = fach$, where $w_{fach,j}$ is the weight (power consumption) of the users packed in the knapsack $ch = fach$ (served by a FACH). This value depends on the farthest user within the FACH user set and the required flow bandwidth. However, for the other users, which are closer to the Node B in UE_{fach}, their weight will be 0, because P_{fach} only depends on the farthest user in this transmission coverage. Equation 3.6c represents the total weight of the knapsack $ch = hs$, which (i.e. HS-DSCH power) depends on the farthest user (determining the SINR) in its user set. Moreover, HS-DSCH is a shared channel, its transmission power depends on the allocated channel bandwidth, which is the sum of the bandwidth received by the users through HS-DSCH: $bw_{f_j} \times$ Number of users in UE_{hs}. Therefore, the weight of the kth item (a user t_k), packed in the knapsack $ch = hs$ is not a constant value, because that $w_{hs,j}$ could not be pre-calculated without knowing the users served by HS-DSCH. Therefore, the first assumption is not in accordance with the MMKP definition, in which the weights of item for each knapsack should be a constant value.

For the second assumption, the knapsack capacities represent the maximum bandwidth capacity of each channel in one cell. According to the channel codes structure, when the OVSF codes are all used for DCH transmission, the accumulated DCH downlink bandwidth is up to 2 Mbps [53]. When the OVSF codes are all used for HS-DSCH transmission, HSPDA offers downlink data speeds up to 8-10 Mbps [54]. However, this bandwidth limitation considers that each channel occupies all OVSF codes. For one FACH, its theoretical maximum bandwidth is 2 Mbps [53], while the accumulated FACH bandwidth depends on the number of users covered by FACH, hence there is no bandwidth limitation for FACH. Therefore, the bandwidth limitations of the channels are not independent from each other, and the FACH knapsack has no bandwidth limitation, hence no maximal capacity. The second assumption is not suitable for the studied problem.

Regarding above analysis, the multi-dimensional MCKP based formulation could

not be accepted.

3.1.2.2   Proposition 2: parallel MCKP

The second proposition defines a single knapsack instead of multiple knapsacks, but a set of classes are constructed associated with each flow. In these classes, each item describes how the transmission modes are allocated for users receiving a given flow: users served by FACH, users served by DCH, users served by HS-DSCH or not served users. Therefore, the second proposition formulates the problem as MCKP. More precisely, each item associated with a given flow, represents a set of users or one user:

1. A set of user served by FACH receiving the same flow or not served.

2. One user served by DCH, HS-DSCH or not served.

The first case leads to one item which includes the users whose distance from the base station do not exceeds the threshold d_{thr}. In the second case, an item refers to one user which is allocated to DCH, HS-DSCH, or not served. Items indicating the same set of terminals for the same flow reception constitute one class. With this proposition, the value/profit of each item could be calculated in advance. Equation 3.7 describes a simple scenario to illustrate the second proposition.

A simple scenario:

$$S(c) = \{s_1\} \tag{3.7a}$$
$$F(s_1) = \{f_1, f_2\} \tag{3.7b}$$
$$M(s_1) = \{t_1, t_2, t_3\} \tag{3.7c}$$

Where $M(s_1)$ is a set of terminals belongs to the multicast group of service s_1, which has $N_f = 2$ flows. The requested flow bandwidth are both 64 kbps. In the multicast group M_{s_1} of service s_1, three users request for the same service content. Users are ordered according to their distance to the base station, t_1 is the closest user, while t_3 is the farthest.

Based on the formulation in the second proposition, let's assume that t_1 and t_2 are served by FACH to receive f_1, i.e. $d_{thr} = d_2$, hence t_3 is beyond the threshold, its transmission could be DCH or HS-DSCH, or not served. Then with predefined d_{thr}, the set of items corresponding to the transmission of flow f_1 are expressed in Equation 3.8.

$$K_1 = \{\{t^f_{1,1} t^f_{1,2}, t^f_{1,1} t^n_{1,2}\}, \{t^d_{1,3}, t^h_{1,3}, t^n_{1,3}\}\}, \ d_{thr} = d_2, \text{and } d_3 > d_{thr} \tag{3.8a}$$

K_1 is the set of classes associated with the flow f_1, where the braces describe the different classes of the set of items. Each class (two in our case) indicates the candidate allocation schemes for a set of users (e.g. t_1 and t_2), or for one user (e.g. t_3). According to the item definition in the 2nd proposition, the item $t_{j,k}^{ch}$ represents the user t_k receiving the flow f_j through the channel ch, with $ch \in \{f, d, h, n\}$ for FACH, DCH, HS-DSCH, and not served respectively. $t_{1,1}^f t_{1,2}^f$ indicates that the users t_1 and t_2 are served by a FACH for receiving f_1, and $t_{1,1}^n t_{1,2}^n$ indicates that they are not served. The value of the item $t_{1,1}^f t_{1,2}^f$ is $th = 2 \times bw_1$; the weight of the item $t_{1,1}^f t_{1,2}^f$ is the power of the FACH channel covering the farthest user t_2 with the channel bandwidth bw_1. While the value and the weight of $t_{1,1}^n t_{1,2}^n$ are both 0. The weight of the item $t_{1,k}^d$ is the power required by a DCH for sending f_1 to user t_k; and its value is $th = bw_1$. The weight of item $t_{1,k}^h$ is the power required by a HS-DSCH for sending f_1 to user t_k; and its value is $th = bw_1$. The value and the weight of $t_{1,k}^n$ are both 0. Therefore, the item value and weight could be calculated in advance. This overcomes the disadvantages of the 1st proposition.

However, during the optimization procedure, the value of d_{thr}, which defines the flexible FACH coverage for flow transmission, is dynamically changed, then the class structure for the associated flows will be different depending on the different value of d_{thr}. To solve this issue, the parallel MCKP is proposed, in which the problem processing is divided into two stages. In the first stage, for all K_j different d_{thr} are randomly generated. N_f is the total number of flows in the cell, i.e. the number of sets of classes. K_j is the set of classes for the flow f_j. Then with each d_{thr}, a corresponding subsets of items for each class K_j is generated, and different d_{thr} results in different items setting, therefore, parallel MCKP formulations. Hence in the first stage, parallel MCKP will be generated, each determined with different d_{thr} for K_j.

In the second stage these MCKPs are separately solved and the optimal solution of each one is recorded. Among these optimum solutions, the best one is selected as the final solution. However, the two-stage method increases the complexity of problem data generation (an input data file for each MCKP), and the execution time of the resolution procedure increases exponentially with the number of flows.

3.1.2.3 Proposition 3: single MCKP

In the 3rd proposition, the item definition is modified to reduce the formulation complexity introduce by the 2nd proposition. Instead of associating an item to one user or a set of users, this proposal defines that each item represents a complete allocation scheme for a given flow $f_{s_i,j}$, i.e. the transmission mode for all users in $M(s_i)$.

Let's take the scenario in Equation 3.7 as an example, this scenario is composed

of $N_f = 2$ flows. Each flow, f_j, gives rise to a class of items, each item concerns the transmission of f_j to all the users in $M(s_1) = \{t_1, t_2, t_3\}$. As shown in Equation 3.9, the FACH coverage, DCH and HS-DSCH allocation are indicated by the item format:

$$it = t_1^{\{f/n\}} t_2^{\{f/n\}} ... t_r^{\{f/n\}} t_{r+1}^{\{n/d/h\}} \ldots t_k^{\{n/d/h\}} \tag{3.9}$$

Where the symbol "/" indicates an alternative: $t_1^{\{f/n\}} t_2^{\{f/n\}} ... t_r^{\{f/n\}}$ indicates that the users $t_1, \ldots, t_r$ either receive the flow f through a FACH ($t_1^f t_2^f \ldots t_r^f$) or do not receive it ($t_1^n t_2^n \ldots t_r^n$). The users are ordered according to their distance from the base station and $t_1, \ldots, t_r$ are the nearest r users. Then the users from t_{r+1} to t_k are separately served by a dedicated channel (d), shared channel (h) or not served (n). In this way, for each item it, its value and weight are computed as an input constant. The item value (profit) v presents the allocated bandwidth th of the item it, and the item weight w is the required transmission power po. With the item definition in Equation 3.9, each scenario can be formulated as one MCKP.

Take the scenario in Equation 3.7 as an example. There are 40 candidate allocation schemes for flow f_1, hence 40 items in the class K_1. These item structures are shown in Table 3.1.

Table 3.1: Items list in class K_1

d_{thr}	Number of items	Items
d_3	1	$\{t_1^f t_2^f t_3^f\}$
d_2	3	$\{t_1^f t_2^f t_3^d, t_1^f t_2^f t_3^h, t_1^f t_2^f t_3^n\}$
d_1	9	$\{t_1^f t_2^d t_3^d, t_1^f t_2^d t_3^h, t_1^f t_2^d t_3^n, t_1^f t_2^h t_3^d, t_1^f t_2^h t_3^h, t_1^f t_2^h t_3^n, t_1^f t_2^n t_3^d, t_1^f t_2^n t_3^h, t_1^f t_2^n t_3^n\}$
0	27	$\{t_1^d t_2^d t_3^d, t_1^d t_2^d t_3^h, t_1^d t_2^d t_3^n, t_1^d t_2^h t_3^d, t_1^d t_2^h t_3^h, t_1^d t_2^h t_3^n, t_1^d t_2^n t_3^d, t_1^d t_2^n t_3^h, t_1^d t_2^n t_3^n,$ $t_1^h t_2^d t_3^d, t_1^h t_2^d t_3^h, t_1^h t_2^d t_3^n, t_1^h t_2^h t_3^d, t_1^h t_2^h t_3^h, t_1^h t_2^h t_3^n, t_1^h t_2^n t_3^d, t_1^h t_2^n t_3^h, t_1^h t_2^n t_3^n,$ $t_1^n t_2^d t_3^d, t_1^n t_2^d t_3^h, t_1^n t_2^d t_3^n, t_1^n t_2^h t_3^d, t_1^n t_2^h t_3^h, t_1^n t_2^h t_3^n, t_1^n t_2^n t_3^d, t_1^n t_2^n t_3^h, t_1^n t_2^n t_3^n\}$

The 1st line shows the item representing all users served by a FACH. In the 2nd line, the distance threshold d_{thr} is the distance of the user t_2. Two users t_1 and t_2 are served by a FACH, then the candidate transmission modes for t_3 are DCH (d), HS-DSCH (h) or not served (n). Similarly, the table lists the candidate solutions for f_1 with different values of d_{thr}.

3.1.2.4   Synthesis: problem formulation by single MCKP

Three propositions have been sequentially discussed to formulate the flexible MBMS RRM problem as MCKP. The first proposition defined the multi-dimensional

MKCP, in which each knapsack represents one transport channel. The knapsack capacities are defined either as maximal bandwidth or as maximal power of each channel. However, the weights of the items for the different knapsacks are not constant values and thus cannot be calculated in advance. Therefore, the proposal is not accepted.

The second proposition defined a single MCKP, and proposed a set of classes for one flow. Every item in one given class represents the transmission mode for a subset of users in the multicast group. Moreover, to illustrate the flexible changed FACH coverage, the second proposition included parallel MCKP defined by different distance thresholds. Each MCKP generates an optimum solution and the final solution is the best solution among these optimum solutions.

To simplify the second proposition, the third proposition formulated the studied problem by mapping it into one MCKP with only one class per flow. Each item in a given class describes the transmission mode of all users to receive the associated flow. The third proposition is selected as the approximation formulation of the MBMS RRM problem presented in Equation 3.10.

$$\text{Maximize } \sum_{j=1}^{N_f} \sum_{i=1}^{n_j} th_{ij} x_{ij} \tag{3.10a}$$

$$\text{Subject to:} \tag{3.10b}$$

$$\sum_{j=1}^{N_f} \sum_{i=1}^{n_j} po_{ij} x_{ij} \leq P_{mbms\_budget} \tag{3.10c}$$

$$N_f = \sum_{i=1}^{N_s} N_f(s_i) \tag{3.10d}$$

$$\forall j \in \{1, \ldots, N_f\}, \sum_{i=1}^{n_j} x_{ij} = 1 \tag{3.10e}$$

$$\forall j \in \{1, \ldots, N_f\}, \forall i \in \{1, \ldots, n_j\}, x_{ij} \in \{0, 1\} \tag{3.10f}$$

Where N_f is the total number of flows in the whole cell, n_j is the number of items in the class associated with the flow f_j. The optimization object is to choose one and only one item from each class such that the aggregated allocated bandwidth is maximized and the weight sum of selected items does not excess P_{mbms_budget}.

$$n_j = 1 + 3^{N_t(s_i)} + \sum_{r=1}^{N_t(s_i)-1} 3^{N_t(s_i)-r} \tag{3.11}$$

Equation 3.11 computes the number of items n_j in the class associated with the flow f_j. Let's assume that $f_j \in F(s_i)$. The value 1 represents the item where all users are served by one FACH. The value $3^{N_t(s_i)}$ computes the number of combinations corresponding to the cases where every user in the multicast group of service s_i is either served by DCH, HS-DSCH or not served. When the nearest r users are served by one FACH, the rest of users have $3^{N_t(s_i)} - r$ allocation assignments. Therefore, the number of items in one class is:

$$n_j = \sum_{r=1}^{N_t(s_i)-1} 3^r + 3^{N_t(s_i)} + 1 = \frac{3^{N_t(s_i)+1} - 1}{2} \qquad (3.12)$$

For a given scenario, the total number of items within this scenario is expressed in Equation 3.13. Therefore, the problem complexity is exponential increasing with the number of users.

$$\prod_{i=1}^{N_s} \prod_{1}^{N_t(s_i)} 2\left(3^{N_f(s_i)} + 1\right) \qquad (3.13)$$

The formulation of F2R2M (with channel code constraint relaxed) based on MCKP is achieved. Since MCKP is NP-Hard [74], this proves that the flexible MBMS RRM problem formulated as F2R2M is also NP-Hard. The complexity of the problem increases with the number of MBMS users and services.

3.1.3 Solution bound generation of F2R2M by solving MCKP

To solve the MCKP formulated MBMS RRM problem, different multimedia diffusion scenarios are transferred to MCKP model parameter. These scenarios are shown in Table 3.2.

In this table, each row presents one scenario, the setting of one scenario is indicated by its name: the number of service s, the total number of users u and the service transmission scheme, where S stands for scalable transmission scheme and N stands for non-scalable transmission scheme. The mixed integer programming (MIP) solver Gurobi [46] is used to solve the MCKP by using input parameters. In this work, the MCKP based formulation does not consider the orthogonality of OVSF code, hence the solutions found by the linear programming solver are the theoretical solution bounds, but not the real optimum solution of the F2R2M scenarios. Such solution bound can be used as an indicator to evaluate the real MBMS RRM algorithm performance.

Table 3.2: Simulation scenarios for MCKP based formulation

Scenario name	Flows of service	User number	Total traffic requirement
1s10u128N	s_1 (f_0: 128 kbps)	10	1280 kbps
1s10u128S	s_1 (f_1: 64 kbps, f_2: 64 kbps)	10	1280 kbps
2s20uSN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_0: 128 kbps)	10 10	2560 kbps
2s20uSS	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_1: 64 kbps, f_2: 64 kbps)	10 10	2560 kbps
3s30uSNN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_0: 128 kbps) s_3 (f_0: 64 kbps)	10 10 10	3200 kbps
3s30uSSN	s_1 (f_1: 32 kbps, f_2: 32 kbps, f_3: 64kbps) s_2 (f_1: 64 kbps, f_2: 64 kbps) s_3 (f_0: 64 kbps)	10 10 10	4480 kbps

3.1.3.1 Parameter transformation

In order to solve the MCKP based problem by Gurobi, the scenarios in F2R2M must be reformulated into MCKP parameters, which is described in a Gurobi model file. Each model file represents all the constraints and the objective described in Equation 3.10a. More precisely, the input model lists the decision variables, the linear objective function, and finally the linear formulation of constraints. Gurobi optimizer considers all possible values of decision variables that satisfy the given constraints, and return the combination of values that optimizes the stated objective function.

To illustrate the nature of the input file of MCKP, an example is given here, which composes one service s_0. s_0 is scalable transmitted with two flows $f_{s_0,1}$ and $f_{s_0,2}$. The multicast group of s_0 includes users t_1, t_2. According to the MCKP based formulation, two classes $K_{0,1}$ and $K_{0,2}$ are associated with the flows $f_{s_0,1}$ and $f_{s_0,2}$, respectively. Each class includes all potential allocation schemes for the corresponding flow. The number of items in each class is 13 (computed by Equation 3.13). The set of items in each class is shown in Equation 3.14.

$$K_{s_0,f_j} = K_{0,j} = \{it_{i,j,k}, \ldots\} = \{it_{0,j,1}, it_{0,j,2}, \ldots it_{0,j,13}\}, j = 1, 2 \qquad (3.14)$$

In Equation 3.14, i is for the service s_i and the kth item $it_{0,j,k}$ is a candidate allocation scheme for the flow $f_{s_0,j}$. The detailed allocation schemes represented by

these items are:

$$K_{0,1} = K_{0,2} = \{t_1^f t_2^f, t_1^f t_2^d, t_1^f t_2^h, t_1^f t_2^n, t_1^d t_2^d, t_1^d t_2^h, t_1^d t_2^n, t_1^h t_2^d, t_1^h t_2^h, t_1^n t_2^d, t_1^n t_2^h, t_1^n t_2^n\}$$
(3.15)

The item $t_1^{ch_m} t_2^{ch_n}$ represents that the user t_1 receives transmission from ch_m, and the user t_2 receives transmission from ch_n. Each item $it_{0,j,k}$ has two parameters according to its allocation scheme: the allocated bandwidth th_k and the transmission power po_k. The solution vector is represented as: $x_{0,1,1} \ldots x_{0,1,13}, x_{0,2,1} \ldots x_{0,2,13}$, with $x_{i,j,k} = 0, 1$ and $x_{i,j,k} = 1$ indicating that the selected allocation scheme is $it_{i,j,k}$, $x_{i,j,k} = 0$ means that the corresponding item is not selected in the solution. There is one and only one $x_{i,j,k}$ set to 1, $\forall k$. With above parameters, the format of model file for Gurobi is shown in Equation 3.16.

MCKP model file:

Maximize $Th_{0,1} + Th_{0,2}$ (3.16a)

Subject to

$T_{0,1} : th_{0,1,1} x_{0,1,1} + th_{0,1,2} x_{0,1,2} + \ldots th_{0,1,13} x_{0,1,13} - Th_{0,1} = 0$ (3.16b)

$P_{0,1} : po_{0,1,1} x_{0,1,1} + po_{0,1,2} x_{0,1,2} + \ldots po_{0,1,13} x_{0,1,13} - Po_{0,1} = 0$ (3.16c)

$K_{0,1} : x_{0,1,1} + x_{0,1,2} + \ldots x_{0,1,13} \leq 1$ (3.16d)

$T_{0,2} : th_{0,2,1} x_{0,2,1} + th_{0,1,2} x_{0,2,2} + \ldots th_{0,2,13} x_{0,2,13} - Th_{0,2} = 0$ (3.16e)

$P_{0,2} : po_{0,2,1} x_{0,2,1} + po_{0,1,2} x_{0,2,2} + \ldots po_{0,2,13} x_{0,2,13} - Po_{0,2} = 0$ (3.16f)

$K_{0,2} : x_{0,2,1} + x_{0,2,2} + \ldots x_{0,2,13} \leq 1$ (3.16g)

$Pt : Po_{0,1} + Po_{0,2} - Ptot = 0$ (3.16h)

$Pm : Po_{0,1} + Po_{0,2} \leq P_{mbms\_budget}$ (3.16i)

Binaries

$x_{0,j,k} \in \{0,1\}, \forall j, k$ (3.16j)

Equation 3.16a shows that Gurobi targets at finding the solution to maximize the allocated throughput of all flows, where $Th_{0,j}$ is the allocated throughput of flow $f_{s_{0,j}}$. Equation 3.16b to Equation 3.16j illustrate the problem constraints. Equation 3.16b and Equation 3.16e state that $Th_{0,j}$ is the profit (bandwidth) of the selected item for the jth class, $j = 1, 2$. Equation 3.16c and Equation 3.16f state that $Po_{0,j}$ is the weight of the selected item in the jth class. Constraint $K_{0,j}$ (Equation 3.16d and Equation 3.16g) guarantees that at least one and only one item is selected from each class. Constraint Pm (Equation 3.16i) guarantees that the total weight of the selected items does not exceed the maximum MBMS power budget. Equation 3.16h shows that the total weight $Ptot$ is the sum of $Po_{0,1}$ and $Po_{0,2}$. In Equation 3.16j, $x_{0,j,k}$ indicates the selection of the kth item in the jth class.

With the MCKP input model file, Gurobi finds solution with a single objective: the maximum throughput. While in the original F2R2M, the optimization objective

has two-dimension (throughput and power). Therefore, the optimization by Gurobi only aims to maximize the allocated throughput (i.e. minimize the throughput loss). To minimize the transmission power, represented by P_{tot} in Equation 3.16h, the value of the maximum power budget P_{mbms_budget} is gradually decreased manually, hence we obtain different optimum solutions for different P_{mbms_budget} settings.

3.1.3.2   Gurobi results

The Gurobi model files for small size scenarios are generated and their optimum solution found by Gurobi are shown in Table 3.3. For large size scenario, for example one service with single flow with 17 users, the number of items in each class will be 1.937×10^9, and the generated model file size is more than 5 Gigabytes. Therefore the generated MCKP model files for large size scenarios are too large to be handled by the solver.

Table 3.3: Solutions found by Gurobi

Scenarios	Gurobi solutions
1s10u128N	0%, 1.352 W
1s10u128S	0%, 1.34 W
2s20uSN	0%, 4.328 W
2s20uSS	0%, 4.263 W
3s30uSNN	0%, 5.297 W
3s30uSSN	0%, 5.231 W

In this work, the MCKP based formulation describes the scenario with relaxed channel code constraint; the MCKP based formulation does not consider the OVSF code orthogonality as in F2R2M, thus the channel code are always fully available. Therefore, the solutions solved by MCKP in Table 3.3 are not practical solutions for each scenario, but the theoretical solution bounds for MBMS RRM, which could be used as the reference for evaluating the practical MBMS RRM algorithms.

3.2   Fitness landscape analysis

Metaheuristic provides efficiency scheme to solve large problem in a reasonable time consuming. In metaheuristic, the search space is the space of all possible solutions that can be visited during the search. In which, a neighborhood operator is used to move from the current solution to a neighbor solution within the search

space. By applying one neighborhood operator on the current solution, a set of neighbor solutions can be generated, constructing the neighborhood of the current solution. Therefore, the operator is an essential parameter in metaheuristic as it acts on the dynamics of the search. Different neighborhood can be constructed by different neighborhood operator.

Generally, for a given combinatorial optimization problem, its fitness landscape is defined as $\mathcal{L} = (X, f, d)$ [66, 39], where f denotes the fitness function, X is the set of solutions depending on the neighborhood structure $\mathcal{N}$ and d represents the distance between two feasible solutions. Therefore, the structure of a fitness landscape is linked to the metaheuristic dynamics such as the solutions of the search space and the objective functions. Hence the fitness landscape allows to characterize the problem structure and provides the understanding of the search strategies. It is an important technique to analyze combinatorial optimization problem [75].

In this section fitness landscape analysis technique will be used for analyzing the solution space of MBMS RRM and evaluating the move operators, which will be later used for metaheuristic searches.

3.2.1   Introduction of fitness landscape analysis

The notion of fitness landscape was firstly proposed in [82], aiming to analyze the gene interaction in biological evolution. Each genotype has a fitness and the distribution of the fitness values over the genotypes space constitutes a fitness landscape.

In chapter 2, two neighborhood operators were proposed to modify the current solution to a new solution. They are the single insert operator δ_{SI} and the multiple insert operator δ_{MI}, which define two associated neighborhood structures. The neighborhood operator is the key element that leads to different optimization performance. To characterize the studied problem, the fitness landscape analysis is conducted to analyze the two neighborhood functions for different scenarios. This study can also help us to determine the better operator in optimization algorithms by studying the operator performance.

Existing studies [39, 66] have proved that a fitness landscape has several important properties. In our work, we focus on the following properties of a fitness landscape:

- the distribution of feasible solutions within the search space.
- the distribution of solution fitness.
- the links between the solution distance and the difference of solutions fitness.

Generally, to achieve such analysis based on these properties, two sets of solutions

are required. S_{ini} is a set of initial solutions randomly chosen from the search space. S_{lo} is a set of local optima solutions found by applying Greedy Local Search (GLS) starting from a solution in S_{ini}. Then the fitness landscape $\mathcal{L}_\mathcal{N}$ is characterized through these two solution sets S_{ini} and S_{lo}.

3.2.2 Solution representations

Two mathematical solution representations are defined, and then the solution distance measurement between two feasible solutions are developed. The solution distance indicates if the solutions are close to each other. Based on the solution distance, then the relationship between the solutions and the landscape can be analyzed.

3.2.2.1 Representation A

Equation 3.17 shows the solution representation A in which the radio resource allocation solution of one cell, $x(c)$, is represented as a matrix.

$$x(c) = \begin{matrix} f_{s_1,1} & f_{s_1,2} & \cdots & f_{s_1,N_f(s_1)} & \cdots & f_{s_{N_s},N_f(s_{N_s})} \\ \begin{pmatrix} ch_{1,1} & ch_{1,2} & \cdots & ch_{1,N_f(s_1)} & \cdots & ch_{1,N_f} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ ch_{i,1} & ch_{i,2} & \cdots & \cdots & \cdots & ch_{i,N_f} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ ch_{N_t,1} & ch_{N_t,2} & \cdots & ch_{N_t,N_f(s_1)} & \cdots & ch_{N_t,N_f} \end{pmatrix} \end{matrix} \quad (3.17a)$$

Subject to: $\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad$ (3.17b)

$$ch_{i,j} \in \{-1, 0, 1, 2, 3\} \quad\quad\quad\quad\quad\quad (3.17c)$$

$$\forall i < i', \ d(t_i) \le d(t_{i'}) \quad\quad\quad\quad\quad\quad (3.17d)$$

$x(c)$ has N_t rows and N_f columns, $N_f = \sum_{i=1}^{N_s} N_f(s_i)$ in the cell, is the number of all flows and $N_t = \sum_{i=1}^{N_s} N_t(s_i)$, is the number of all users in the cell. As shown in Equation 3.17d, the users are ordered in ascending order of distance from the Node B, therefore, the user t_i is closer than the user $t_{i'}$, hence $i < i'$ and $d_i \le d_{i'}$. An element of this matrix, $ch(i,j)$, indicates the transport channel allocated to the user t_i to receive the flow f_j. The value of $ch(i,j)$ could be 0, 1, 2 or 3 if the user t_i belongs to $UE_{fach}(f_{s_i,j})$, $UE_{dch}(f_{s_i,j})$, $UE_{hs}(f_{s_i,j})$ or $UE_{noch}(f_{s_i,j})$, respectively. -1 means that the user does not belong to this multicast group.

3.2.2.2   Representation B

In Equation 3.18, the solution representation B defines that the solution for the flow $f_{s,j}$ is a vector of users.

$$x(f_{s,j}) = (\underbrace{-1, 1, 2, \ldots,}_{UE_{fach}(f_{s,j})} \underbrace{-1, i, \ldots,}_{UE_{dch}(f_{s,j})} \underbrace{-1, j, \ldots,}_{UE_{hs}(f_{s,j})} \underbrace{-1, k, \ldots, t_{Nt}}_{UE_{noch}(f_{s,j})}) \qquad (3.18)$$

The four user subsets are delimited by the number -1, and indicates the transmission channel for the users in each subset. In each subset, users are ordered in ascending order of distance from the base station.

Subsequently, the solution of the cell $x(c)$ is a vector of solutions for the service $x(s_i)$:

$$x(c) = \{x(s_1), \ldots, x(s_{Ns})\}, \forall s_i \in S(c) \qquad (3.19a)$$

$$x(s_i) = \begin{cases} \{x(f_{s_i,0})\} & \text{if } s_i \text{ is a single layer service} \\ \{x(f_{s_i,1}), [x(f_{s_i,2}), [x(f_{s_i,3}) \ldots]]\} & \text{otherwise} \end{cases}$$

$$(3.19b)$$

3.2.3   Distance measurements

The solution distance should evaluate the number of elementary moves required to move from one solution to another. In order to conduct the landscape analysis, the distance measurement between the solutions is required.

3.2.3.1   Hamming distance

The hamming distance is a well-known distance in combinatorial optimization, it corresponds to the number of different elements between two solutions. The hamming distance d_{Ham} is used to measure the distance between two feasible solutions in the format of the representation A.

3.2.3.2   Comparative distance

For the representation B, the solution distance is measured according to structural comparisons between solutions, named comparative distance d_{Com}. Let's assume that the current solution x and the new solution x' are based on the representation B. This measurement compares the solutions for the same flow in x and x', and then the users allocated in the different channel in x and x' are counted, this number is expressed by d_{Com}.

The comparative algorithm measures the exact minimum number of applications of the single insert operator δ_{SI} to move from x to x'. It could also be used to measure the approximate distance of solutions generated by the multiple insert operator δ_{MI}. d_{Ham} compares the different allocated values for all users in cell. While d_{Com} only considers the users within one multicast group for each flow. Therefore the representation B requires less memory space and the associated comparative distance computation is faster. Hence the solution representation B and the comparative distance are used for the following analysis.

3.2.4 Greedy Local Search and its application on landscape analysis

A Greedy Local Search (GLS) algorithm is used to generate the local optimum solutions from the randomly generated initial solutions. The search procedure is described in Algorithm 2. In each single trial of GLS, the search begins from an initial solution x_{ini}. The neighborhood operator specifies a set of allowable modifications to the current solution x at each iteration. Among these neighborhood solutions, the best solution offering the minimum fitness will be selected to replace the current solution. The search procedure terminates when the new solution is worse than the current solution. The final solution is named local optimum solution x_{lo}.

Algorithm 2 Greedy Local Search [45]

randomly generate an initial solution x_{ini}
$x \leftarrow x_{ini}$
repeat
   generate $x' \in N(x)$ // the best solution from $N(x)$
   if $f(x') \leq f(x)$ **then**
      $x \leftarrow x'$ // downhill move
   end if
until $x \neq x'$
$x_{lo} \leftarrow x$ // local optimum

By applying GLS to each solution $x_{ini} \in S_{ini}$, a population of local optimum solution is obtained.

3.2.5 Experiments and results analysis

Six scenarios are selected to conduct the fitness landscape analysis. Their user distributions are illustrated in Figure 3.1.

Each figure illustrates two scenarios with the same multicast groups and traffic load but the scalability is different. For example, in 3.1(a), the service s_1 at 128
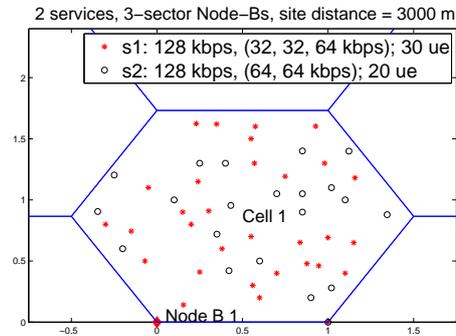
(a) scenarios 1,2: 2s50uSN/SS
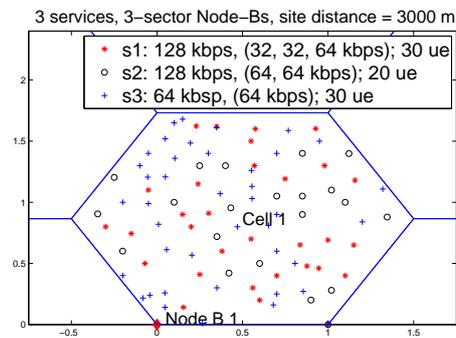
(b) scenarios 3,4: 3s80uSNN/SSN
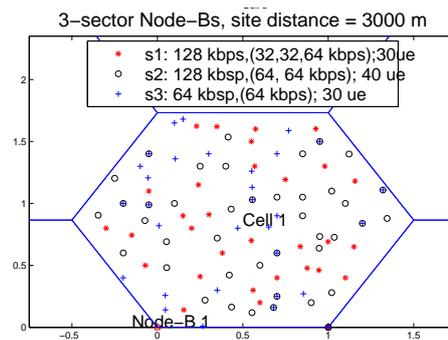
(c) scenarios 5,6: 3s100uSNN/SSN

Figure 3.1: User distributions of six scenarios

kbps is transmitted as three flows at 32, 32 and 64 kbps, but the service s_2 of 128 kbps is transmitted as a single 128 kbps flow in one scenario (*2s50uSN*), while in the other scenario it is transmitted as two flows at 64 kbps (*2s50uSS*).

3.2.5.1 Analysis of search space

It is not possible to exhaustively enumerate all the solutions from the search space, as the solutions number is exponential. The solution distribution could only help to understand if the solutions are close or not, if they are similar or different, if the local optima are far away from each other in the search space, etc. Then the indicators are computed to estimate the width or the diversity of the search space.

To study the distribution of feasible solutions in the search space, we define two kinds of distance: d_{ini} and d_{lo} are a set of solution distances. In d_{ini}, each value is the distance measurement among any two solutions in S_{ini}. In d_{lo}, each value is the distance among any two solutions in S_{lo}. Table 3.4 and Table 3.5 present the minimum, the maximum and the median values (first and third quartile are also given) of these distances. L_{SI} and L_{MI} are the search spaces constructed by the operators δ_{SI} and δ_{MI}, respectively.

Table 3.4: Distance between solutions of L_{SI}

Scenarios	d_{ini} in $S_{ini,SI}$			d_{lo} in $S_{lo,SI}$		
	Min	Median$_{Q1,Q3}$	Max	Min	Median$_{Q1,Q3}$	Max
2s50uSN	15	$62_{54,72}$	109	17	$62_{54,71}$	108
2s50uSS	20	$64_{55,74}$	114	20	$65_{55,75}$	117
3s80uSNN	20	$64_{55,73}$	116	20	$64_{55,74}$	116
3s80uSSN	26	$66_{56,76}$	148	20	$66_{56,76}$	148
3s100uSNN	10	$53_{44,63}$	92	12	$54_{45,63}$	99
3s100uSSN	21	$64_{55,73}$	130	19	$65_{55,76}$	180

The first three columns list the statistics of d_{ini} and show that the random initial solutions for both search spaces are quite homogeneous. Comparing the statistics of d_{lo}, the local optima are different for two landscapes. In $S_{lo,MI}$, the minimum value of 0 indicates that the identical local optimum is found from different initial solutions. While in $S_{lo,SI}$ no local optimum solutions have the same value. The quartiles (median, Q1 and Q3) show that the search space of $S_{lo,MI}$ is more concentrated than for $S_{lo,SI}$ even if the maximum values are larger. Therefore, for the population of local optima, L_{MI} appears closer than L_{SI}.

Table 3.5: Distance between solutions of L_{MI}

Scenarios	d_{ini} in $S_{lo,MI}$			d_{lo} in $S_{lo,MI}$		
	Min	$Med_{Q1,Q3}$	Max	Min	$Med_{Q1,Q3}$	Max
2s50uSN	19	$62_{53,71}$	108	0	$34_{23,45}$	109
2s50uSS	23	$64_{54,73}$	121	0	$33_{22,46}$	127
3s80uSNN	21	$65_{54,73}$	115	0	$43_{30,57}$	138
3s80uSSN	20	$64_{55,73}$	116	0	$35_{24,49}$	156
3s100uSNN	24	$55_{64,73}$	114	0	$63_{39,86}$	157
3s100uSSN	21	$64_{55,73}$	129	0	$38_{23,56}$	194

3.2.5.2   Analysis of fitness space

The fitness value represents the quality of a solution. Figure 3.2 shows the distribution of fitness values of S_{ini} and S_{lo} for *3s100uSNN* for both objectives and for both move operators.
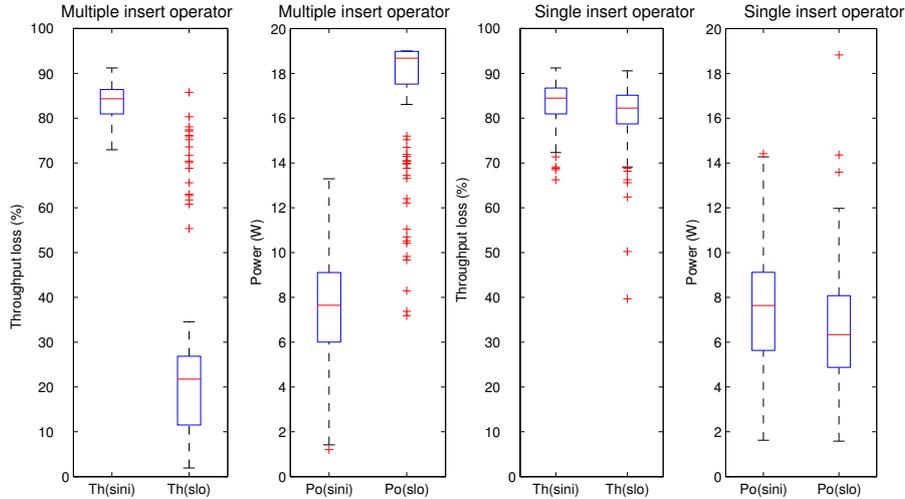
Figure 3.2: Two fitness spaces of scenario 5 (3s100uSNN)

Firstly, it can be observed that the fitness values of S_{ini} are well diversified. According to the similarity of the statistics in S_{ini} for all scenarios (Table 3.4 and Table 3.5), it can be observed that the random initial populations are uniformly distributed and in a similar way both for δ_{SI} and δ_{MI} tests. As well, the fitness of all local optimum for both operators are better than the associated random initial

solutions on Th criteria.

Table 3.6: Fitness values of $S_{lo,SI}$

Scenarios		Min	Med$_{Q1,Q3}$	Max	Mean
2s50uSN	Th(c) %	4.5	57.5$_{52.25,62}$	69	56.21
	Po(c) w	1.5	6.53$_{5.03,8.33}$	18.54	6.44
2s50uSS	Th(c) %	0	58.75$_{54,62}$	69.5	10.74
	Po(c) w	1.69	6.24$_{4.69,7.93}$	16.3	6.44
3s80uSNN	Th(c) %	37	87.5$_{82,92}$	99	86.61
	Po(c) w	2.10	6.56$_{4.96,8.21}$	18.96	6.74
3s80uSSN	Th(c) %	20.5	88$_{83.5,91.5}$	99	86.34
	Po(c) w	2.1	6.56$_{4.96,8.21}$	18.96	6.75
3s100uSNN	Th(c) %	20.5	88$_{83.5,91.5}$	99	86.34
	Po(c) w	2.1	6.56$_{4.96,8.21}$	18.96	6.75
3s100uSSN	Th(c) %	15.88	75.88$_{72.64,77.94}$	82.35	74.4
	Po(c) w	1.37	65.36$_{4.82,8.24}$	18.95	6.65

Moreover, in the first and third subfigures of Figure 3.2, the fitness values of the local optima obtained by the two operators are not flat. Actually, the quality of $S_{lo,MI}$ is better than that of $S_{lo,SI}$ ($Th_{MI} < Th_{SI}$). Table 3.7 and Table 3.6 give the fitness of the local optima in the two landscapes L_{SI} and L_{MI}. It is shown that δ_{MI} finds solutions with smaller $Th(c)$, hence better solutions than δ_{SI}. This can be explained by the fact that, unlike δ_{SI}, δ_{MI} moves a block of elements and explores a larger search space, so that it can more easily escape from the local optima.

3.2.5.3 Analysis of links between distance and fitness

The step length is the number of moves from an initial solution to its associated local optimum. In F2R2M, the step length is defined as the number of times that the greedy local search method calls the neighborhood operator.

Table 3.8 presents the statistics of the step lengths to find local optima through δ_{MI} and δ_{SI}. In average, δ_{SI} moves to shorter distances than δ_{MI}, which may make δ_{SI} walks nearby the initial solution without exploring much better solution even if the maximum step length is higher for δ_{SI}. Therefore, L_{SI} seems to have less capacity of combining exploration and exploitation than L_{MI}, which explains that in Table 3.4 the values of $S_{lo,SI}$ are close to the $S_{ini,SI}$.

To investigate how the population of local optima is distributed in the search

Table 3.7: Fitness values of $S_{lo,MI}$

Scenarios		Min	Med$_{Q1,Q3}$	Max	Mean
2s50uSN	Th(c) %	0	$0_{0,0}$	58	1.7
	Po(c) w	7.08	$14.89_{12.95,18.18}$	19.0	15.28
2s50uSS	Th(c) %	0	$0_{0,0}$	62.5	0.94
	Po(c) w	5.81	$14.63_{14.36,15.02}$	16.58	14.57
3s80uSNN	Th(c) %	0	$4_{0,15}$	91.5	9.43
	Po(c) w	5.543	$17.29_{15.92,18.79}$	18.99	17.2
3s80uSSN	Th(c) %	0	$1_{0,2}$	93.5	3.8
	Po(c) w	6.38	$17.57_{17.22,17.63}$	19.0	17.2
3s100uSNN	Th(c) %	0	$1_{0,2}$	93.5	3.8
	Po(c) w	6.38	$17.57_{17.22,17.63}$	19.0	17.2
3s100uSSN	Th(c) %	1.76	$4.12_{4.12,5.88}$	78.24	7.29
	Po(c) w	5.67	$18.38_{18.25,18.41}$	18.99	18.11

space relative to the best solution found, Figure 3.3, Figure 3.4 and Figure 3.5 present the plots of fitness distance scatter of all scenarios.

In each subfigure, e.g. Figure 3.3(a), in the left side the y-axis illustrates the fitness difference in terms of throughput loss, and the x-axis illustrates the solution distance with d_{Com}. In the right side, the y-axis illustrates the fitness difference in terms of power consumption, and the x-axis illustrates the solution distance. These plots depict the fitness gap between local optima and best solution found according to their distances in terms of required moves number. The plots determine the relationship between the fitness gap and the move distance. When the distance to

Table 3.8: Step lengths of the two landscapes

Scenario	step length in L_{SI}				step length in L_{MI}			
	Min	Median	Max	Mean	Min	Median	Max	Mean
2s50uSN	1	7	80	9.70	1	18	57	25.97
2s50uSS	1	7.5	126	10.74	1	21	62	26.93
3s80uSNN	1	6	85	8.78	2	43	100	42.69
3s80uSSN	1	7	127	11.1	1	61	114	67.61
3s100uSNN	1	7	126	8.94	1	19.5	93	99.88
3s100uSSN	1	11	181	12.62	5	105	166	99.88
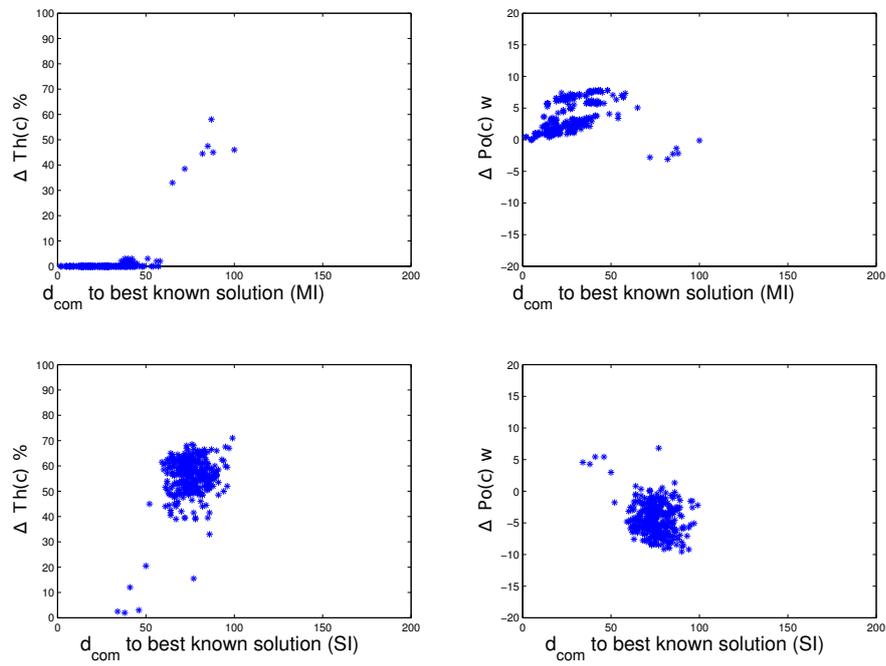
(a) scenario 1: 2s50uSN
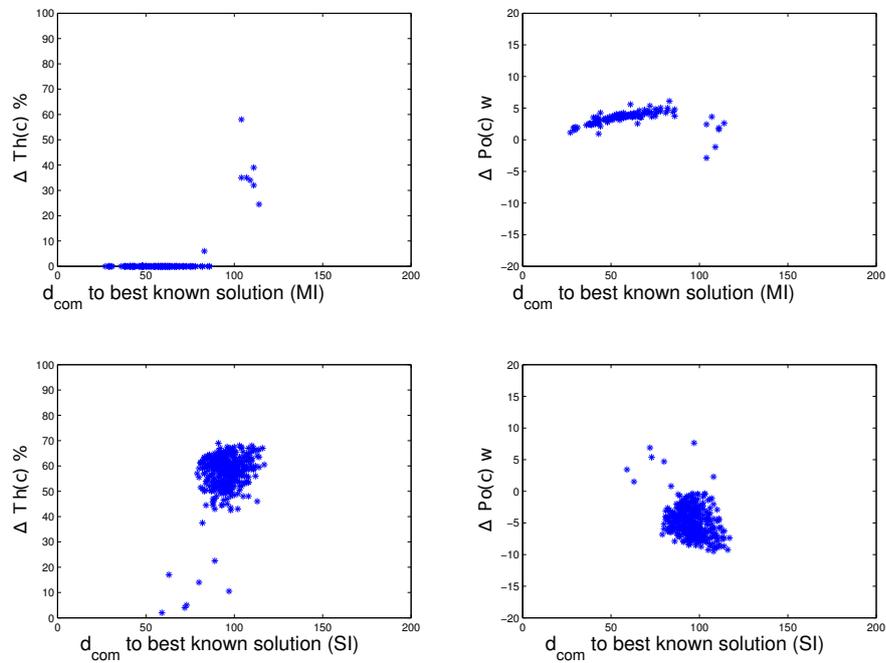
(b) scenario 2: 2s50uSS

Figure 3.3: Fitness distance scatter: scenarios 1, 2

the best solution found becomes smaller (the fitness difference is decreased), the search procedure is expected to be easier.

Figure 3.3, Figure 3.4 and Figure 3.5 reveal that all local optima converge in a small region of the search space and the local optima found by δ_{MI} are closer to the best solution found than the local optima found by δ_{SI}. When the points are located in different distance from the best solution found, their fitness difference are varied, which means that both search space are rugged. But in the search space of δ_{SI}, the fitness gap and the distance from the best solution found is less correlated than in the search space of δ_{MI}. The search difficulty with L_{SI} is harder than L_{MI}.

Moreover, the scenarios 1 and 2 (Figure 3.3(a) and Figure 3.3(b)) show higher correlation than the scenarios 5 and 6, which indicates that the difficulty of exploring the search space is increased with the increasing of scenario complexity.

Finally, the study of fitness landscape reveals that the local optima in L_{MI} are closer to each other than in L_{SI}, and that δ_{MI} can explore larger neighborhood space to reach better solution than δ_{SI}. Therefor L_{MI} outperforms L_{SI}.

3.2.5.4 Comparison of two neighborhood operators

Table 3.9 shows the best performance with the two neighborhood operators through the Greedy Local Search algorithm. The fitness value of the solutions are presented in terms of the percentage of loss throughput and the consumed power in Watts of all MBMS multicast services within one cell.

Firstly we prove the feasibility of some scenarios with F2R2M. Then δ_{MI} always offers good solution with higher satisfied throughput and less power consumption than δ_{SI}, which proves that δ_{MI} has more ability to escape from local optima, while δ_{SI} can only stay in basins.

3.3 Synthesis

In this chapter, the mathematical analysis was conducted in the proposed flexible radio resource allocation model for MBMS RRM. In section 3.1, a mathematical formulation of MBMS RRM is investigated on the basis of the knapsack problem variants. Three propositions were done to discuss the possibility of this formulation, and the work shows that the single MCKP formulation is the best choice. Then the F2R2M is approximated as a MCKP, with the relaxation of the channel code constraint. In this formulation, a set of classes are associated with the flows of the MBMS service. Each class includes a set of items, each item is a candidate allocation schemes of the corresponding flow. The problem target is to select one and only one
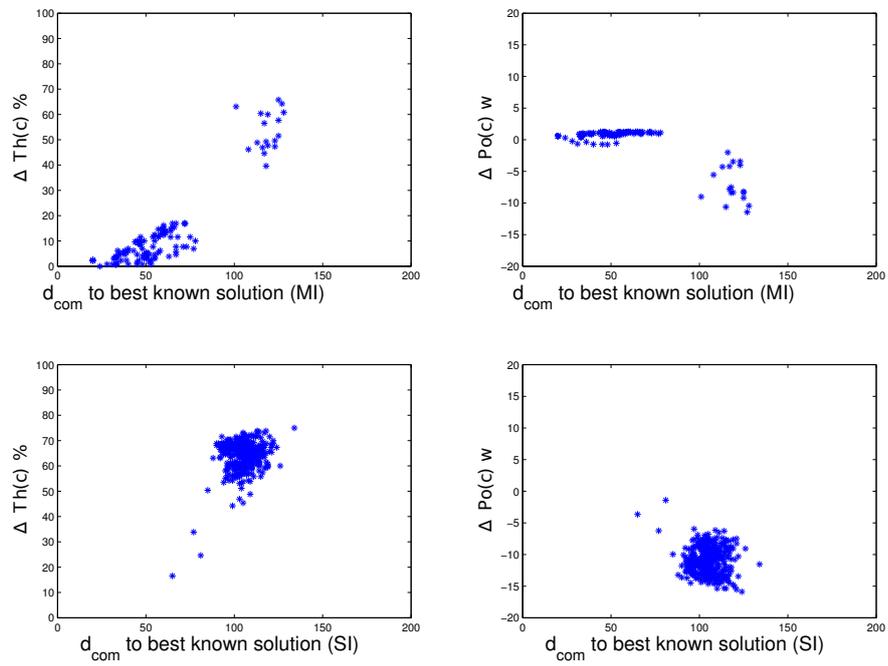
(a) scenario 3: 3s80uSNN
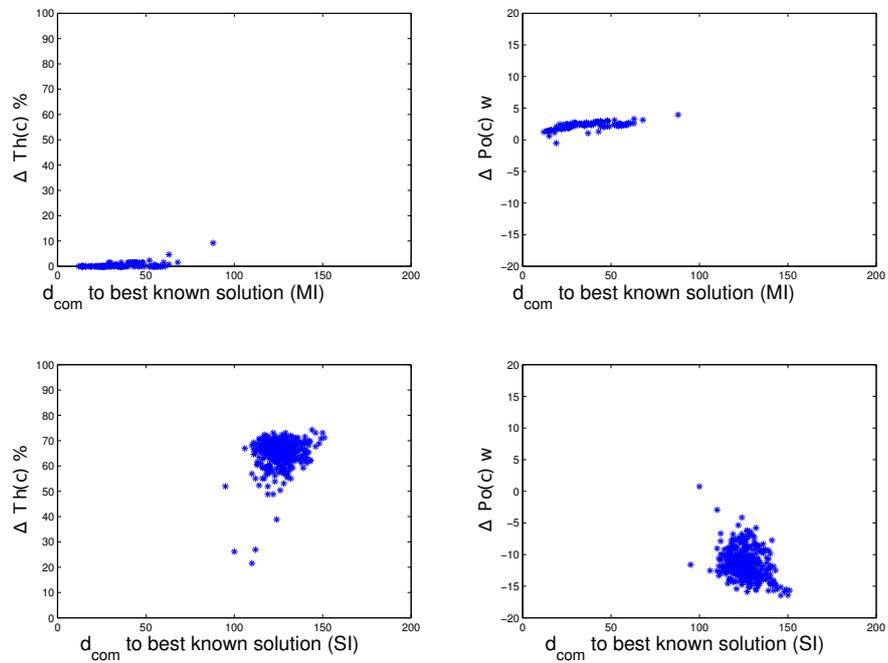
(b) scenario 4: 3s80uSSN

Figure 3.4: Fitness distance scatter: scenarios 3, 4
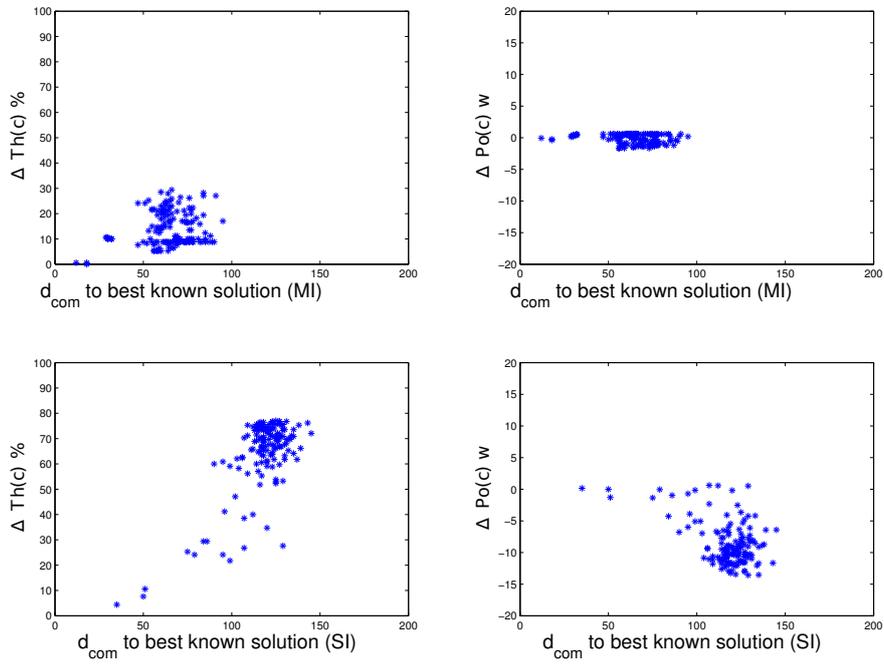
(a) scenario 5: 3s100uSNN
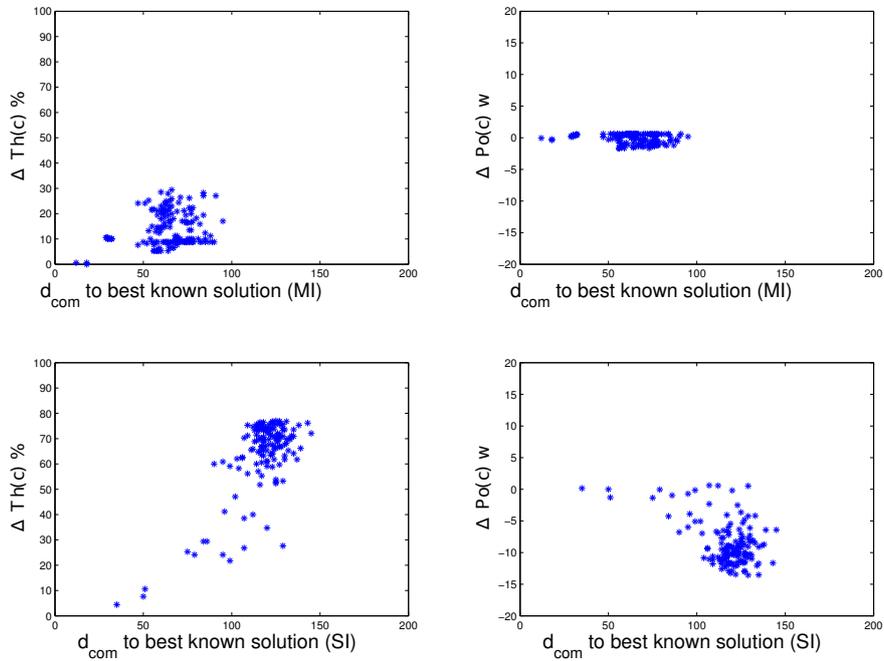
(b) scenario 6: 3s100uSSN

Figure 3.5: Fitness distance scatter: scenarios 5, 6

Table 3.9: Performance of the Greedy Local Search with δ_{SI} and δ_{MI}

Scenarios	F2R2M with δ_{SI}			F2R2M with δ_{MI}		
	best	mean	std.	best	mean	std.
2s50uSN	4.5%	56.2%	8.12	0%	1.7%	8.18
	18.5	6.4	8.12	10.19	15.28	2.56
2s50uSS	0%	57.29%	7.82	0%	0.94%	6.46
	15.58	6.44	2.27	13.06	14.57	0.9
3s80uSNN	25.4%	66.62%	5.53	0%	7.25%	11.77
	16.9	6.73	2.29	15	17.12	2.0
3s80uSSN	15.4%	66.42%	6.73	0%	2.9%	11.26
	16.5	6.75	2.45	14.4	17.2	1.23
3s100uSNN	36.47%	74.98%	4.5337	1.76%	19.57%	12.61
	18.82	6.58	2.3244	18.39	18.03	1.5856
3s100uSSN	15.9%	74.4%	7.13	1.76%	7.29%	11.93
	18.12	6.65	2.55	17.5	18.12	1.4

item from each class in such manner that the sum of the values (throughput) of the selected items is maximized and the knapsack capacity (power) is not exceeded.

The multiple-choice knapsack problem is a NP-hard optimization problem, hence the problem of flexible radio resource allocation for scalable MBMS is proved as NP-Hard problem. Six small-sized scenarios are designed and re-formulated as MCKP model parameters. These model parameters are input of the Gurobi solver, which is used to solve the MCKP. Because the MCKP based formulation does not consider the orthogonality of OVSF code, the linear programming solver is not a practical approach and the solutions found are only theoretical solution bounds of the F2R2M scenarios.

In section 3.2, two solution representations were defined, and the corresponding distance measurements between two feasible solutions were proposed. Based on the solution distance and two neighborhood operators, the fitness landscape analysis was conducted to study the characteristics of the proposed problem and the operators. The fitness distance plots showed that both search spaces generated by the neighborhood operators are rugged. Among two proposed neighborhood operators, the multiple insert operator δ_{MI} seems to be more efficient than the single insert operator δ_{SI} even if its computation is more complicated.

Solving MBMS RRM Problem by Simulated Annealing

In chapter 3, the studied MBMS RRM problem has been proved as NP-Hard. This combinatorial optimization problem thus cannot be handled by exhaustive search. To solve this problem, the metaheuristic is selected since it requires reasonable amount of effort to get a good (but might be non-optimal) solution. In this chapter, a popular metaheuristic algorithm, the Simulated Annealing (SA) is investigated to solve the MBMS RRM problem based on F2R2M. This work shows that it is feasible to modify and map the SA algorithm on the proposed model, named as F2R2M-SA. The simulations are conducted and show that the proposed F2R2M-SA can generate better solutions than the state-of-the-art approaches within acceptable time.

Contents

4.1   Introduction of simulated annealing algorithm

The complexity of the studied problem has been analyzed in chapter 3, and it is proved that the size of the search space increases exponentially with the number of users. To solve the combinatorial optimization (CO) problem, an algorithm is required to find a good (eventually) near-optimal solution, with a reasonable amount of effort. In the past three decades, metaheuristic algorithms tried to combine basic heuristic methods offering efficiency and rapidity [27]. The simulated annealing algorithm is one of the first metaheuristic that has a strategy to avoid local optima. The origins of this algorithm for solving CO are first presented in [61] and [83].

Simulated annealing is so named because of its similarity to the solid annealing process, in which a heated crystalline solid is allowed to cool very slowly until it achieves its minimum lattice energy state, and thus is free of crystal defects [45]. The general framework of the simulated annealing is outlined in algorithm 3.

Algorithm 3 General SA algorithm [36]
---
   select an initial solution x_0
   select an initial temperature $T = T_0, T_0 \geq 0$
   set change counter $c = 0$
   select a repetition schedule, M_k, that defines the number of iterations executed at each temperature, T_k
   repeat
     repeat
       set repetition counter $c = 0$
       generate $x\prime \in N(x)$
       calculate $\delta = f(x\prime) - f(x)$
       if $\delta < 0$ **then**
         $x \leftarrow x\prime$ // downhill move
       else if $random(0,1) < \exp(\frac{-\delta}{T_k})$ **then**
         $x \leftarrow x\prime$ // uphill move
       end if
       $c \leftarrow c + 1$
     until $c = M_k$
     $k \leftarrow k + 1$
     compute T_k
   until stopping criterion is met

To describe the features of the general simulated annealing algorithm, several definitions need to be presented. Let Ω be the solution space, let f be an objective function defined on the solution space. The goal is to find an optimum solution x^*, $x^* \in \Omega$ such that $f(x^*) \leq f(x), \forall x \in \Omega$. Let's define $N(x)$ to be the neighborhood function for $x \in \Omega$. The neighborhood solution $x\prime \in N(x)$ can be reached in a single

iteration by applying a neighborhood operator to the current solution x.

As shown in algorithm 3, SA starts with an initial solution $x_0 \in \Omega$. The temperature parameter T is initialized by a predefined value T_0. At each decreasing temperature T_k, a series of Metropolis chains is performed (the inner loop in algorithm 3). The aim of each Metropolis chain is to permit the system to reach the thermal equilibrium, in which the energy state is being minimized. At each iteration within the Metropolis chain, a neighborhood solution $xⅠ \in N(x)$ is generated, either randomly or using some pre-specific rules. The candidate solution $xⅠ$ is accepted as the current solution x based on the acceptance probability.

The major purpose of the acceptance probability p_{acc} is to accept $xⅠ$ with the worse result than x, in order to explore new solutions and to escape from local optimum. p_{acc} is generally computed following the Boltzmann distribution function [45], by Equation 4.1.

$$p_{acc} = \begin{cases} \exp\left(-\frac{f(xⅠ)-f(x)}{T_k}\right) & \text{if } f(xⅠ) - f(x) > 0 \\ 1 & \text{if } f(xⅠ) - f(x) \leq 0 \end{cases} \tag{4.1}$$

where T_k is the temperature parameter at the kth iteration (the outer loop in algorithm 3). This acceptance probability is the basic element of the search mechanism in the simulated annealing. $f(xⅠ)$ and $f(x)$ denote the energies (objective function values) associate with the solutions $x \in \Omega$ and $xⅠ \in N(x)$, respectively. With a fixed temperature, the higher the difference $\delta = f(xⅠ) - f(x)$, the lower the probability to accept a move from x to $xⅠ$. Moreover, the higher is the temperature, the higher is the probability to receive non-improving moves.

In the SA framework, a full annealing schedule is defined by an initial temperature T_0, a cooling schedule to reduce the temperature, and a stopping criterion.

The initial temperature should be chosen such that (nearly) every neighborhood solution is accepted, thus the entire solution space could be captured. A high temperature could increase the search space, however it will result in a larger number of iterations without guarantee on the solution quality. Therefore, T_0 is chosen by experimentation depending upon the problem feature. Van Laarhoven has proposed a method to analytically determine T_0 for a given number of solution sampling [48]. This procedure starts from a random initial solution with an initial acceptance ratio χ_0, and then a series of random walk of solution modifications is performed. In this random walk procedure, for minimization problem, the uphill move, also named bad move, brings non-improving solution comparing to the current solution. While the downhill move is good move, bringing improvement of the current solution. Equation 4.2 shows the calculation of T_0. In Equation 4.2a, Δf_0 is the average increase of the objective function between x' and x during the random walk, and χ_0 is the number of *accepted* bad moves divided by the number of *attempted* bad moves. In Equation 4.2b, Δf^+ is the increase of fitness between x' and x, brought

by these bad moves.

$$T_0 = -\frac{\Delta f_0}{ln(\chi_0)} \tag{4.2a}$$

$$\Delta f_0 = \frac{\Delta f^+}{\text{number of accepted bad moves}} \tag{4.2b}$$

A similar method of selecting initial temperature, named *find-T_0* has been proposed in [50]. Firstly, T_0 is set to zero and a sequence of moves is implemented. After each move, a new value of T_0 is calculated according to Equation 4.3.

$$T_0 = \frac{\Delta^+}{\ln(\frac{m_2}{m_2\chi_0 - m_1(1-\chi_0)})} \tag{4.3a}$$

$$\Delta^+ = \frac{f(x) - f(x')}{m_2} \tag{4.3b}$$

In Equation 4.3a, χ_0 is the initial acceptance ratio of bad moves (e.g. set 0.8 in [50]). m_1 and m_2 are the number of good (downhill) and bad (uphill) moves obtained so far to solve a minimization problem. For solving a maximization problem, downhill and uphill should be swapped. In Equation 4.3b, x and x' are respectively the current solution and the new solution in each iteration, Δ^+ is the average fitness difference brought by uphill moves so far. After a given iteration of moves, the final value of T_0 is used as the initial temperature.

The cooling schedule determines the decreasing of the temperature in the SA search procedure. The earliest annealing schedules were based on the analogy with the physical annealing. In early SA study, a proportional temperature is used, $T_{k+1} = \alpha T(k)$. α is a constant value known as the cooling factor, most of the time it ranges from 0.80 to 0.99 [76]. At the end of the search, the temperature does not allow any higher energy level, which is called the frozen state. However, [76] concluded that a proportional temperature cooling schedule does not lead to equilibrium at low temperature. Nowadays there are three important annealing schedules to reduce the SA temperature:

- Logarithmic schedule [42], $T_k = \frac{T_0}{LOG(k)}$. The temperature decreases quickly at the beginning of the search, then it decreases slowly. It allows large variation of f for bad solutions at the beginning and small at the end of the search.
- Cauchy schedule [77] is a faster schedule in which $T_k = \frac{T_0}{k}$. The Cauchy schedule allows a regular decreasing of T_k along the search procedure.
- Exponential [76] schedule, $T_k = T_{k-1} \exp(-K)$. It is the fastest schedule and K is a constant value.

Besides, a commonly implemented strategy for reducing the temperature in SA is the linear approach. This strategy reduces the temperature by the same amount throughout the annealing process: $T_{k+1} = T_k - t$. In which t is a constant which describes the amount of the temperature decrement after each iteration.

4.2   Algorithm design

To solve the MBMS RRM problem, the general SA algorithm is modified and mapped onto the proposed model, named as F2R2M-SA. We choose SA as it is simple to implement and efficient in term of time cost, it statistically guarantees finding a good (enough) solution, and it is adaptable to a variety of problems including telecommunications [69]. The detailed design of F2R2M-SA algorithm is presented in this section.

4.2.1   Algorithm framework

Algorithm 4 illustrates the pseudo code of the proposed F2R2M-SA algorithm. The input parameters are the service and user informations collected in the first phase of model (subsection 2.1.1). The output is the allocation solution for whole cell $x(c)$, which consists the solutions for all flows of all service.

Algorithm 4 F2M2R-SA algorithm

Require: $T(c)$, $D(c)$, $S(c)$, $F(s_i)$, $M(s_i), \forall s_i \in S(c)$
Ensure: $x(c) = \{x_{f_{s_i,j}}, \forall s_i \in S(c), \forall j \in F(s_i)\}$

  $x_0 \leftarrow$ **initialize()**   // initialize solution
  $x(c) \leftarrow x_0, x(c)^* \leftarrow x(c), f^* \leftarrow f(x^*(c))$
  $T_k \leftarrow T_0$ // initialize temperature
  while $T_k >$ threshold **do**
    for 1 **to** M **do**
      $f_{s_i,j} \leftarrow$ **select_flow**$(F(s_i), M(s_i))$ // randomly select one service and flow
      $x_f \leftarrow$ extract $(x(c), f_{s_i,j})$ // the flow solution is a part of the cell solution
      $x'_f \leftarrow$ **random_move**(x_f), $x'_f \in N(x_f)$ // new flow solution
      calculate $F(x'_f) = (Th(x'_f), Po(x'_f))$
      if $Po(x'_f) + \sum Po(x_{\tilde{f}_{s_i,j}}) > P_{mbms\_budget}, \forall \tilde{f}_{s_i,j} \neq f_{s_i,j}$ **then**
        reject x'_f
      end if
      evaluation$(F(x_f), F(x'_f))$ // evaluation is conducted only on the moved flow
      if x'_f is accepted **then**
        update the cell solution $x(c)$ with x'_f
      end if
    end for
    $T_{k+1} \leftarrow$ **cooling**(T_k)
  end while
  return $x(c)$

The algorithm starts by generating an initial solution $x(c)$ for the whole cell. The

initial temperature T_0 is a pre-defined value. At different decreasing temperatures, a series of Metropolis chains is performed. The length of each Metropolis chain is M, i.e. the number of iterations at each temperature T_k. The value of M is set at 100 times the size of the search space for the current flow, i.e. the number of possible channel assignment. This value is expressed as $4^{N_t(s_i)}$, $N_t(s_i)$ is the number of users requesting the service s_i. At each iteration within the Metropolis procedure, a flow is firstly selected by the function $select\_flow()$. Then a neighborhood solution x'_f is randomly generated by applying a move operator to the solution of the current flow, named flow solution x_f. Then the function $Evaluation()$ determines to accept or reject x'_f by comparing it with x_f. Therefore, the process of the SA algorithm is the result of two strategies: random walk and iterative improvement. The first strategy improves slowly the current solution and explores gradually the search space. The second strategy decreases the temperature thus leads the search to converge to a (local) minimum.

4.2.2 Solution initialization

The pseudo code for the temperature initialization is described in algorithm 5. The input parameters are the service and user informations collected in the first phase of model (subsection 2.1.1). The output is the initial solution for the whole cell $x(c)$, which consists the solutions for all flows of all service.

Algorithm 5 Initialize()

Require: $T(c)$, $D(c)$, $S(c)$, $F(s_i)$, $M(s_i), \forall s_i \in S(c)$
Ensure: $x(c) = \{x_{f_{s_i,j}}, \forall s_i \in S(c), \forall j \in F(s_i)\}$
    for each service $s_i \in S(c)$ **do**
        if s_i is nonscalable transmission **then**
            apply S-MPC to $f_{s_i,0}$
        else if s_i is scalable transmission **then**
            apply S-MPC to $f_{s_i,j}$, such that $j > 0$
        end if
    end for
    while $Po(x(c)) > P_{\text{mbms\_budget}}$ **do**
        random select $s_i \in S(c)$
        gradually reduce the transmission coverage by 10% for $f_{s_i,j}$, such that $j = N_f(s_i)$
    end while

As shown in algorithm 5, $x(c)$ is determined by applying the MBMS Power Counting approach (MPC) for each flow, named S-MPC. That is because F2R2M supports the flow-based channel assignment. In S-MPC, firstly, all users are separately assigned to one of the three candidate transport channels, then the channel which consumes the minimum power is selected as the allocation mode in the initial

solution. If the power consuming of the selected (pure) transmission mode is over the MBMS power budget, a service will be randomly selected, and the farthest users for the advanced flow will be gradually rejected until a feasible solution is obtained.

4.2.3 Annealing schedule

As mentioned in section 4.1, a full annealing schedule is defined by the initial temperature T_0, a cooling schedule reducing the temperature, and a stop criteria.

In F2R2M-SA, the method proposed by Van Laarhoven in Equation 4.2 is used to calculated T_0 as in Equation 4.4.

$$T_0 = -\frac{\Delta f_0}{ln(\chi_0)} = -\frac{Po(x') - Po(x) + Th(x') - Th(x)}{ln(\chi_0)} \quad (4.4)$$

Δf_0 is the average increase in the objective function brought by the uphill moves.

Table 4.1 shows the example of T_0 values for 18 scenarios in 10 trials by the method in Equation 4.4. The number of random walk iterations is 100, and the initial probability of accepting a bad solution is 0.7 [48]. In each trial, the random walk procedure is turned n times, the number of users in each scenario. Then the initial temperature for each trial is the rounded integer of the average value: $T_0 = \lceil \frac{1}{n} \sum_{i=1}^{n} T_0^i \rceil$. As shown in Table 4.1, the more the number of users in one scenario, the larger the value of T_0.

For cooling the temperature, we select the logarithmic schedule [42] to decrease the temperature in F2R2M-SA. In which, the temperature in kth step is determined by $T_k = \lceil \frac{T_0}{LOG(k)} \rceil$. The stop criterion is when T_k is below a given threshold.

4.2.4 Select_flow(): flow selection

To determine the flow $f_{s_i,j}$ to modify, two methods are investigated. The first method randomly selects a flow among all N_f flows in a given scenario, N_f is the total number of flows in this scenario. This method randomly selects one service $s_i, i \leq N_s$, and then it randomly selects one flow $f_{s_i,j}$ from $F(s_i)$.

The second method determines $f_{s_i,j}$ in quasi-determined order. Firstly the service s_i is selected in a predefined order (i.e. i is set from 1 to N_s, N_s is the total number of service in one scenario). Then the flow $f_{s_i,j}$ is determined:

- $j = 0$, if s_i is a non-scalable transmission scheme service.
- $j = random(1, N_f(s_i))$, if s_i is a scalable transmission scheme service.

Four scenarios are selected to apply SA with the random flow selection and the

Table 4.1: Values of T_0 with Equation 4.2

Scenarios	T_0 in 10 trials										Average T_0
1s30uN	33	28	29	27	72	13	34	16	29	33	31
1s30uS	51	42	83	31	65	63	28	71	38	63	54
1s60uN	32	98	7	31	47	187	55	17	12	17	51
1s60uS	120	73	104	17	13	132	44	32	122	91	75
1s150uN	344	407	154	86	301	389	196	382	163	155	258
1s150uS	43	106	160	137	299	86	46	191	79	81	123
2s20uSN	22	20	19	12	13	15	5	12	17	16	16
2s20uSS	15	15	7	28	17	18	10	16	23	24	18
2s50uSN	100	53	67	92	75	162	49	45	57	42	75
2s50uSS	112	46	98	48	90	114	35	98	37	138	82
3s30uSNN	88	30	61	37	106	41	37	71	68	64	61
3s30uSSN	27	193	43	20	30	29	46	45	50	85	57
3s50uSNN	35	140	65	74	181	106	55	47	41	45	79
3s50uSSN	98	63	125	32	87	65	75	32	56	61	70
3s80uSNN	55	95	22	54	51	136	106	76	81	163	84
3s80uSSN	114	64	64	58	61	45	78	46	274	79	89
3s100uSNN	167	132	94	85	109	89	71	83	49	50	93
3s100uSSN	75	40	59	36	100	119	65	123	58	92	78

sequential flow selection methods, respectively. Each algorithm is run 50 trials, the best solutions found and average solutions are illustrated in Table 4.2. For each scenario, the best solution is underlined.

Table 4.2: Comparison between two flow selection methods

Scenarios	Method 1: random flow selection		Method 2: sequential flow selection	
	Best solution	Average solution	Best solutions	Average solution
1s60uS	0%, 12.94 W	0%, 13.68 W	<u>0%, 12.82 W</u>	0%, 13.18 W
1s150uS	0%, 15.52 W	0%, 16.07 W	<u>0%, 15.38 W</u>	0%, 15.53 W
2s20uSN	0%, 13.81 W	9.5%, 14.97 W	<u>0%, 7.83 W</u>	0%, 13.18 W
3s50uSSN	5%, 18.87 W	15.19%, 18.26 W	<u>2%, 17.26 W</u>	0%, 18.79 W

Table 4.2 shows that the sequential flow selection method finds better solution than the random flow selection method. That is because the sequential flow selection method provides quasi-arbitrary decision instead of full arbitrary decision in the search procedure. In the F2R2M-SA implementation, the sequential flow selection method is adopted.

4.2.5 Random_move(): solution generation

In each iteration during the SA procedure, for a given flow $f_{s_i,j}$, a new flow solution will be randomly selected from the neighborhood of the current flow solution $x_{f_{s_i,j}}$. Algorithm 6 shows the random generation which randomly moves the current solution to generate the new solution $x'_{f_{s_i,j}}$.

As illustrated in algorithm 6, for the flow solution $x_{f_{s_i,j}}$, according to the UE partition in $x_{f_{s_i,j}}$, a channel ch_o will be firstly randomly determined as "output" set and UE_{ch_o} should not be an empty set. Secondly, a channel ch_i will be randomly chosen as "input" set, $ch_i \neq ch_o$ and ch_i could be an empty user set. Thirdly, a user t_u will be randomly selected from UE_{ch_o}. Then t_u or a set of several users including t_u, constitute the list of users which will be moved from UE_{ch_o} to UE_{ch_i}.

4.2.6 Evaluation(): solution evaluation

Algorithm 7 illustrates the criteria to accept or reject the new flow solution. The comparison is conducted between the current flow solution x_f (i.e. $x_{f_{s_i,j}}$), and the new flow solution x'_f (i.e. $x'_{f_{s_i,j}}$). Each of them should be a part of a feasible cell solution.

Algorithm 6 Random_move()

Require: $x_{f_{s_i,j}}$
Ensure: generate new flow solution $x'_{f_{s_i,j}}$
  service ID $= i$, flow ID $= j$
  randomly choose ch_o from $x_{f_{s_i,j}}$, $UE_{ch_o} \neq \phi$ // output (source) channel set
  randomly choose ch_i, $(ch_i \neq ch_o)$ // input (target) channel set
  randomly choose user t_u from UE_{ch_o}
  if $ch_o \neq$ FACH **and** $ch_i \neq$ FACH **then**
    move t_u from ch_o to ch_i
  else
    if $ch_o =$ FACH **then**
      move user(s) t_v and t_u from FACH to ch_i, $t_v \in UE_{fach}, d_v \geq d_u$ // the FACH coverage is reduced
    else if $ch_i =$ FACH **then**
      move user(s) t_v and t_u from non FACH channels to FACH, $t_v \in UE_{nofach}, d_v \leq d_u$ // the FACH coverage is increased
    end if
  end if

Algorithm 7 Evaluation()

Require: x'_f, x_f
Ensure: accept or reject x'_f
  calculate $Th(x'_f), Po(x'_f), Th(x_f), Po(x_f)$
  if $Th(x'_f) < Th(x_f)$ **or** $(Th(x'_f) = Th(x_f)$ **and** $Po(x'_f) < Po(x_f))$ **then**
    accept x'_f
  else
    $p \leftarrow random(0,1)$
    if $Th(x'_f) > Th(x_f)$ **then**
      $pn = -K_t(Th(x'_f) - Th(x_f))$
    else if $Th(x'_f) = Th(x_f)$ **and** $Po(x'_f) \geq Po(x_f)$ **then**
      $pn = -K_p(Po(x'_f) - Po(x_f))$
    end if
    $p_a \leftarrow e^{\frac{pn}{T_k}}$
    if $p < p_a$ **then**
      accept x'_f // accept x'_f with acceptance probability
    else
      reject x'_f
    end if
  end if

Based on the proposed lexicographic evaluation criteria in subsection 2.2.4, the evaluator firstly compares the throughput loss because the throughout requirement has higher priority than the power consumption. Therefore, the solution with less throughput loss is the better solution regardless its power consumption. Then for solutions having the same throughput loss, the solution with less power is better, so if x'_f has a better fitness, it is accepted, else it will be accepted with the acceptance probability, calculated by Equation 4.5.

$$
p_a = \begin{cases} \exp\left(-K_p \frac{Po(x'_f)-Po(x_f)}{T_k}\right) & \text{if } Th(x'_f) = Th(x_f) \text{ and } Po(x'_f) \geq Po(x_f) \\ \exp\left(-K_t \frac{Th(x'_f)-Th(x_f)}{T_k}\right) & \text{if } Th(x'_f) > Th(x_f) \end{cases}
$$

$$(4.5)$$

The probability factor p_a is a function of the current temperature T_k, the fitness difference, two factors K_p and K_t. In the proposed model, the two-dimensional fitness consists in the bandwidth difference in kilobits per second, and the transmission power in Watts; K_t and K_p are defined to calculate p_a separately. K_t is set larger than K_p ($K_p = 10, K_t = 50$) because the higher constant leads to less acceptance probability, then more restriction in the acceptance of the solution with additional throughput loss.

4.3 Results and comparison with existing approaches

F2R2M-SA is implemented in the simulator proposed in chapter 2. Together with the competing approaches from the state of the art, they are simulated and compared under 18 scenarios. The competing approaches are MBMS Power Counting (MPC), Dual Transmission mode (Dual Tx), and Scalable FACH transmission (S-FACH). Besides, to prove the advantages of the flow based channel allocation, we apply the MBMS power counting to each flow (S-MPC). To observe the power saturation, the solutions of MPC, Dual Tx, S-FACH and S-MPC are determined for their minimum power consumption, i.e. their solution are accepted no matter if the power consumption is over the MBMS power budget.

Table 4.3 shows the algorithm solutions for all scenarios. To apply MPC and Dual Tx, the services in all the scenarios use non-scalable transmission scheme. Therefore the scenarios have the same service and user distributions share the same solution, e.g. *1s30uN* and *1s30uS*. S-FACH allocates the FACH for each flow with a fixed coverage [31]: 95% for f_1, or f_0 of nonscalable encoded service, 50% for f_2 and 33% for f_3 (if the service has f_3).

In Table 4.3, the solutions are presented as two-dimensional fitness value: the lost of throughout in percentage and the consumed power in Watts for all MBMS multicast services within one cell. For example for scenario *1s30uS*, S-FACH applies

Table 4.3: Best solutions found by different MBMS RRM algorithms

Scenarios	MPC	Dual Tx	S-FACH	S-MPC	F2R2M-SA
1s30uN	0%,	0%,	14%, 10.8 W	0%, 13.77 W	0%, 7.02 W
1s30uS	13.77 W	13.77 W	25%, 9.45 W	0%, 10.56 W	0%, 8.41 W
1s60uN	0%,	0%	29.2%, 10.8 W	0%, 16.68 W	0%, 11.62 W
1s60uS	16.68 W	16.68 W	29.2%, 10.8 W	0%, 17.16 W	0%, 12.82 W
1s150uN	0%,	0%,	29.3%, 10.8 W	0%, 16.68 W	0%, 14.65 W
1s150uS	16.68 W	16.68 W	29.3%, 10.8 W	0%, 17.16 W	0%, 15.38 W
2s20uSN	0%,	0%,	81.25%, 5.82 W	0%, 6.82 W	0%, 7.8291 W
2s20uSS	17.58 W	19.32 W	45.3%, 15.3 W	0%, 6.18 W	0%, 8.80 W
2s50uSN	0%,	0%,	65%, 10.23W	28%, 21.51 W	0%, 11.12 W
2s50uSS	27.19 W	30.45 W	47%, 15.4 W	16%, 18.4 W	0%, 13.06 W
3s30uSNN	0%,	0%,	85%, 5.82 W	0%, 8.17 W	0%, 8.72 W
3s30uSSN	8.93 W	22.92 W	57%, 15.3	0%, 7.53 W	0%, 10.97 W
3s50uSNN	0%,	0%,	24.3%, 12.7 W	0%, 18.8 W	0%, 17.32 W
3s50uSSN	18.08 W	32.16 W	41.4%, 15.8 W	24.3%, 12.9 W	0%, 18.79 W
3s80uSNN	0%,	0%,	23.6%, 26.9 W	41.4%, 21.5 W	0%, 14.99 W
3s80uSSN	32.47 W	37.68 W	22.6%, 22.6 W	32.9%, 18.4 W	0%, 14.36 W
3s100uSNN	0%,	0%,	89.3%, 5.9 W	0%, 31.79 W	11.8%, 18.96 W
3s100uSSN	35.73 W	37.68 W	62.1%, 15.4 W	68.8%, 12.1 W	15.9%, 18.8 W

the FACH with 95% coverage to transmit the first flow $f(s_1, 1)$ and 50% coverage for the second flow $f(s_1, 2)$; hence it loses 576 kbps required bandwidth ($576/3840 = 15\%$), its solution fitness is (15%, 10.8 W). The feasible solutions ($Po \leq 19W$) are emphasized in boldface.

Table 4.3 shows that, when the service transmission is a non-scalable mode (i.e. MPC and Dual Tx), 6 scenarios can be transmitted through feasible solution with MPC, while the Dual Tx obtains feasible solutions for 4 scenarios. Dual Tx costs more power than MPC since the former does not consider using HS-DSCH. Such inefficiency is confirmed in the solution allocation for *3s80uSNN* (Table 4.4), where MPC consumes less power than Dual Tx because the users of s_2 and s_3 receive the services through HS-DSCH.

S-FACH saves the power consumption by reducing the coverage for the advanced flow(s), it also guarantees the service coverage (all users can be covered). However, when the power is not saturated, such throughout sacrifice in S-FACH is unnecessary. For scenarios having 1 service and *2s20uSN/SS*, both S-FACH and MPC could obtain feasible solutions. S-FACH costs less power than MPC but loses more bandwidth due to the smaller coverage for the advanced flow(s). According to the lexicographic evaluation criteria, MPC is still better than S-FACH. Moreover, when the service demand is higher (i.e. *3s80uSNN/SSN*), with the fixed flow coverage, S-FACH is not flexible in terms of trade-off between the service quality and power consumption. Therefore, for the last eight scenarios, S-FACH still achieves power saturation although certain throughput has been lost. Such power saturation actually could be avoided by (additionally) decreasing the user coverage from advanced flow(s). We can conclude that MPC is more efficient than S-FACH.

However, MPC does not consider the multimedia scalability, therefore, it always achieve saturated transmission power for scenarios with heavier traffic load, i.e. *2s50uSN/SS* and *3s80uSNN/SSN*. By comparing the MPC and S-MPC solutions, we find that S-MPC costs less power than MPC, hence S-MPC achieves more feasible solutions; in particular, for most of the scenarios (expect *3s50uSSN* and *3s100u*), S-MPC obtains the same throughput requirement (0% throughput loss) and less power consumption than MPC. Therefore, the comparison between MPC and S-MPC reveals the advantage of scalable transmission. Besides, we observe the S-MPC solutions for the scenarios with the same user distribution and total traffic load, i.e. *2s20uSN* vs *2s20uSS*, *2s50uSN* vs *2s50uSS*, *3s30uSNN* vs *3s30uSSN*, *3s80uSNN* vs *3s80uSSN* and *3s100uSNN* vs *3s100uSSN*, we can find that the scenarios with scalable transmission of s_2 consume less power. This also proves the advantage of scalable transmission. Among these scenarios, however, when the users are more than than 20, e.g. *2s50u*, *3s50u*, *3s80u* and *3s100u*, the S-MPC solutions have throughout loss more than 20%. That is because S-MPC only allocates pure transmission mode for each flow, and when DCH users are numerous, S-MPC causes expensive channel code consumption, that increases the possibility of channel codes

saturation, and leads to high throughput loss.

Table 4.4: Detailed solutions of *3s80uSNN*

Algorithms	flow	Number of users in UE sets			
		UE_{fach}	UE_{dch}	UE_{hs}	UE_{noch}
MPC	$f_{s_1,0}$	30	0	0	0
	$f_{s_2,0}$	0	0	20	0
	$f_{s_3,0}$	0	0	30	0
Dual Tx	$f_{s_1,0}$	30	0	0	0
	$f_{s_2,0}$	20	0	0	0
	$f_{s_3,0}$	30	0	0	0
Scalable FACH	$f_{s_1,1}$	30	0	0	0
	$f_{s_1,2}$	14	0	0	16
	$f_{s_1,3}$	5	0	0	25
	$f_{s_2,0}$	20	0	0	0
	$f_{s_3,0}$	30	0	0	0
Scalable MPC	$f_{s_1,1-3}$	0	0	30	0
	$f_{s_2,0}$	0	0	20	0
	$f_{s_3,0}$	0	0	30	0
F2R2M-SA	$f_{s_1,1-3}$	0	20	10	0
	$f_{s_2,0}$	0	10	10	0
	$f_{s_3,0}$	0	18	12	0

The detailed allocation of different algorithms for *3s80uSNN* is illustrated in Table 4.4, which shows that F2R2M-SA avoids the saturation of channel codes by applying combinational channel assignment of DCH and HS-DSCH for the advanced flow $f_{s_1,2-3}$.

In conclusion, F2R2M-SA outperforms the other algorithms. For small size scenarios, when the conventional approaches can allocate the radio resources properly, F2R2M-SA consumes less power (47% of MPC solution) by coordinating the throughput request of the flows and the channel allocation. For large size scenarios, F2R2M-SA avoids the unnecessary throughput loss by more flexibility on the allocation of the users to each flow, which allows the algorithm to serve more users for the advanced flow.

4.4 Synthesis

MBMS RRM problem has been proved to be NP-Hard, which cannot be tackled by exhaustive approach within practical computation time. The selection of meta-heuristic is reasonable, since it requires reasonable amount of effort to get a good solution. As a popular and efficient metaheuristic, the Simulated Annealing (SA) has been studied to solve the target problem.

In this chapter, the concept of general SA is firstly introduced. Then a modified SA algorithm, named F2R2M-SA is proposed to solve the MBMS RRM problem. In the proposed algorithm, the MBMS power counting approach is used to generate the initial solution. The temperature initialization is applied by a predefined sampling procedure. For the cooling schedule, the logarithmic cooling schedule is used to decrease the temperature. At each temperature, the Metropolis chain is implemented with $100 \times 4^{N_t(s_i)}$ iterations, $N_t(s_i)$ is the number of users receiving the flow $f_{s_i,j}$. At each iteration, one application of the neighborhood operator, named the multiple insert operator, is applied to the current flow solution to randomly generate a new flow solution. When the new solution offers a non-improving fitness (according to the lexicographic evaluation method), an acceptance probability is calculated by the difference of solution fitness and two factors K_p and K_t, which are related to the two-dimensional fitness value.

Together with the existing MBMS RRM algorithms, the proposed F2R2M-SA is implemented and tested in the same simulation environment. The algorithm performance is evaluated. Simulation results show that, F2R2M-SA outperforms the other UMTS existing algorithms: MPC, S-FACH and Dual Tx Mode. The proposed model provides the best trade off between the power consumption and the service quality by applying a flow based channel allocation, and reduces the possibility of radio resource saturation by adapting the combinational channel assignment.

Solving MBMS RRM Problem by Tabu Search

In this chapter, a Tabu Search (TS) algorithm is investigated for solving the MBMS RRM problem. The general TS algorithm is adapted and mapped on the F2R2M, named F2R2M-TS. Comparing with F2R2M-SA which is proposed in chapter 4, F2R2M-TS can avoid revisiting previous solutions by keeping memory of previous search steps, which helps to improve the search efficiency. For algorithm design, three tabu memory structures are proposed based on the model characteristics. Their performance are compared and evaluated. Furthermore, a tabu repair mechanism is proposed as extension of the classic TS procedure. This mechanism improves the search efficiency by self-adapting tabu moves during the search iterations. The proposed F2R2M-TS is compared with F2R2M-SA and Greedy Local Search (GLS). Simulation results show that F2R2M-TS outperforms the other two algorithms. The best solutions found by F2R2M-TS are also compared with the theoretical solution bounds. Results show that the solutions found by F2R2M-TS are very close or equal to the theoretical optimum solutions.

Contents

5.1   Introduction of tabu search algorithm

In chapter 3 and chapter 4, the GLS and F2R2M-SA algorithms have been proposed to solve the MBMS RRM problem. The simulations show that they have already outperformed the existing UMTS approaches. However, both of the algorithms have the shortcoming that they have difficulties to avoid local optima, making that the best solution found so far is not always found by the algorithms. The Tabu Search (TS) algorithm is studied to overcome this inconvenience.

Tabu search is a metaheuristic search method, originally proposed by Glover in 1986 [43]. This algorithm guides a local search procedure to explore the solution space beyond local optimality by using memory structures to avoid revisiting previous solutions. In the past decade, several applications of TS have been presented in the field of wireless networks: frequency and channel assignment [63, 59], node planning and network topology optimization [35], routing optimization [57], etc. This makes TS also an attractive candidate for solving MBMS RRM problem. The basic concept of TS is described in this section.

Tabu search enhances the performance of the classical local search method. It prevents cycling back to previously visited solutions by using memory structure, called tabu list L. The tabu list records the recent search history, describing the recently visited solutions or pre-defined sets of rules. If a potential solution has been previously visited within a certain period, it is marked as "tabu" (forbidden) so that TS does not consider this solution repeatedly. This memory structure aims to overcome local optima. Another important element in TS is the neighborhood structure. At each iteration, TS starts from the current solution x. The transformations that can be applied to x define a set of neighboring solutions, denoted $N(x)$. $N(x)$ is named the neighborhood of x. For any given solution, there are many possible neighborhood structures defined by different optimization operators.

Algorithm 8 General TS algorithm [44]

    select an initial solution x_0
    $x \leftarrow x_0, f^* \leftarrow f(x_0), x^* \leftarrow x_0, L \leftarrow \emptyset$ // tabu list is empty
    repeat
        select $x\prime$ in $\arg\min_{x\prime \in N(x) \backslash L}[f(x\prime)]$ // the best solution among non tabu
        neighboring solutions
        update the tabu list L, $L \leftarrow L \cup \{x\}$
        $x \leftarrow x\prime$
        if $f(x) < f^*$ **then**
            $f^* \leftarrow f(x), x^* \leftarrow x$
        end if
    until stopping criterion is met

Algorithm 8 illustrates the general framework of the most commonly used version of TS. Where x denotes the current solution, x^* is the best-known solution during the TS search procedure. f^* is the fitness value of x^*. $N(x)$ is the neighborhood of x. At each iteration, the best neighboring solution $x\prime$ is selected from all the neighboring solutions of x. The selected solution should not appear in the tabu list. Once a new solution $x\prime$ is selected, the tabu list is updated with the current solution and the new solution replaces the current solution. The solution recorded in the tabu list is named tabu. The number of iterations that one tabu stays in the tabu list is named tabu tenure. The best-known solution x^* is updated every time the current solution is better than the best solution ever found. When the stopping criterion is satisfied, x^* is returned as the result of TS algorithm.

5.2   Algorithm design

To better get rid of local optimum and improve the search efficiency, the general TS algorithm is modified and mapped onto the proposed F2R2M model, named F2R2M-TS. This section presents the detailed design of F2R2M-TS algorithm. The top level algorithm framework is described in Algorithm 9.

Algorithm 9 F2R2M-TS algorithm

Require: $T(c)$, $D(c)$, $S(c)$, $F(s_i)$, $M(s_i)$, $\forall s_i \in S(c)$
Ensure: $x(c) = \{x_{f_{s_i,j}}, s_i \in S(c), j \in F(s_i)\}$
  randomly generate x_0
  $x(c) \leftarrow x_0, x(c)^* \leftarrow x_0, f^* \leftarrow f(x^*(c)), L \leftarrow \emptyset$
  while termination criterion not satisfied **do**
    for all $s_i \in S(c)$ **do**
      for all $f_j \in F(s_i)$ **do**
        $x_{f_j} \leftarrow x_{f_{s_i,j}}$ // sequentially select one flow to modify its assignment
        select x'_{f_j} in $\arg\min_{x'_{f_j} \in N(x_{f_j}) \setminus L}[(F(x'_{f_j})]$, $F(x'_{f_j}) = (Th(x'_{f_j}), Po(x'_{f_j}))$
        // the best neighborhood non tabu assignment for flow $f_{s_i,j}$
        update tabu list L with the move from x'_{f_j} to x_{f_j}, $L \leftarrow L \cup \{x'_{f_j} \rightarrow x_{f_j}\}$
        update current flow solution $x_{f_j} \leftarrow x'_{f_j}$
        update current cell solution $x(c)$ with x_{f_j}
        if $Th(x(c)) < Th(x^*(c))$
        or $(Th(x(c)) = Th(x^*(c))$ **and** $Po(x(c)) < Po(x^*(c)))$ **then**
          $x^*(c) \leftarrow x(c), f^* \leftarrow f(x^*(c))$
        end if
      end for
    end for
  end while

Algorithm 9 describes the F2R2M-TS procedure. $f(x)$ is the fitness value of

the solution x, it includes the throughput loss and the transmission power $f(x) = (Th(x), Po(x))$. To optimize the fitness value of the cell solution $f(x(c))$, F2R2M-TS begins from an initial solution x_0. At each iteration, a given flow f_j is sequentially selected. The partitioning of UE for the selected flow, x_{f_j}, is then explored. The associated neighborhood of x_{f_j}, named $N(x_{f_j})$ is generated. Each solution in $N(x_{f_j})$ could be reached from x_{f_j} by an operator named move. The best neighborhood solution x'_{f_j} is then generated from the feasible and admissible subset of $N(x_{f_j})$. The tabu list, L, is updated with the reversal move which transfers x_{f_j} to x'_{f_j}. The tabu list, L, records the active tabu moves (forbidden moves) for a certain number of iterations (i.e. tabu tenure). This process is repeated until satisfying the termination criterion.

In Algorithm 9, a key step is to determine the best flow solution x'_{f_j} from the admissible non tabu neighborhood of the current flow solution $N(x_{f_j})$, that is to generate x'_{f_j} from $N(x_{f_j})$. Figure 5.1 illustrates the selection of the best neighboring solution x'_{f_j} in $\arg\min_{x'_{f_j} \in N(x_{f_j}) \backslash L}$, and the update of the tabu list L.
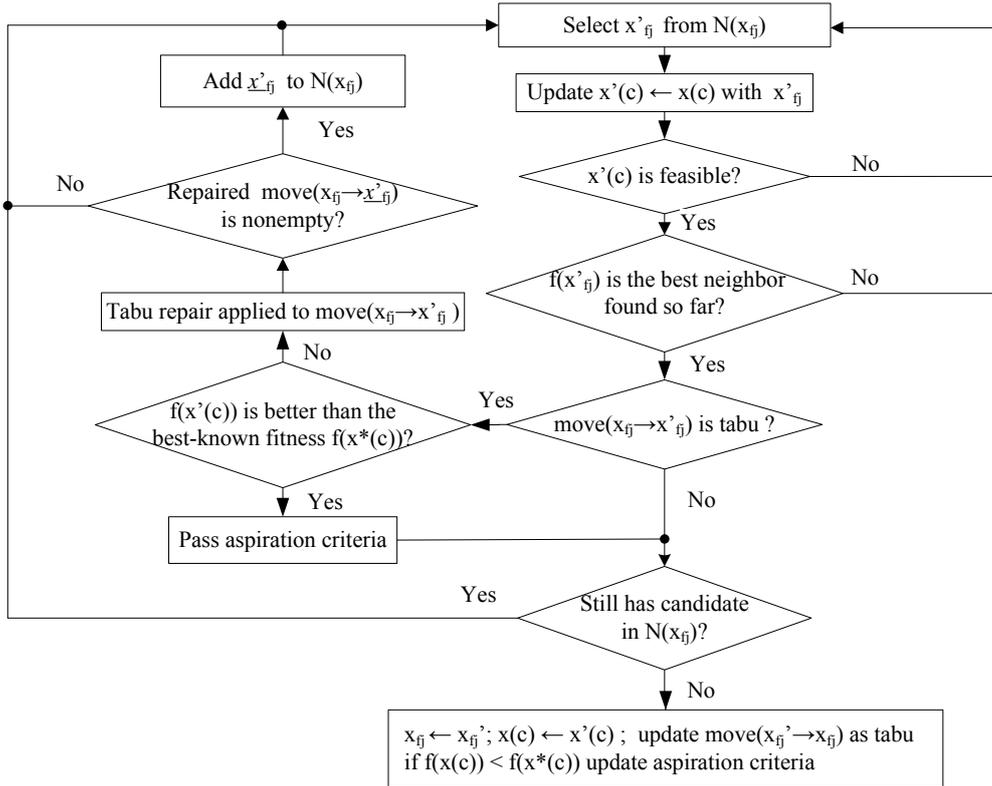

Figure 5.1: Best neighborhood selection in F2R2M-TS

As shown in Figure 5.1, at each iteration, for each candidate neighborhood solution x'_{f_j}, its feasibility and its admissibility is verified. x'_{f_j} is called "admissible"

when it is not classified as tabu or is a tabu move but pass the aspiration criteria. For feasibility control, the new candidate neighborhood solution x'_{f_j} will firstly replace the current flow solution x_{f_j} within the current cell solution $x(c)$. Then the power consumption of the obtained new candidate cell solution, $x'(c)$, is calculated. When $Po(x'(c))$ is less than the maximum power budget P_{mbms_budget}, $x'(c)$ (as well as x'_{f_j}) is called "feasible". It is necessary that an admissible solution should first be a feasible solution.

As shown in Figure 5.1, if x'_{f_j} is tabu, the fitness of the solution x'_{f_j} will be firstly calculated, and then the fitness of the other flow solutions $x'_{f_k}, k \neq j$, hence the aggregated fitness value $f(x'(c))$ is obtained. $f(x'(c))$ will then be compared with the best-known fitness $f(x^*(c))$. If $f(x'_c)$ is better than $f(x^*_c)$, x'_{f_j} will be accepted even if it is tabu. This procedure is named aspiration criterion, which overrides a solution tabu state when it obtains a fitness value which is better than the best ever known flow solution. Then the best known solution and the tabu list will be updated. When $x'(c)$ does not satisfy the aspiration rule, a tabu repair mechanism is proposed to extend the classic tabu search. This mechanism deletes the tabu attributes from the tabu-declared move that transfers x_{f_j} to x'_{f_j}, which eventually leads to a new admissible solution $\underline{x}'_{f_j}$. $\underline{x}'_{f_j}$ is then be packed in the candidate solution lists. The aspiration criteria and the repaired tabu mechanism are important elements that introduce flexibility in the solution exploration process of F2R2M-TS.

5.2.1 Move operator and neighborhood generation

In the study of optimization strategies in chapter 3, two move operators are designed: the Single Insert (SI) operator δ_{SI} (one user moves at a time) and the Multiple Insert (MI) operator δ_{MI} (several users move at a time). It is showed that δ_{MI} can easily escape from local optimal solutions while δ_{SI} could not escape from the initial solution space. Hence it is harder for δ_{SI} to get rid of local optimum than δ_{MI}.

Hereby the tabu search neighborhood is considered based on δ_{MI}. As introduced in section 2.2, a MI move operation is implemented in five steps:

1. Choose a flow $f_{s_i,j}$ to change its current allocation solution $x_{f_{s_i,j}}$.

2. From the UE partition in $x_{f_{s_i,j}}$, choose a channel category $iniCh$ as a "source" set, $UE_{iniCh} \neq \emptyset$.

3. From the UE partition in $x_{f_{s_i,j}}$, select a channel category $tarCh$ as a "desination" set, $tarCh \neq iniCh$.

4. Select one user t_k from UE_{iniCh}.

5. According to $iniCh$ and $tarCh$, move t_k or a block of users including t_k from UE_{iniCh} to UE_{tarCh}.

Based on these five steps, the move structure is designed, including five variables, shown in Table 5.1.

Table 5.1: Move structure

Variable	Note	Value
sId	service ID	$1, \ldots, N_s$
fId	flow ID	$1, \ldots, N_f(sId)$
iniCh	channel ID of source user set	$0, 1, 2, 3$
tarCh	channel ID of destination user set	$0, 1, 2, 3$
uList	list of users to be moved	$1, \ldots, N_t(sId)$

In the move structure, integer values are used to represent the channel ID, hence $0, 1, 2, 3$ represent the channels FACH, DCH, HS-DSCH and no transmission (NOCH). $uList$ includes the moved users ID, and depends on the values of $iniCh$ and $tarCh$. When the move concerns the source channel is ch_1 (DCH) or ch_2 (HS-DSCH), the δ_{MI} operator moves only one user. While for the source channel ch_0 (FACH), the moved users depends on the changed FACH coverage. For example, if the user t_k is moved from UE_{hs} to UE_{fach}, the FACH coverage is extended to reach t_k and all the users nearer than t_k can now hear from FACH transmission and should be (if currently they are not in) moved to FACH user set. Therefore, when t_k is chosen to be inserted into UE_{fach}, F2R2M-TS will first check the distribution of users served by the other channels, and then pick out the users within the enlarged coverage to UE_{fach}. At the contrary, if the source channel $iniCh$ is FACH, the move will leads to reducing the FACH coverage and all users farther than t_k within UE_{fach} should be inserted to UE_{tarCh}.

$$x_1 : \left(\underbrace{-1\ 1\ \underline{2\ 3\ 4\ 5}}\ \underbrace{-1\ 7\ 8}\ \underbrace{-1\ 6\ 9\ 10}\ \underbrace{-1\ 11} \right) \tag{5.1a}$$

$$x_2 : \left(\underbrace{-1\ 1}\ \underbrace{-1\ 7\ 8}\ \underbrace{-1\ \underline{2\ 3\ 4\ 5}\ 6\ 9\ 10}\ \underbrace{-1\ 11} \right) \tag{5.1b}$$

$$x_3 : \left(\underbrace{-1\ 1}\ \underbrace{-1\ \underline{2}\ 7\ 8}\ \underbrace{-1\ 3\ 4\ 5\ 6\ 9\ 10}\ \underbrace{-1\ 11} \right) \tag{5.1c}$$

$$x_4 : \left(\underbrace{-1\ 1}\ \underbrace{-1\ 2\ \underline{3}\ 7\ 8}\ \underbrace{-1\ 4\ 5\ 6\ 9\ 10}\ \underbrace{-1\ 11} \right) \tag{5.1d}$$

$$x_5 : \left(\underbrace{-1\ 1}\ \underbrace{-1\ 2\ 3\ \underline{4}\ 7\ 8}\ \underbrace{-1\ 5\ 6\ 9\ 10}\ \underbrace{-1\ 11} \right) \tag{5.1e}$$

In Equation 5.1, five examples of δ_{MI} operators are conducted. This example

shows a case where the cell includes only one service with one flow and the solution is changed from x_1 to x_2 and so on until x_5. According to the solution representation introduced in chapter 3, solutions are represented as a vector of four user sets: UE_{fach}, UE_{dch}, UE_{hs} and UE_{noch}, these user subsets are delimited by -1. In each subset, users are ordered in ascending order of their distances from the base station.

Table 5.2 describes the first move, m_1, modifying the solution x_1 to x_2.

Table 5.2: A move example m_1: from x_1 to x_2

Variable	sId	fId	iniCh	tarCh	uList
Value	1	1	0	2	2, 3, 4, 5

The users t_2, t_3, t_4 and t_5 are moved from the FACH channel to a HS-DSCH channel with $t_k = t_2$. The coverage threshold d_{thr} is set to d_2. t_1 is kept within the FACH coverage because it is the nearest user from the base station. Whereas if $d_{thr} < d_1$, the FACH coverage would be further reduced, bringing out the t_1 from the FACH coverage. Then from x_3 to x_5, one user is selected to move from the HS-DSCH to the DCH channel.

Algorithm 10 shows the neighborhood generation of solution $x_{f_{s_i,j}}$, based on δ_{MI}.

5.2.2 Design of tabu memory structures

The tabu list is a memory structure dynamically modified during the search procedure. The tabu search records the characteristics of previous moves. Once a solution is updated, the modified attributes by the current move will be set tabu active, and recorded in the tabu list. In the following iterations, the tabu evaluation procedure compares the attributes of a potential move operation with the tabu attributes stored in the tabu list. The move is considered as tabu if a matching is found. The moves declared as tabu will be ignored except if it satisfies the aspiration criterion. By doing this, re-visiting of recent solutions through non-improving moves are prevented. In TS heuristic, tabu memory structures plays an important role in the algorithm performances. Based on the move operator, three tabu structures are defined. As shown in Table 5.3, these move based tabu structures include three, four and five variables, respectively.

As shown in Table 5.3, the reverse move of the selected one is considered as tabu. In addition to definitions in Table 5.1, uId refers to the user that recently moved in a recent iteration. *tabu-tri* means that the current transmission channel of the user uId on the flow fId belonging to the service sId is tabu-active (should temporarily not be modified). *tarCh* represents the target channel for uId, therefore

---

Algorithm 10 Neighborhood_generate()

Require: $x_{f_{s_i,j}}, M(s_i), i \in S(c), j \in N_f(s_i)$
Ensure: $N(x_{f_{s_i,j}})$
  for all user $t \in M(s_i)$ **do**
    $UE_{nofach} = \cup_{ch \neq fach} UE_{ch}(f_{s_i,j})$
    if $t \in UE_{fach}(f_{s_i,j})$ **then**
      for all $tar = 1$ or 2 or 3 **do**
        new move: $sId = i$; $fId = j$; $iniCh = 0$; $tarCh = tar$
        $uList = \{v \in UE_{fach}(f_{s_i,j}), d_v < d_t\}$
      end for
    end if
    if $t \in UE_{dch}(f_{s_i,j})$ **then**
      for all $tar = 0$ or 2 or 3 **do**
        new move: $sId = i$; $fId = j$; $iniCh = 1$; $tarCh = tar$
        if $tar = 0$ **then**
          $uList = \{v \in UE_{nofach}, d_v \leq d_t\}$
        else
          $uList = t$
        end if
      end for
    end if
    if $t \in UE_{hs}(f_{s_i,j})$ **then**
      for all $tar = 0$ or 1 or 3 **do**
        new move: $sId = i$; $fId = j$; $iniCh = 2$; $tarCh = tar$
        if $tar = 0$ **then**
          $uList = \{v \in UE_{nofach}, d_v \leq d_t\}$
        else
          $uList = t$
        end if
      end for
    end if
    if $t \in UE_{noch}(f_{s_i,j})$ **then**
      for all $tar = 0$ or 1 or 2 **do**
        new move: $sId = i$; $fId = j$; $iniCh = 3$; $tarCh = tar$
        if $tar = 0$ **then**
          $uList = \{v \in UE_{noch}(f_{s_i,j}), d_v \leq d_t\}$
        else
          $uList = t$
        end if
      end for
    end if
  end for

Table 5.3: Three tabu structure definitions

tabu-tri	tabu-qua	tabu-pen
sId	sId	sId
fId	fId	fId
uId	uId	uId
	tarCh	tarCh
		iniCh

tabu-qua stipulates that a given destination channel for *uId* is tabu-active, which means that the reallocation of a given type of channel to the user is not allowed for this flow. Finally, *tabu-pen* sets that the transfer of user *uId* from *iniCh* to *tarCh* is tabu-active. *tabu-tri* is the most restrictive tabu structure.

Following the example in Equation 5.1, m_1 transfers x_1 to x_2. Because x_2 is the best neighbor, it replaces x_1, and the reverse move of m_1 is recorded as tabu, as shown in Table 5.4. The *tarCh* and *iniCh* in tabu elements are, respectively, the initial channel and the target channel in the move operator.

Table 5.4: Tabus states (type *tabu-pen*) based on move m_1

Variables	tabu1	tabu2	tabu3	tabu 4
sId	1	1	1	1
fId	1	1	1	1
uId	2	3	4	5
iniCh	2	2	2	2
tarCh	0	0	0	0

5.2.2.1 Tabu evaluation

During the search iterations, the potential moves are evaluated by comparing the move attributes with the tabu elements stored in the tabu list. A move is declared as tabu when a matching is found. Only one type of tabu structure is applied in each TS procedure. Assume that *tabu-pen* is decided to evaluate the tabu moves, a move for x_5 cannot put the users t_2 or t_3 from the *iniCh* 2 to the *tarCh* 0 as those moves are tabu.

Table 5.5: A tabu move for x_5

Variable	sId	fId	iniCh	tarCh	uList
Value	1	1	2 (HS-DSCH)	0 (FACH)	2, 3

5.2.3   Adaptive tabu tenure

The number of iterations that an attribute remains tabu-active is called its tabu tenure. In general it could be fixed as predefined value or dynamically calculated. Several studies [34, 71] have shown that a dynamic tabu tenure could be more effective than a static value. Here, the tabu tenure varies for different tabu attributes and different problem sizes, which provides a dynamic and robust form of search [44]. In this work, the dynamic tabu tenure is randomly selected from the interval $[\frac{\sqrt{N}}{2}, \frac{3\sqrt{N}}{2}]$, where the value of N depends on the type of tabu attribute. Assume that an instance has N_s services, each has $N_f(s_i) = N_f$ flows and $N_t(s_i) = N_t$ users. In the case of *tabu-pen* structure there are $N = 12 \cdot N_s \cdot N_f \cdot N_t$ possible moves (channel modification) from a current solution. Similarly, for *tabu-qua*, N is the number of possible target channel setting for the whole problem, hence $N = 3 \cdot N_s \cdot N_f \cdot N_t$. Finally, for *tabu-tri*, $N = N_s \cdot N_f \cdot N_t$. Therefore, the more restrictive the tabu structure is, the shorter the tabu tenure, and consequentially the shorter the tabu list.

5.2.4   Tabu repair mechanism

To explore new candidate solutions, the tabu repair mechanism is proposed based on the definitions of tabu and the move structure. It is done by re-allowing a tabu-declared move by modifying a sub-set of its attributes. In this application, the user attribute is modified by the tabu repair mechanism.

Figure 5.2: Tabu repair mechanism: an example

As illustrated in Figure 5.2, a tabu move is selected to be applied to x_5: move the users t_5, t_6, t_9 and t_{10} from HS-DSCH to FACH. The selected move is tabu

because it includes one tabu element, i.e. users from HS-DSCH to FACH. To repair this move, the user t_5 is deleted from *uList*, then the repaired move is added to the candidate list. The tabu repair mechanism improves the traditional TS search in such a way that, it not only prevents re-visiting a previous non-improving move, but also introduces the self-adaptation mechanism for the current moves during the search iterations. Therefore it could further improve the search efficiency.

5.3   Experiments and results analysis

In this section, simulations are conducted to evaluate the proposed F2R2M-TS algorithm. First, all the six combinations of tabu search strategies (three different tabu memory structures, with and without tabu repair mechanism) are compared, and the best strategy is selected. Then, F2R2M-TS is compared with the other two metaheuristic algorithms, in terms of both the solution quality and the computation time. Finally, F2R2M-TS performance is compared with the theoretical solution bound generated by the multiple-choice knapsack problem based modeling. The results show that F2R2M-TS can find optimum or near-optimal solutions for MBMS RRM problem.

5.3.1   Analysis of tabu search strategies

The performance of the different TS strategies are conducted with two scenarios: *3s50uSSN* and *3s80uSSN*. We define a metric called rank value to compare the best solution found of different approaches. The ranked values are obtained by comparing the power and throughput loss of different best solutions found in lexicographic order: x' is evaluated as a better solution than x when $Th(x') = Th(x)$ and $Po(x') \leq Po(x)$, or $Th(x') < Th(x)$. For each scenario, each TS method is applied 1000 trials. 6000 solutions are then collected together, then compared and listed in descending order of quality. Each solution obtains a rank value corresponding to its index, indicating its quality level in the solutions population. The lower the rank is, the better the quality.

Figure 5.3 draws the curves of solutions ranks obtained by each TS approaches. The x-axis orders the 1000 solutions obtained by one TS approach from the best to the worst. The y-axis shows the solutions ranking among the 6000 final solutions, where the lower ranks correspond to the best solutions. In Figure 5.3, the red, black and blue lines are used to represent the TS solutions applying *tabu-tri*, *tabu-qua* and *tabu-pen* tabu structure. The dash lines are solutions of TS without Tabu Repair Mechanism (TRM), while the solid lines are solutions applying TRM.

According to the results shown in Figure 5.3, *tabu-pen* provides the best solu-
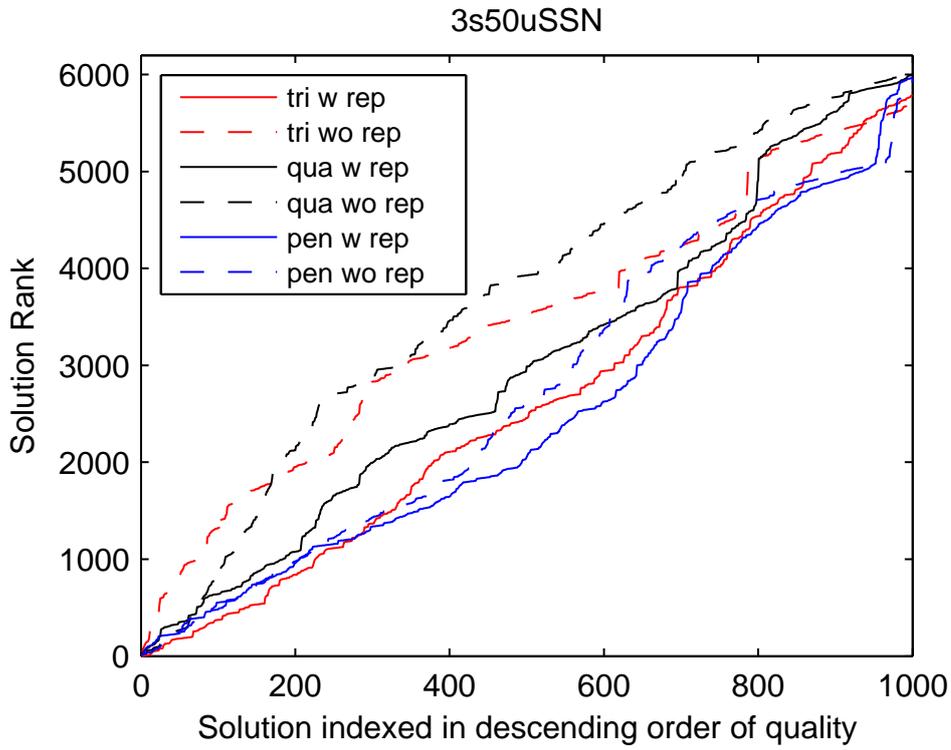
3s50uSSN

(a) 3s50uSSN
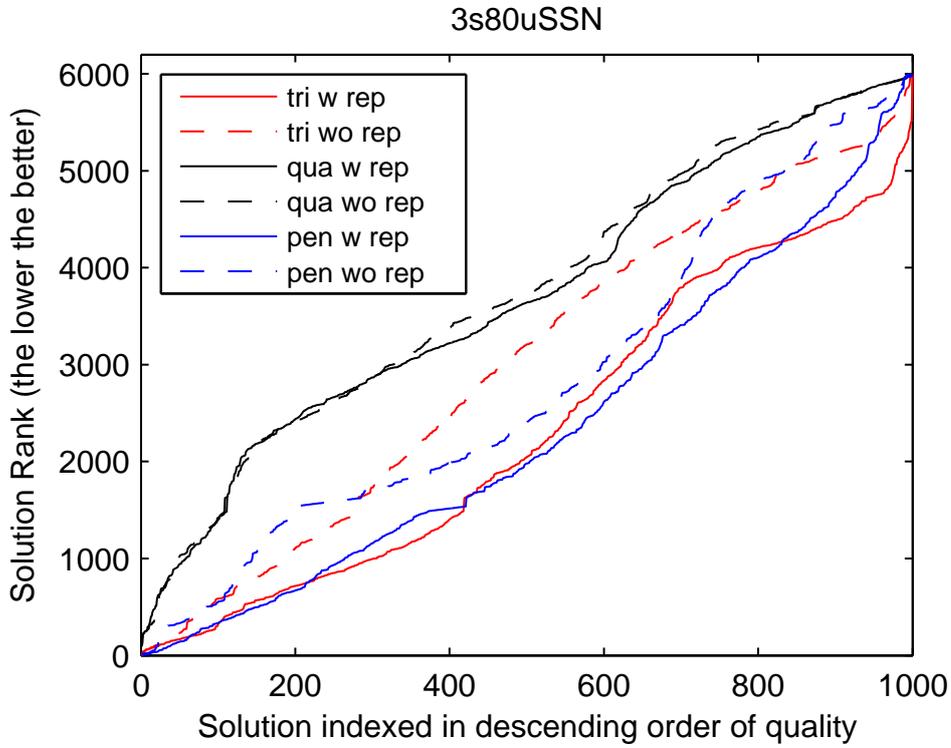
3s80uSSN

(b) 3s80uSSN

Figure 5.3: TS strategies comparison

tions set among the three tabu structures and *tabu-qua* provides the worse solutions quality. The comparison of the six TS strategies according to the tabu repair mechanism shows that this mechanism significantly improves the solution whatever the tabu structure.

Table 5.6 and Table 5.7 illustrate the best solutions found (among 150 trials) of tabu search with and without tabu repair mechanism. The solution fitness are represented in two dimensions: $Th(x)$ in percentage and $Po(x)$ in Watts. The maximum power budget is set to 19 W (solutions with $Th(x) < 19$ are feasible) and all solutions are feasible. The solutions found by TRM approach (Table 5.6) consume less power than without TRM in Table 5.7, which confirms the conclusions of Figure 5.3(a) and Figure 5.3(b).

Table 5.6: TS solutions without tabu repair mechanism

Scenarios	tabu-tri	tabu-qua	tabu-pen
1s60uS		0%, 12.46 W	
1s150uS		0%, 15.38 W	
2s50uSN	0%, 11.16 W	0%, 11.27 W	0%, 11.12 W
2s50uSS	0%, 10.71 W	0%, 10.74 W	0%, 10.70 W
3s50uSNN	0%, 9.27 W	0%, 9.50 W	0%, 9.27 W
3s50uSSN	0%, 9.27 W	0%, 9.53 W	0%, 9.27 W
3s80uSNN	0%, 14.07 W	0%, 14.46 W	0%, 14.06 W
3s80uSSN	0%, 13.45 W	0%, 13.59 W	0%, 13.50 W

Table 5.7: TS solutions with tabu repair mechanism

Scenarios	tabu-tri	tabu-qua	tabu-pen
1s60uS		0%, 12.46 W	
1s150uS		0%, 15.38 W	
2s50uSN	0%, 11.12 W	0%, 11.17 W	0%, 11.12 W
2s50uSS	0%, 10.70 W	0%, 10.73 W	0%, 10.54 W
3s50uSNN	0%, 9.27 W	0%, 9.5 W	0%, 9.27 W
3s50uSSN	0%, 9.27 W	0%, 9.40 W	0%, 9.27 W
3s80uSNN	0%, 14.07 W	0%, 14.26 W	0%, 14.06 W
3s80uSSN	0%, 13.43 W	0%, 13.50 W	0%, 13.43 W

Table 5.7 indicates that TS with *tabu-pen* always provides the best solution

(underlined). This tabu declaration method is less strict than the two others, hence it provides more flexible neighborhood exploration. Moreover, although *tabu-tri* is the most strict tabu type, the adaptive tabu tenure mechanism loosens its stringency. The tabu list for *tabu-tri* is shorter since it has smaller tenure. Therefore, it can sometimes also find good solutions as *tabu-pen*. Besides, the solutions in Table 5.7 are better than those of Table 5.6, which shows the advantage of the tabu repair mechanism.

5.3.2   Results comparison with other metaheuristic algorithms

The proposed TS algorithm is compared with the Greedy Local Search (GLS) in chapter 3 and the Simulated Annealing (SA) in chapter 4. In Figure 5.4, the bars show the rank distributions of F2R2M-TS, GLS and F2R2M-SA solutions for four scenarios. Here F2R2M-TS applies the tabu repair mechanism and uses the *tabu-pen* structure to record the tabu moves.

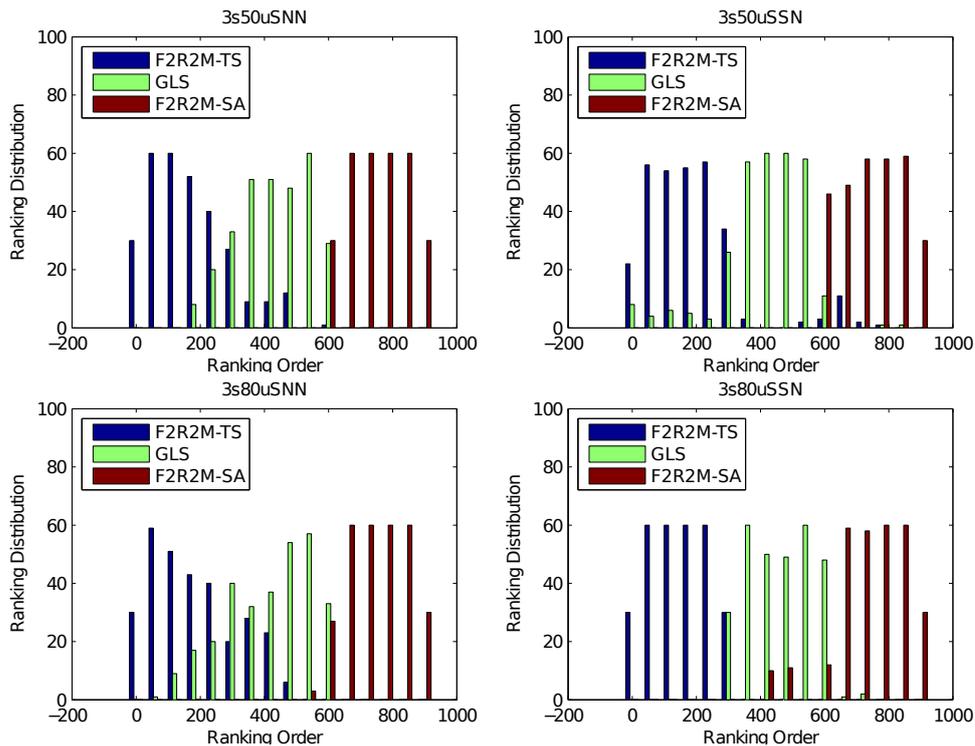Figure 5.4: Rank distributions of three metaheuristic algorithms

GLS starts from a random initial solution, evaluates all its feasible neighbors and replaces the current solution by the neighborhood solution that has the best fitness. GLS stops at the local optimum when no neighbor is better than the current solution. F2R2M-SA generates an initial solution by scalable MBMS power counting, then

it randomly modifies the current solution in each iteration. The new solution will be accepted if it has better fitness than the current one, else it is accepted with an acceptance probability depending on the current temperature. $100 \times 4^{N_t(s_i)}$ iterations are performed at each temperature until the temperature declines under a given threshold. The following experiments are conducted to perform the algorithm comparison. For each given scenario, each algorithm is run 300 times. All solutions of these algorithms are ordered in the same database. The ranking number ranges from 1 to 900 and the smaller the rank is, the better the solution. This range is divided into 16 subsets, the number of ranking values in each subset is then counted. Such number reflects the ranking distribution and shows the distribution of solution ranking for a corresponding algorithm.

Table 5.8 lists the best solutions found by the three metaheuristics over the 18 scenarios (each algorithm runs 300 trials per scenario). The feasible solutions ($Po < 19W$) are underlined.

In Table 5.8, for the small-sized scenarios (one service), all algorithms can find the same solutions, for both scalable and non-scalable transmission schemes. That is because the landscape of these scenarios is quite smooth, having few local optima.

However, for the large-sized scenarios, F2R2M-TS outperforms GLS while F2R2M-SA obtains worse solution than GLS. GLS is the most simple heuristic, but is not robust in terms of solution quality. A well-known disadvantage of GLS is that, when it reaches a state where no further improvement can be found, GLS stops the solution modification and output the current solution, such states are referred to as local optimum.

Besides, although SA is easy to implement and adaptable [69], the randomness of the solution update mechanism and the complex search parameter setting (i.e. cooling schedule and acceptance possibility) bring instability in solution quality. Because the solution spaces of the large-sized scenarios are rugged, SA is not suitable for F2R2M for its randomness characteristics in new solution selection and acceptance. It should need more tests with different parameter setting to improve it.

In this work, the simulations are conducted on Linux PC with Intel(R) Core(TM) i5 CPU 750 @ 2.67GHz. Table 5.9 shows the average computation time of the three approaches for all scenarios on 300 runs. The consuming time of the search procedure in F2R2M-TS is about 1 to 3 seconds in the regular Linux PC. It is acceptable for a RNC server. A stop criterion is defined both in F2R2M-SA and F2R2M-TS; they stop when the consecutive number of iterations with non-improving solutions (comparing to the best solution found) is more than $100N_f$, where N_f is the total number of flows in the cell. GLS costs the shortest time since it does not conduct the tabu evaluation. For simple scenarios, F2R2M-TS costs longer time than F2R2M-SA because it requires to evaluate and update the tabu list. For complex scenarios,

Table 5.8: Best solutions found by three metaheuristic algorithms

Scenarios	F2R2M-TS	GLS	F2R2M-SA
1s30uN	0%, 5.18 W	0%, 5.18 W	0%, 5.18 W
1s30uS	0%, 5.31 W	0%, 5.31 W	0%, 5.31 W
1s60uN	0%, 11.62 W	0%, 11.62 W	0%, 11.62 W
1s60uS	0%, 12.82 W	0%, 12.82 W	0%, 12.82 W
1s150uN	0%, 14.65 W	0%, 14.65 W	0%, 14.65 W
1s150uS	0%, 15.38 W	0%, 15.38 W	0%, 15.38 W
2s20uSN	0%, 4.328 W	0%, 4.66 W	0%, 7.83 W
2s20uSS	0%, 4.35 W	0%, 5.12 W	0%, 8.80 W
2s50uSN	0%, 11.12 W	0%, 11.17 W	0%, 10.19 W
2s50uSS	0%, 10.94 W	0%, 10.92 W	0%, 13.06 W
3s30uSNN	0%, 5.33 W	0%, 6.15 W	0%, 8.72 W
3s30uSSN	0%, 5.31 W	0%, 6.27 W	0%, 10.97 W
3s50uSNN	0%, 9.27 W	0%, 11.18 W	0%, 18.8 W
3s50uSSN	0%, 9.27 W	0%, 11.38 W	0%, 18.79 W
3s80uSNN	0%, 14.06 W	0%, 14.49 W	0 %, 15 W
3s80uSSN	0%, 13.43 W	0%, 14.52 W	0%, 14.36 W
3s100uSNN	2.35 %, 18.11 W	4.12 %, 17.51 W	11.77 %, 18.96 W
3s100uSSN	0.59 %, 17.86 W	1.76 %, 18.11 W	15.89 %, 18.79 W

Table 5.9: Average time cost (second)

Scenario	F2R2M-TS	GLS	F2R2M-SA
1s60uS	0.65	0.02	0.81
1s150uS	1.01	0.04	0.92
2s50uSN	1.36	0.16	1.24
2s50uSS	1.42	0.26	1.35
3s50uSNN	1.87	0.32	1.82
3s50uSSN	1.84	0.28	1.95
3s80uSNN	2.88	0.37	3.52
3s80uSSN	2.99	0.34	3.78

F2R2M-SA costs much longer time than F2R2M-TS because at each iteration the probability of randomly generating an unfeasible solutions is too high, therefore, the most computation time is spent on searching for (randomly pick out) a feasible solution. As F2R2M-TS can quickly find a good enough solution, the non-improving iteration number larger than $100N_f$ is earlier satisfied than for F2R2M-SA for the complex scenarios.

5.3.3 Performance comparison with theoretical lower bounds

In chapter 3, it has been shown that MBMS RRM problem can be approximated as a Multiple-Choice Knapsack Problem (MCKP) when relaxing the OVSF code constraints. Solving the MCKP gives the solution lower bound for the MBMS RRM problem, which can be considered as the approximation of the MBMS RRM optimum solution. In this section, to better evaluate the performance of F2R2M-TS, a comparison is conducted between F2R2M-TS and the theoretical solution bound. Table 5.10 shows the lower bounds by solving the MCKP using Gurobi solver, and the best solutions found by using F2R2M-TS. It shows that, in the given scenarios, F2R2M-TS can find the optimum or near-optimum solutions.

Table 5.10: Solution lower bounds found by Gurobi and the best solutions found by F2R2M-TS

Scenarios	Gurobi	F2R2M-TS
1s10u128N	0%, 1.352 W	0%, 1.352 W
1s10u128S	0%, 1.34 W	0%, 1.34 W
2s20uSN	0%, 4.328 W	0%, 4.328 W
2s20uSS	0%, 4.263 W	0%, 4.346 W
3s30uSNN	0%, 5.297 W	0%, 5.328 W
3s30uSSN	0%, 5.231 W	0%, 5.314 W

5.4 Synthesis

In this chapter, the Tabu Search (TS) algorithm is conceived to solve the MBMS RRM problem. First, the general concepts of TS are introduced then an adaptation of TS algorithm, named F2R2M-TS, is proposed to solve the MBMS RRM problem.

In the proposed algorithm the move structure is designed based on the multiple insert operator δ_{MI}. Each move includes five parameters. From the move structure,

the neighborhood generation is illustrated. This procedure explores all the candidate moves and constructs the candidate move list.

The tabu memory structure plays a key role for the search efficiency. Three tabu structures are proposed for tabu declaration. *Tabu-tri* stipulates that the modification of the allocated channel of a certain user for a given flow is tabu-active. *Tabu-qua* stipulates that a certain target channel for a given user is tabu-active, hence this tabu structure forbids to insert a given user back to a given channel. *Tabu-pen* sets that transferring a given user from an initial channel to a target channel is tabu-active. *Tabu-tri* is the most restrictive tabu structure. An aspiration mechanism is used to authorize a tabu move when the candidate solution that can be obtained is the best ever known solution. Moreover, based on the characteristics of the application, a mechanism named tabu repair is proposed to re-allow a tabu-declared move by modifying a sub-set of its attributes. This mechanism makes the moves self-adaptive during the search iterations and therefore improves the search efficiency.

The performance of six alternatives of F2R2M-TS strategies are compared: three tabu declaration methods, with and without tabu repair mechanism. A performance metric named rank distribution is proposed for the comparison. The comparative study shows that *tabu-pen* is more efficient than the other two tabu structures, and the proposed tabu repair mechanism leads to improve the search performance significantly. Then comparison simulations among F2R2M-TS, F2R2M-SA and GLS are carried out in a variety of scenarios. The simulations show that F2R2M-TS provides higher robustness in terms of solution quality, and higher efficiency in terms of time cost. Besides, F2R2M-TS can find optimum and near-optimum solutions comparing to the theoretical solution bounds found with the MCKP model solved exactly with the Gurobi solver.

Conclusion and Future Work

The conflict between the demand of high throughput multimedia multicast service and the reality of limited radio resources results in strong requirement for efficient radio resource allocation schemes in UMTS MBMS system. Some research work has been done on this field, but in general there are several shortcomings. Firstly, almost all existing schemes try to solve the issue in a perceptional way but lacking systematic analysis of the targeting problem. Secondly, some existing schemes lack enough flexibility so that they can only work in certain scenarios but has limitations in other scenarios. Thirdly, the existing schemes do not fully explore the solution space, therefore the obtained allocation solutions are not optimal. To address these issues, this Ph.D. dissertation proposed a systematic approach to solve the MBMS RRM problem. This approach covers a formal mathematical modeling of the targeting problem, an in-depth analysis of the solution boundary and the solution space based on the proposed model, as well as practical algorithms which can achieve optimal or near-optimal solutions. In the following two sections, first, the contributions accomplished by this Ph.D work will be summarized, and then, the perspectives of further work will be outlined.

A summary of contributions

The first contribution of this thesis is the proposal of a formal modeling of the MBMS RRM problem, named as F2R2M, as well as a C++ based simulator which implemented the proposed model. The work shows that MBMS RRM problem can be modeled as a combinational optimization problem. To better explore the solution space, the model proposes the flexibility of transmission mode selection and scalable multimedia streaming. F2R2M stipulated the definition of input parameters, decision parameters and optimization objectives. Seven modules were abstracted in this model, constructing an iterative architecture to explore the solution space. The allocation procedure designed in this model targets at finding good or optimum solution for two-dimensional objective: a) the quality of resources allocation in terms of service satisfaction; b) the resource consumption in terms of transmission power and channel code occupation. Guided with a proposed lexicographic-order evaluation criteria, this model aimed to find the optimal radio resource assignment with satisfiable throughput requirement and minimal transmission power. The proposed model has been implemented as a simulator containing the abstraction of MBMS RRM procedure, as well as the power emulator and the OVSF code allocator. In particular, the transmission power for different MBMS channels are simulated. Simulations showed that the transmission power is non-linear and quite dynamic depending on the user scenarios. Such results justified that, the concept of flexible radio resource

management is a reasonable choice for minimizing the power consumption as well as maximizing the cell capacity.

The second contribution is the mathematical approximation of the MBMS RRM problem. To understand the problem complexity and the solution boundaries, the work proved that the studied problem could be approximated as a Multiple-Choice Knapsack Problem (MCKP) by relaxing the OVSF code constraints. In the MCKP based formulation, each class is associated with a given flow of service. Each class includes multiple items, which are potential allocation scheme for the associated flow. Such finding generates two outcomes. First, the NP-Hard proof for MBMS RRM is self-contained because MCKP is NP-hard, in which the solution size is exponentially increased with the number of users. Second, solving MCKP could give theoretical solution bounds for the MBMS RRM solver.

In the third contribution, to illustrate the structure of solution space for the targeting problem, the fitness landscape analysis technique was conducted on F2R2M. In this work, the solution representation and solution distance measurement were defined. Then two landscapes differentiated by two neighborhood operators, single insert operator and multiple insert operator, were generated. The characteristics of the studied problem were analyzed regarding the search space and the solution distribution. Simulations showed that the solution space is rugged for the targeting problem, meaning that the search space is filled with local optima. Simulations also showed that the multiple insert operator is better than the single insert operator for efficient search.

The forth contribution is the proposal of a modified Simulated Annealing (SA) algorithm to solve the MBMS RRM problem. The work showed that it is feasible to modify and map SA algorithm on the proposed model, named F2R2M-SA. F2R2M-SA generates the initial solution based on the MBMS power counting algorithm for each flow of services. In each iteration of the search procedure, the new solution is randomly generated by one application of the multiple insert neighborhood operator. The new solution is accepted when it is better than the current solution, otherwise, to prevent the search procedure to stick in a local minimum, the new solution is rejected with an acceptance probability. The acceptance probability function is composed by two parameters K_p and K_t, which are related to the two-dimensional fitness values (the power and the allocated throughput). Simulation results show that F2R2M-SA can find better solutions than the other UMTS existing algorithms. F2R2M-SA obtained all feasible solutions while the existing algorithms obtained either unfeasible solutions (power over the maximum limit), or feasible solutions with low QoS (loss of throughput). The simulation results of F2R2M-SA proved that, the proposed model could provide the solution balancing the power consumption and the service quality, and reduce the possibility of radio resource saturation by adapting the combinational channel assignment.

In the fifth contribution, the general Tabu Search (TS) algorithm was modified and mapped on F2R2M, named F2R2M-TS. Comparing with F2R2M-SA, this algorithm further improves the search efficiency and solution quality by avoiding to re-visit the previous found solutions. Specific to this application, three tabu structures for tabu evaluation are defined to conduct the tabu evaluation. Simulation results show that the tabu structure *tabu-pen* offers the best performance. Besides, a mechanism named tabu repair is designed to extend the classic TS algorithm. Simulation results show that, this mechanism explores more potential solutions and provides higher solution quality. Such mechanism re-allows a tabu-declared move by modifying a sub-set of the tabu attributes, so that the search not only avoids re-visiting a previous bad solution but also self-adjusts to the optimal solution. To better compare the different metaheuristic algorithms, a metric called rank distribution is proposed to compare the statistics of the best solutions found by different algorithms. Simulations are conducted on the F2R2M simulator and the results show that F2R2M-TS could find better solutions than F2R2M-SA and Greedy Local Search (GLS) within acceptable computation time. The best solutions found by F2R2M-TS are also compared with the theoretical solution bounds found by the MCKP solver. It shows that they are equivalent with or closely approaching the theoretical optimal solutions.

Perspectives for further work

For further research, there are sill some improvements which can be done based on the current work.

Scenario tracking. In the current work, RNC assumes that the user scenario is unchanged within a very short amount of time (e.g. 1 seconds). Therefore, the search space stays constant before the best solution found is obtained by the MBMS RRM solver. However, in hot-spot locations, for example the city center, the user scenario partially changes in a very short time. It means that the search space is gradually changed before a search is finished. It will be interesting to study how robust our current algorithms can be in such situation. On the other hand, in reality, there are similarities between two subsequent user scenarios. This gives the opportunity that the best solution found for the previous scenario could be used to speed up the search to solve the next scenario. It results in an iterative solver which is tracking the field scenario changes. A new dimensional research can be explored on such tracking mechanism.

Computation speed up by GPGPU programming. In the current implementation, the averaged computation time of the proposed metaheuristic algorithms for

solving MBMS RRM is about 1 second in regular Linux PC. It is acceptable for a RNC server but the computation time can be significantly shortened by programming on General-Purpose Graphics Processing Units (GPGPU). This requires the adaption of MBMS RRM algorithms such that it can be parallelized and then mapped into a GPGPU architecture, e.g. CUDA [41]. Some literatures about parallel metaheuristics are listed [62, 33, 58]. A speed up of the MBMS RRM solver will give performance improvement in dynamical scenarios.

Extend the concept to E-MBMS system.   The Evolved MBMS (E-MBMS) has been standardized in various groups of 3GPP as part of LTE release 9 [7]. The physical layer of E-MBMS in LTE is different comparing with MBMS in UMTS but the similar concept of F2R2M can be extended for E-MBMS by replacing OVSF codes with time/frequency resource blocks. More flexible variants E-MBMS standards will better fit the flexible radio resource management concept and therefore better performance gain can be expected.

Appendix A: Acronyms

3GPP 3rd Generation Partnership Project

4G 4th Generation

AMC Adaptive Modulation and Coding

AUC Authentication Center

AGWN Additive Gaussian White Noise

BLER Block Error Rate

BMSC Broadcast Multicast Service Centre

BS Base Station

BTS Base Transceiver Station

CAGE Compound Annual Growth Rate

CBS Cell Broadcast Service

CDMA Code Division Multiple Access

CM Connection Management

CN Core Network

CNIR Carrier Noise and Interference Ratio

CPICH Common Pilot Channel

CQI Channel Quality Information

CS Circuit Switched

CUDA Compute Unified Device Architecture

DCH Dedicated Channel

DPS Dynamic Power Setting

DTM Dual Transmission Mode

Eb/No Energy per Bit to Noise power spectral density ratio

Ec/No Energy per Chip to Noise power spectral density ratio

E-MBMS Evolved Multimedia Broadcast Multicast Service

F2R2M Flexible Radio Resource Management Model

FACH Forward Access Channel

FDD Frequency Division Duplex

FDMA Frequency Division Multiple Access

GGSN Gateway GPRS Support Node

GLS Greedy Local Search

GMM GPRS Mobility Management

GMSC Gateway Mobile Switching Center

GPGPU General-Purpose Graphics Pocessing Units

GPRS General Packet Radio Service

GSM Global System for Mobile communications

KP Knapsack Problem

HARQ Hybrid Automatic Repeat reQuest

HLR Home Location Register

HSDPA High Speed Downlink Packet Access

HS-DSCH High Speed Downlink Shared Channel

IGMP Internet Group Management Protocol

IP Internet Protoco

ITU International Telecommunication Union

LTE Long Term Evolution

MAC Media Access Control

MBMS Multimedia Broadcast Multicast Service

MCCH MBMS Control Channel

MCKP Multiple-Choice Knapsack Problem

MKP Multi-dimensional Knapsack Problem

ML Multilayer (transmission scheme)

MM Mobility Management

MMKP Multi-dimensional Multiple-choice Knapsack Problem

MPC MBMS Power Counting

MSC Mobile Switching Center

MSCH MBMS Scheduling Channel

MTCH MBMS Point-to-Multipoint Traffic Channel

OVSF Orthogonal Variable Spreading Factor

PDP Packet Data Protocol

PHY Physical layer

PS Packet Switched

PTM Point-to-Multipoint

PTP Point-to-Point

QAM Quadrature Amplitude Modulation

QoS Quality of Service

RLC Radio Link Control

RNC Radio Network Controller

RNS Radio Network Subsystem

RRC Radio Resource Control

RRM Radio Resource Management

SA Simulated Annealing

S-CCPCH Secondary Common Control Physical Channel

SF Spreading Factor

SGSN Serving GPRS Support Node

SINR Signal to Interference Noise Ratio

SL Single Layer (transmission scheme)

SM Session Management

STTD Spatio-Temporal Transmit Diversity

TDMA Time Division Multiple Access

TFCI Transport Format Combination Indicator

TFI Transport Format Indicato

TS Tabu Search

TTI Transmission Time Interval

UE User Equipment

UMTS Universal Mobile Telecommunications System

USIM UMTS Subscriber Identity Module

UTRAN UMTS Terrestrial Radio Access Network

VLR Visitor Location Register

WCDMA Wideband Code Division Multiple Access

Appendix B: CQI mapping table

Table B.1: CQI mapping table [13]

CQI value	Transport block size (bits)	Num. of codes	Modulation	Total bit rate (kbps)
1	137	1	QPSK	480
2	173	1	QPSK	480
3	233	1	QPSK	480
4	317	1	QPSK	480
5	377	1	QPSK	480
6	461	1	QPSK	480
7	650	2	QPSK	960
8	792	2	QPSK	960
9	931	2	QPSK	960
10	1262	3	QPSK	1440
11	1483	3	QPSK	1440
12	1742	3	QPSK	1440
13	2279	4	QPSK	1920
14	2583	5	QPSK	1920
15	3319	5	QPSK	2400
16	3565	5	16-QAM	4800
17	4189	5	16-QAM	4800
18	4664	5	16-QAM	4800
19	5287	5	16-QAM	4800
20	5887	5	16-QAM	4800
21	6554	5	16-QAM	4800
22	7168	5	16-QAM	4800
23	9719	7	16-QAM	6720
24	11418	8	16-QAM	7680
25	14411	10	16-QAM	9600
26	17237	12	16-QAM	11520
27	21754	12	16-QAM	14400
28	23370	15	16-QAM	14400
29	24222	15	16-QAM	14400
30	25558	15	16-QAM	14400

Publication list during PhD study

Journal papers

Qing Xu, Frédéric Lassabe, Hakim Mabed, Alexandre Caminada, Fitness Landscape Analysis for Scalable Multicast RRM Problem in Cellular Network, In *Electronic Notes in Discrete Mathematics*, ELSEVIER, Volume 41, Pages 407-414, ISSN 1571-0653.

International conference papers

(PIMRC 2014) Qing Xu, Hakim Mabed, Alexandre Caminada and Frédéric Lassabe, 'Solving MBMS RRM Problem by Metaheuristics', In *2014 IEEE 25th International Symposium on Personal, Indoor and Mobile Radio Communications* Washington DC, USA, 2-5 September, 2014.

(VTC 2014-Fall) Qing Xu, Hakim Mabed, Frédéric Lassabe and Alexandre Caminada, MBMS Radio Resource Optimization by Tabu Search, In *2014 IEEE 80th Vehicular Technology Conference*, Vancouver, Canada, 14-17 September, 2014.

(VTC 2013-Spring) Qing Xu, Hakim Mabed, Frédéric Lassabe and Alexandre Caminada, Optimization of Radio Resource Allocation for Multimedia Multicast in Mobile Networks, In *2013 IEEE 77th Vehicular Technology Conference*, Dresden, Germany, 2-5 June 2013.

(WiOpt'13) Qing Xu, Hakim Mabed, Frédéric Lassabe and Alexandre Caminada, Modeling and Fitness Landscape Analysis for Flexible MBMS Radio Resource Allocation, In *11th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, Tsukuba Science City, Japan, 13-17 May, 2013.

National conference papers

(ROADEF'13) Qing Xu, Frédéric Lassabe, Hakim Mabed and Alexandre Caminada, Allocation par métaheuristique des ressources radio pour service de multi-diffusion hiérarchique, In *14e conference ROADEF de la société Française de Recherche Opérationnelle et Aide à la Décision*, Troyes, France, 13-15 February, 2013.

(JDIR11) Qing Xu, Hakim Mabed, Frédéric Lassabe and Alexandre Caminada, 'Ef-

ficient Dynamic Radio Resource Management for Multimedia Broadcast Multicast Services', In *Proceedings of 10eme Journées Doctorales en Informatique et Réseaux*, Belfort, France, 25 November, 2011.

Bibliography

[1] 3GPP TR 25.992 Multimedia Broadcast/Multicast Service (MBMS) ; UTRAN/GERAN requiremenst.

[2] 3GPP TR 30.03U Selection procedures for the choice of radio transmission technologies of the UMTS.

[3] 3GPP TS 22.060 General Packet Radio Service (GPRS); Service description; Stage 1.

[4] 3GPP TS 23.060 General Packet Radio Service (GPRS); Service description; Stage 2,.

[5] 3GPP TS 25.211 Physical channels and mapping of transport channels onto physical channels (FDD).

[6] 3GPP TS 25.213 Spreading and modulation (FDD).

[7] 3GPP TS 25.346 Introduction of the Multimedia Broadcast/Multicast Service (MBMS) in the Radio Access Network (RAN); Stage 2 (version 9.1.0 Release 9).

[8] OPNET University Program: http://www.opnet.com /services/university/.

[9] 3rd Generation Partnership Project R1-021240 TSG-RAN WG128 Power Usage for Mixed FACH and DCH for MBMS, 2002.

[10] 3GPP TS 23.041: Technical Realization of Cell Broadcast Service (CBS), Jan. 2004.

[11] *Usage of link-level performance indicators for HSDPA network-level simulations in E-UMTS*, Sydney, Australia, 2004. Proceedings of IEEE ISSSTA 04.

[12] 3GPP TR 25.803 S-CCPCH performance for MBMS25.922, September 2005.

[13] 3GPP TS 25.214 Physical layer procedures (FDD), June 2005.

[14] 3GPP TS 25.324 Broadcast/Multicast Control (BMC), June 2006.

[15] 3GPP TR 25.922 Radio resource management strategies, March 2007.

[16] 3GPP TS 23.246 Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description (Release 6), June 2007.

[17] 3GPP TS 23.246 Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description (Release 7), September 2007.

[18] 3GPP TS 25.346 Introduction of the Multimedia Broadcast/Multicast Service (MBMS) in the Radio Access Network (RAN); Stage 2, March 2011.

[19] UMTS Forum Summary of Recent Activities. Technical report, UMTS Forum, 2011.

[20] VNI mobile forecast highlights 2012 - 2017, 2013.

[21] A. Alexiou, C. Bouras, and V. Kokkinos. An enhanced MBMS power control mechanism towards long term evolution. In *Wireless Telecommunications Symposium, 2009. WTS 2009*, pages 1–6, April 2009.

[22] A. Alexiou, C. Bouras, V. Kokkinos, and E. Rekkas. MBMS power planning in macro and micro cell environments. In *12th IEEE Symposium on Computers and Communications, (ISCC 2007)*, pages 33– 38, July 2007.

[23] Antonios Alexiou, Christos Bouras, and Vasileios Kokkinos. Evaluation of different power saving techniques for MBMS services. *EURASIP Journal on Wireless Communications and Networking*, pages 1–15, January 2009.

[24] Antonios Alexiou, Christos Bouras, and Evangelos Rekkas. A power control scheme for efficient radio bearer selection in MBMS. In *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, pages 1–8, June 2007.

[25] Antonios Alexiou, Christos Bouras, and Evangelos Rekkas. An improved mbms power counting mechanism towards long term evolution. *Telecommunication Systems*, 43(1-2):109–119, 2010.

[26] Cristina Bazgan, Hadrien Hugot, and Daniel Vanderpooten. Solving efficiently the 0-1 multi-objective knapsack problem. *Computers and Operations Research*, 36(1):260 – 279, 2009.

[27] Christian Blum and Andrea Roli. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys*, 35(3):268– 308, September 2003.

[28] Christophoros Christophorou, Andreas Pitsillides, and Tomas Lundborg. Enhanced radio resource management algorithms for efficient mbms service provision in UTRAN. *Computer Networks*, 55(3):689–710, 2011.

[29] PC Chu and JE Beasley. A genetic algorithm for the multidimensional knapsack problem. *Joural of Heuristics*, 4(1):63–86, Jun 1998.

[30] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction To Algorithms*. MIT Press, 2001.

[31] Amrico M. C. Correia, Joo C. M. Silva, Nuno M. B Souto, Lusa A. C. Silva, Alexandra B. Boal, and Armando B. Soares. Multi-resolution broadcast/multicast systems for MBMS. *IEEE Transactions on Broadcasting*, 53(1):224–234, March 2007.

[32] P.J. Czerepinski, T.M. Chapman, and J. Krause. Coverage and planning aspects of MBMS in UTRAN. In *Fifth IEE International Conference on 3G Mobile Communication Technologies*, 2004.

[33] I. De Falco, R. Del Balio, E. Tarantino, and R. Vaccaro. Improving search by incorporating evolution principles in parallel Tabu Search. In *Proceedings of the First IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence,*, pages 823–828 vol.2, Jun 1994.

[34] Isabelle Devarenne, Hakim Mabed, and Alexandre Caminada. Adaptive tabu tenure computation in local search. In Jano Hemert and Carlos Cotta, editors, *Evolutionary Computation in Combinatorial Optimization*, volume 4972 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg, 2008.

[35] M.M. Diallo, Samuel Pierre, and R. Beaubrun. A tabu search approach for assigning node bs to switches in umts networks. *IEEE Transactions on Wireless Communications*, 9(4):1350–1359, April 2010.

[36] R.W. Eglese. Simulated annealing: A tool for operational research. *European Journal of Operational Research*, 46(3):271 – 281, 1990.

[37] France Telecom University of Surrey RWTH Aachen RAI Ericsson, Motorola. IST-2001-35125, Deliverable of the project (D08), Spectrum efficient multicast and asymmetric servicesin UMTS. Technical report, 2003.

[38] Thomas Erlebach, Hans Kellerer, and Ulrich Pferschy. Approximating multi-objective knapsack problems. In *Algorithms and Data Structures*, Lecture Notes in Computer Science, pages 210–221. Springer Berlin Heidelberg, 2001.

[39] Cyril Fonlupt, Denis Robilliard, Philippe Preux, and El-Ghazali Talbi. Fitness landscapes and performance of meta-heuristics. In *Meta-Heuristics: Advances and Trends in Local Search Paradigms for Optimization*, page 257268. Kluwer Academic Publishers, 1999.

[40] Arnaud Fréville. The multidimensional 01 knapsack problem: An overview. *European Journal of Operational Research*, 155(1):1 – 21, 2004.

[41] M. Garland. Parallel computing with cuda. In *Parallel Distributed Processing (IPDPS), 2010 IEEE International Symposium on*, pages 1–1, April 2010.

[42] Stuart Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, Nov 1984.

[43] Fred Glover. Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, 13(5):533 – 549, 1986.

[44] Fred Glover. Tabu search part I. *ORSA Journal on computing*, 1(3):190–206, 1989.

[45] Fred W. Glover and Gary A. Kochenberger, editors. *Handbook of Metaheuristics*. Springer, 1st edition, January 2003.

[46] Gurobi Optimization. *Gurobi Optimizer Reference Manual*, December 2009.

[47] Hans-Otto Gunther and Kap Hwan Kim. *Container terminals und cargo systems: design, operations management, and logistics control issues*. Springer, Berlin, 2007.

[48] H. Hamacher. Theoretical and computational aspects of simulated annealing: Theory and applications. *SIAM Review*, 32(3):504–506, 1990.

[49] Antti Toskala Harri Holma. *HSDPA/HSUPA for UMTS: High Speed Radio Access for Mobile Communications*. John Wiley and Sons, 2006.

[50] S. Mehdi Hashemi, Mohsen Rezapour, and Ahmad Moradi. Two new algorithms for the min-power broadcast problem in static ad hoc networks. *Applied Mathematics and Computation*, 190(2):1657 – 1668, 2007.

[51] Mariann Hauge. Multicast in 3G networks: Employment of existing IP multicast protocols. pages 96–103, 2002.

[52] Cornelius Hellge, Thomas Schierl, Jörg Huschke, Thomas Rusert, Markus Kampmann, and Thomas Wiegand. Graceful degradation in 3GPP MBMS mobile TV services using H.264/AVC temporal scalability. *EURASIP Journal on Wireless Communications and Networking*, 2009:207210, 2009.

[53] Harri Holma and Antti. Toskala. *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. Wiley, 2000.

[54] Harri Holma and Antti Toskala. *WCDMA for UMTS: HSPA Evolution and LTE*. Wiley Online Library: Books. Wiley, 2010.

[55] IST-2003-507607 (B-BONE). Final results with combined enhancements of the air interface, deliverable of the project (D2.5). Technical report.

[56] A. Van Ewijk J. De Vriendt, I. Gomez Vinagre. Multimedia broadcast and multicast services in 3G mobile networks. Technical Report 4-1, Alcatel Telecommunications Review, 2004.

[57] Kil-Woong Jang. A tabu search algorithm for routing optimization in mobile ad-hoc networks. *Telecommunication Systems*, 51(2-3, SI):177–191, NOV 2012.

[58] A. Janiak, W. ladys law Janiak, and M. Lichtenstein. Tabu search on GPU. *Journal of Universal Computer Science*, 14(14):24162427, 2008.

[59] Hany Kamal, Marceau Coupechoux, and Philippe Godlewski. Tabu search for dynamic spectrum allocation in cellular networks. *IEEE Transactions on Emering Telecommunications Technologies*, 23(6):508–521, oct 2012.

[60] Richard M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, page 85103. Plenum, New York, 1972.

[61] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[62] The Van Luong, N. Melab, and E.-G. Talbi. Gpu computing for parallel local search metaheuristic algorithms. *Computers, IEEE Transactions on*, 62(1):173–185, Jan 2013.

[63] Hakim Mabed, Alexandre Caminada, and Jin-Kao Hao. Genetic Tabu search for robust fixed channel assignment under dynamic traffic data. *Computational Optimization and Applications*, 50(3):483–506, DEC 2011.

[64] Raïd Mansi, Cláudio Alves, J. M. Valério de Carvalho, and Saïd Hanafi. A hybrid heuristic for the multiple choice multidimensional knapsack problem. *Engineering Optimization*, 45(8):983–1004, 2013.

[65] A. Medeisis and A. Kajackas. On the use of the universal okumura-hata propagation prediction model in rural areas. In *Vehicular Technology Conference Proceedings, (VTC-Spring) Tokyo.*, volume 3, pages 1815–1818 vol.3, 2000.

[66] Peter Merz and Bernd Freisleben. Fitness landscapes and memetic algorithm design. In *New Ideas in Optimization*, pages 245–260. McGraw-Hill, 1999.

[67] S. Parkvall, E. Englund, M. Lundevall, and J. Torsner. Evolving 3g mobile systems: broadband and broadcast services in wcdma. *Communications Magazine, IEEE*, 44(2):30–36, Feb 2006.

[68] A. Raschella, A. Umbert, G. Araniti, A. Iera, and A. Molinaro. SINR-Based transport channel selection for mbms applications. In *IEEE 69th Vehicular Technology Conference, (VTC Spring) 2009*, pages 1–5, April 2009.

[69] C. R. Reeves. *Modern Heuristic Techniques for Combinatorial Problems.* John Wiley & Sons, 1 edition, April 1993.

[70] Jordi Perez Romero, Oriol Sallent, Ramon Agusti, and Miguel Angel Diaz-Guerra. *Radio Resource Management Strategies in UMTS*. Wiley, 1 edition, August 2005.

[71] Said Salhi. Defining tabu list size and aspiration criterion within tabu search methods. *Computers and Operations Research*, 29(1):67–86, 2002.

[72] Jochen H Schiller. *Mobile Communications.* Addison-Wesley, London, 2011.

[73] Luísa Silva and Américo Correia. Hsdpa delivering MBMS video streaming. In *Proceedings of the 2006 ACM CoNEXT Conference*, CoNEXT '06, pages 20:1–20:2, New York, NY, USA, 2006. ACM.

[74] Prabhakant Sinha and Andris A. Zoltners. The multiple-choice knapsack problem. *Operations Research*, 27(3):503–515, 1979.

[75] PeterF. Stadler and SantaFe Institute. Towards a theory of landscapes. In *Complex Systems and Binary Networks*, volume 461-461 of *Lecture Notes in Physics*, pages 78–163. Springer Berlin Heidelberg, 1995.

[76] B. Suman and P. Kumar. A survey of simulated annealing as a tool for single and multiobjective optimization. *Journal of the Operational Research Society*, 57(18):1143–1160, 2006.

[77] Harold Szu and Ralph Hartley. Fast simulated annealing. *Physics Letters A*, 122(34):157 – 162, 1987.

[78] Lucent Technologies. 3GPP TSG RAN WG1 No.29; R1-02-1324; Power Control for FACH, November 2002.

[79] Lucent Technologies. TSG-RAN WG1 No.33 r1-030778; Rate Splitting for MBMS, August 2003.

[80] V. Vartiainen and J. Kurjenniemi. Point-to-multipoint multimedia broadcast multicast service (MBMS) performance over HSDPA. In *IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, (PIMRC 2007)*, pages 1–5, 2007.

[81] Kun Wang, Fan Yang, Qian Zhang, Dapeng Oliver Wu, and Yinlong Xu. Distributed cooperative rate adaptation for energy efficiency in IEEE 802.11-based multihop networks. *IEEE Transactions on Vehicular Technology*, 56(2):888–898, MAR 2007.

[82] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proc of the 6th International Congress of Genetics*, volume 1, page 356366, 1932.

[83] V. erný. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, 1985.