공학박사학위논문

Performance Enhancement of Video Delivery Services in LTE Networks

LTE 네트워크에서 비디오 전달 서비스의 성능 향상

2015년 2월

서울대학교 대학원
전기 · 컴퓨터공학부
이지훈

Performance Enhancement of Video Delivery Services in LTE Networks

지도교수 권태경

이 논문을 공학박사 학위논문으로 제출함

2014년 10월

서울대학교 대학원

전기 · 컴퓨터공학부

이지훈

이지훈의 박사 학위논문을 인준함

2014년 12월

위 원 장	김 종 권	(인)
부위원장	권 태 경	(인)
위 원	전 병 곤	(인)
위 원	엄 현 상	(인)
위 원	백 상 헌	(인)

Abstract

Performance Enhancement of Video Delivery Services in LTE Networks

Ji Hoon Lee
School of Computer Science & Engineering
The Graduate School
Seoul National University

LTE includes an enhanced multimedia broadcast/multicast service (eMBMS); but delay-sensitive real-time video streaming requires the combination of efficient handling of wireless link bandwidth and reduced handover delays, which remains a challenge. The 3GPP standard introduces a Multimedia Broadcast and multicast service over a Single Frequency Network (MBSFN) area which is a group of base stations broadcasting the same multicast packets. It can reduce the handover delay within MBSFN areas, but raises the traffic load on LTE networks.

In this dissertation, we first presents an MBSFN architecture based on location management areas (LMAs) which can increase the sizes of MBSFN areas to reduce the average handover delay without too much bandwidth waste. An analytical model is developed to quantify service disruption time, bandwidth usage, and blocking probability for different sizes of MBSFN areas and LMAs while considering user

mobility, user distribution, and eMBMS session popularity. Using this model, we also propose how to determine the best sizes of MBSFN areas and LMAs along with performance guarantees. Analytical and simulation results demonstrate that our LMA-based MBSFN scheme can achieve bandwidth-efficient multicast delivery while retaining an acceptable service disruption time.

We next propose to transmit the real-time video streaming packets of eMBMSs proactively and probabilistically, so that the average handover delay perceived by a user is stochastically guaranteed. To quantify the tradeoff between the perceived handover delay and the bandwidth overhead of proactive transmissions, we develop an analytical model considering user mobility, user distribution, and session popularity. Comprehensive simulation is carried out to verify the analysis.

On the other hand, hypertext transfer protocol (HTTP) based adaptive streaming (HAS) is expected to be a dominant technique for non-real-time video delivery in LTE networks. In this dissertation, we first analyze the root causes of the problems of the existing HAS techniques. Based on the insights gained from our analysis, we propose a network-side HAS solution to provide a fair, efficient, and stable video streaming service. The key characteristics of our solution are: (i) unification of video- and data-users into a single utility framework, (ii) direct rate control conveying the assigned rates to the video client through overwritten HTTP Response messages, and (iii) rate allocation for stability by a stateful approach. By the experiments conducted in a real LTE femtocell network, we compare the proposed solution with state-of-the-art HAS solutions. We reveal that our solution (i) enhances the average video bitrates, (ii) achieves the stability of video quality, and (iii) supports the control of the balance

between video- and data-users.

Keywords : Network planning, Video, Streaming, eMBMS, MBSFN, HAS, LTE

Student Number : 2006-21263

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

The rapid and widespread deployment of broadband wireless networks has raised the expectation of high-quality video services in mobile environments. However, supporting bandwidth-intensive video applications requires efficient handling of network resources. When many users want to receive the same real-time video content (e.g., news or live sport) simultaneously, even high-bandwidth wireless link resources would fall short if a separate point-to-point channel is required for each user. The need for resource- and cost-efficient delivery of video content to many users has motivated the Third Generation Partnership Project (3GPP) to support an efficient network-wide video multicast service [1].

The 3GPP defined multimedia broadcast/multicast service (MBMS) to optimize the distribution of video traffic [2]. This standard covers the terminal, radio, core network, and user service aspects. This MBMS standard has evolved into enhanced MBMS (eMBMS) that builds on top of the 3GPP Long Term Evolution (LTE) standard. The eMBMS evolution brings improved performance thanks to higher and more flexible LTE bit rates and single frequency network (MBSFN) operations [3].

With the eMBMS support, the LTE network can extend the number of the services offered to the end-users and utilizes its resources in a more efficient way, but delay-sensitive applications like real-time video and audio streaming require the reduced handover delays, so that the disruption in the meantime can be ignorable or

tolerable to mobile users.

One approach to overcome this is deploying an 'MBSFN Area', which is defined by the 3GPP [1]. The MBSFN area is a specific region allocated by a group of cells transmitting the same video content, which allows all User Equipments (UEs) to use the same multicast bearer connnection and security keys during handovers within the same MBSFN area. As a result, the UE can receive the eMBMS packets while moving within the MBSFN area; its handover delay can be minimized. However, since every Evolved Node B (eNB) in the same MBSFN area broadcasts the same packets regardless of the presence of a user, the wireless link bandwidth can largely be wasted. Seeing that service disruption and wireless bandwidth usage are necessarily conflict in MBSFN area planning, we were motivated to study how to determine the best size of MBSFN areas.

On the other hand, video-on-demand (VoD) over LTE networks has become an immensely popular service in recent years. In VoD services, UEs are serviced individually by allocating and dedicating a transmission channel and a set of radio resources to each UE. Thus, an application-layer solution is gaining attention for VoD services over LTE networks. The hypertext transfer protocol (HTTP) based adaptive streaming (HAS) has become one of the most cost-effective solutions in delivering video content due to the abundance of Web platforms, and received great attention from both industry and research communities [4–7]. However, the performance of HAS solutions are rarely evaluated in LTE networks.

In this dissertation, we have focused on the MBSFN and the HAS technique enhancement issues to improve the performance of the real-time and non-real-time video delivery in LTE networks. First, we propose an MBSFN architecture based on

location management areas (LMAs), in order to save wireless link bandwidth while keeping the service disruption time at an acceptable level. We also studied how to decide MBSFN area and LMA sizes, which can make the best use of bandwith in maintaining the quality of eMBMS services. Second, we also propose to transmit eMBMS packets proactively and probabilistically to stochastically bound the average service disruption time for an eMBMS user. Third, we investigate the problems of the current HAS techniques, such as unfairness and unstability. To provide a fair, efficient, and stable video delivery, we propose a novel network-side HAS solution that optimizes the total utility of all users in a cell including video- and data-users, while maintaining the stable video quality.

The remainder of this dissertation is organized as follows. Chapter 2 introduces a novel scheme that improves the performance of real-time multicast video delivery in LTE networks. Chapter 3 presents a proactive transmission approach based on the framework discussed in Chapter 2. An improved HAS technique for non-real-time video delivery in LTE networks is introduced in Chapter 4. The summary and future work are briefly described in Chapter 5.

Chapter 2

Performance Improvements on Real-time Multicast Video Delivery

2.1 Introduction

Long Term Evolution (LTE) systems have gained much attention for high transmission rates in cellular environments. One of the promising applications that can leverage high bit rates is real-time video service based on the enhanced Multimedia Broadcast and Multicast Service (eMBMS) [1, 2]. Although the expectation of IP-based real-time multimedia streaming services in mobile environments is rising, QoS provisioning for those delay sensitive services is still challenging since mobility should be supported without intolerable delay.

One crucial issue to be addressed in the eMBMS architecture is the provision of seamless multimedia streaming to mobile receivers [8]. That is, a User Equipment (UE) should be able to receive a multimedia stream without noticeable disruption while it is moving across cells, even though a handover process is required when a UE moves from one cell of a cellular network to another. During the handover process, the path to the UE is transferred from the serving cell to the target cell, and the time required for this process is referred to as the handover delay. LTE normally performs hard handovers, in which all connections to the serving cell are broken before new connections are made to the target cell. As a result, packets being sent

through the serving cell during the handover may not be delivered to the UE[1]. In the case of unicast traffic, this so-called 'service disruption' can be overcome by packets being stored and forwarded from the serving cell to the target cell for a fast handover and retransmission [9, 10]. However, eMBMSs cannot rely on such techniques since packets are destined for multiple receivers. Therefore, the handover delay in eMBMS should be minimized.

LTE eMBMS introduces a new transmission scheme called Multimedia Broadcast and multicast service over a Single Frequency Network (MBSFN). In MBSFN operation, eMBMS data is transmitted simultaneously over the air from multiple tightly time-synchronized cells. A group of those cells which are targeted to receive the broadcast eMBMS data constitute a so called *MBSFN area* [1]. This allows not only better signal reception at UEs, but also the handover delay reduction. A handover between eNBs in the same MBSFN area involves a reduced delay because packets with the same content will be received from the target eNB immediately after the completion of a link-level handover. However, a handover that crosses a boundary between different MBSFN areas requires not only link-level handover signaling but also eMBMS-related signaling which takes a much longer time.

Obviously, larger MBSFN areas will yield a better quality of service for a given level of mobility, but handover delay cutback comes at the cost of amplified traffic. This wastes the link capacity of the air interface[2], because every eNB in the same MBSFN area broadcasts the same eMBMS packets regardless of the presence of a user in its coverage; moreover, requesting a new eMBMS session can be blocked due

[1]The arrival of eMBMS packets cannot be guarantted, since an unacknowledged transfer mode is used for LTE eMBMS.

[2]An MBSFN area will also waste the link bandwidth of the wired backhaul network, but we focus on the wireless link.

to lack of available bandwidth. Seeing that service disruption and wireless bandwidth usage are necessarily conflict in MBSFN area planning, we were motivated to study how to determine the best size of MBSFN areas.

In this chapter, we propose an MBSFN architecture based on location management areas (LMAs), each of which is a set of geographically adjacent BSs within an MBSFN area. Then multicast and broadcast packets only need to be transmitted to the LMAs which have eMBMS users reducing the requirement for wireless link bandwidth. Using LMAs allows large MBSFN areas to be used, so that the number of inter-MBSFN area handovers can be reduced; in this way we can reduce the average handover delay without too much bandwidth waste. We analyze the performance of our LMA-based MBSFN scheme by means of an analytical model for different sizes of MBSFN areas and LMAs. We go on to propose how the sizes of the MBSFN areas and LMAs can be determined while retaining an acceptable service disruption time. This is a comprehensive analytical study for MBSFN area planning; our model of the service disruption time, bandwidth usage, and blocking probability in terms of user mobility, distribution, and session popularity is novel, as is our approach to determining the sizes of the MBSFN areas and LMAs.

The rest of this chapter is organized as follows. In Section 2.2, we summarize related works in the literature and highlight the major differences between existing works and our work. Section 2.3 presents our LMA-based MBSFN area planning scheme and explains how it affects handover delay and a performance analysis follows in Section 2.4. Numerical and simulation results are presented in Sections 2.5 and 2.6, respectively, and Section 2.7 concludes this chapter.

2.2 Related Work

The UMTS multicast architecture [11] employs standard IP multicast protocols. A multicast mechanism for UMTS has been proposed [12] which establishes multicast tunnels throughout the UMTS network that allow multicast packets to be transferred on shared links toward multiple destinations. The tradeoffs between the broadcast, multiple unicast, and multicast approaches for one-to-many packet delivery services in UMTS have been investigated [13]. But this work lacks an analysis of user mobility handling; it is assumed that mobility is handled by the standard UMTS mobility mechanisms which are similar to conventional unicast packet forwarding. As an alternative [14], routing lists can be introduced into the nodes of the UMTS to support resource-efficient multicast transmissions combined with a reassessment of the handover types and the mobility management mechanism in UMTS. However, multicast service continuity still cannot be assured unless handover delay issues are taken into account.

Mobile WiMAX has a Multicast and Broadcast Service (MBS) zone which is similar with the MBSFN area in LTE eMBMS [15]. There has been some research on the support of real-time services such as voice over IP and video streaming (IPTV) over WiMAX [16, 17]. The hard handovers mandated by IEEE 802.16e make seamless mobility with imperceptible interruption of service difficult to achieve in Mobile WiMAX. A fast handover scheme has been proposed [18] along with a new transport connection identifier (CID) mapping strategy for real-time applications in order to reduce handover delay and the probability of packet loss. This approach could be an option for unicast services (e.g., video on demand), but it is not suitable for multi-

cast and broadcast services. The efficient delivery of video broadcasts over WiMAX has been studied [19], especially the issue of synchronous transmission over multiple base stations. The effectiveness of data delivery in intra-MBS zone operations is shown to be improved by macro diversity [20], and hence seamless handover is also feasible. But this work lacks an analysis of the effects of the various MBS zone sizes, and the mobility scenario that goes out of an MBS zone coverage is not considered at all.

There also have been efforts on investigating the performance of MBSFN, especially in the optimal resource allocation, overlapping MBSFN areas, and transmission cost analysis [21–23]. Each of the above issues can be boiled down to an MBSFN area planning criterion. This work is a comprehensive analytical study of MBSFN area planning which has the aim of reducing both bandwidth usage and service disruption. Nevertheless, some previous studies have addressed the issues of network planning for wireless multicast and broadcast services. An efficient multicast mechanism for heterogeneous wireless networks has been proposed [24] which reduces the total bandwidth requirement of the IP multicast tree by adaptively selecting the cell and the wireless technology for each mobile stations. Although this is not suitable for multicast services in a homogeneous wireless network, it does allow more mobile stations to cluster together, and leads to the use of fewer cells, thus saving bandwidth. A network operator might use this approach for network planning in small fixed or nomadic wireless networks, but it is impractical for large mobile wireless networks in which the network frequently needs to recompute its low-cost multicast tree due to mobility. In another multicasting mechanism for UMTS [25], multicast packets are distributed to location areas (LAs) which are groups of cells. When a mobile station

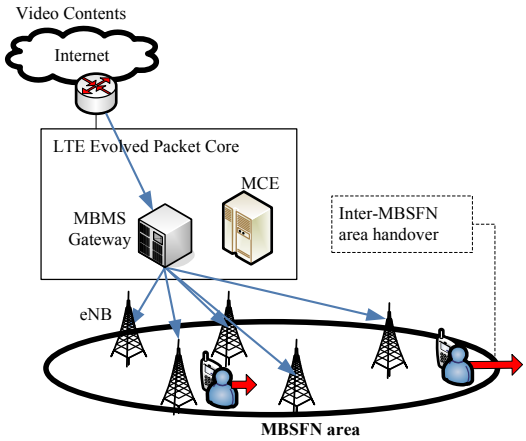


Fig. 2.1. An illustration of MBSFN.

moves between LAs, the change is reported to a location server so that the position of the mobile station is tracked for paging purposes. This scheme's primary concern is with the delivery of short messages to multiple users in these LAs to minimize paging cost, but the solution proposed includes the location tracking of multicast receivers which is relevant to our work.

2.3 Location Management Area Based MBSFN

To support MBSFN, several eNBs (normally adjacent to each other) construct an MBSFN area between them as shown in Fig. 2.1, which is managed by the MBMS Gateway and multi-cell/multicast coordination entity (MCE). Once MBSFN area sizes are determined (refer to Section 2.3.3), MBS areas can be deployed by a network operator and all eNBs in an MBSFN area have a shared multicast bearer connection

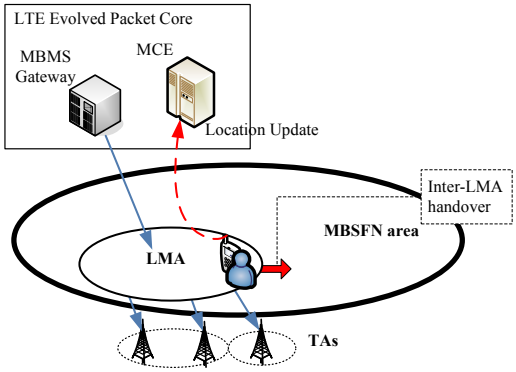


Fig. 2.2. An illustration of LMA-based MBSFN.

for the same multicast transmission[3]. Therefore, a UE does not need to create a new bearer connection during handovers between eNBs in the same MBSFN area which reduces the handover delay. Accordingly, increasing the sizes of MBS areas will reduce service disruptions for a given level of mobility while wireless link bandwidth will be wasted in the sense that every eNB in the same MBSFN area broadcasts the same eMBMS packets regardless of the presence of a user. This also results in increasing the probability of blocking a new eMBMS session.

2.3.1 Location Management Area (LMA)

To efficiently balance the tradeoff between bandwidth usage and service disruption in MBSFN, we define a location management area (LMA) as a set of geographically adjacent eNBs which is used to track the location of eMBMS users. That is, the network is always aware of the current LMA of each eMBMS user. Then, we introduce an LMA-based MBSFN that partitions an MBSFN area into multiple

[3]An eNB transmits its MBSFN area identifier(s) using the System Information Block (SIB) type 13 [1].

LMAs and then selectively transmits packets to the LMAs in which eMBMS users currently reside. Using LMAs decouples the requirement for wireless link bandwidth from the size of an MBSFN area. This allows for large MBSFN areas to be used, so that service disruptions will be reduced as shown in Fig. 2.2.

Our scheme relies on the network keeping track of the location of every ('eMBMS-enabled') UE at the granularity of an LMA. In LTE, a tracking area (TA) [1], which is analogous to the location area (LA) in 3G cellular networks, is used to track the locations of UEs. A LTE Evolved Packet Core (EPC) manages information that tracks which UEs are currently located in each TA. An LMA might correspond to one or more TAs, but it is possible that the coverage of an LMA is determined independently of those of TAs.

The location of a UE in normal mode is tracked from one eNB to another by the EPC. Whenever a handover is done, the target eNB reports this event to the EPC, which therefore knows the current eNB and TA of each UE. In idle mode[4] however, the location of a UE is handled at a coarser level. A UE, even in idle mode, can acquire the TA information of current and neighboring cells as long as it receives eMBMS packets from an eNB. Therefore each time an idle UE crosses a TA boundary, it can inform the EPC of its new TA. To flexibly determine the coverage of an LMA, we suggest that every eNB should broadcast an LMA information as well as the TA information. This enables a UE in idle mode to update its location upon every inter-LMA movement. Since the MCE controls which LMAs will transmit eMBMS packets, it needs to receive the locations of eMBMS users from the EPC.

[4]This mode is defined in 3GPP to conserve power and network resources. A UE in idle mode performs no handover when it crosses a cell boundary, but it can receive eMBMS packets [1].

2.3.2 Handover Delays

The handover of an eMBMS session involves delays due to the link-level messages exchanged during the LTE handover and due to the eMBMS signaling messages. The former delay occurs whenever a UE with an ongoing eMBMS session switches to a new eNB irrespective of its MBSFN area or LMA. However, the latter delay only occurs when a UE moves from one MBSFN area to another. First, eMBMS handovers can therefore be classified into *intra-MBSFN area handovers* and *inter-MBSFN area handovers*.

As shown in Fig. 2.1, the MBMS Gateway transmits eMBMS packets to all eNBs within an MBSFN area, using IP multicasting. Note that MBSFN areas which currently contain users of the current eMBMS session are called *active* MBSFN areas, whereas those without such users are called *inactive* MBSFN areas. If the target MBSFN area is already active, the IP multicast distribution tree does not have to be updated.

In an LMA-based MBSFN, intra-MBSFN area handovers are sub-classified into: intra-LMA handovers which result from a change of eNB within the same LMA; and inter-LMA handovers between eNBs in different LMAs. Similarly, LMAs containing current users of a particular eMBMS session are called *active* LMAs with remainder called *inactive* LMAs, with respect to that session.

2.3.3 LMA-based MBSFN Area Planning

Seeing that service disruption and wireless bandwidth usage are necessarily conflicting when deciding the sizes of MBSFN areas, we now formally define the MBSFN area planning problem. Let N_Z be the number of cells in an MBSFN area and

Algorithm 2.1 Algorithm to determine N_Z and N_L.

```
// Assumption: N_{Z,max} ≥ N_{Z,min} ≥ N_{L,min} ≥ 1.
N_L = N_{L,min}.
while N_L ≤ N_{Z,max} do
    N_Z = max{N_{Z,min}, N_L}
    while N_Z ≤ N_{Z,max} do
        if Pr[Y > D_th]_i ≤ δ then
            return (N_Z, N_L).
        end if
        increase N_Z.
    end while
    increase N_L.
end while
```

N_L be the number of cells in an LMA. We present a method that determines the values of N_Z and N_L which minimize the bandwidth usage while keeping the average handover delay below a specified value.

Let Y be the random variable expressing the length of a handover delay, and D_{th} be a threshold value of the handover delay above which an eMBMS user experiences a noticeable session disruption. The probability that receiving session i is disrupted, such that Y exceeds D_{th}, can be written as $Pr[Y > D_{th}]_i$. How to compute the probability will be elaborated in Section 2.5.2.

The problem of MBSFN area planning for session i can now be considered as a search problem for values of N_Z and N_L which satisfy $Pr[Y > D_{th}]_i \leq \delta$, where δ is the tolerable disruption ratio of session i[5]. We assume that a service provider enforces the MBSFN area and LMA size constraints by introducing min/max bounds: $N_{Z,min}$ is the minimum number of cells in an MBSFN area; $N_{Z,max}$ is the maximum number of cells in an MBSFN area; $N_{L,min}$ is the minimum number of cells in

[5]Different sessions may have different values of N_A and N_L, and different MBSFN areas can be overlapped.

an LMA. Then we present an algorithm in Algorithm 2.1 which determines N_A and N_L. It takes the initial results of the MBSFN area planning produced by minimizing the bandwidth usage and tries to reduce the disruption probability by increasing the number of cells in MBSFN areas and LMAs[6]. This process is repeated until the disruption probability is below a specified value δ. Implementing a dynamic MBSFN system may be possible by running the algorithm periodically at MCE. However, since it will incur large overhead due to the synchronous transmission and scheduling problems [19] in a dynamic set of eNBs, we will only consider a static MBSFN area system; once MBSFN areas and LMAs are planned, the configuration will not be changed dynamically.

2.4 Performance Analysis

We will now analyze the service disruption time and bandwidth usage in the LMA-based MBSFN Scheme (denoted by LMS), while considering user distribution, mobility and eMBMS session popularity. We make the following assumptions and use the notations summarized in Table 2.1:

- The eMBMS session duration time follows an exponential distribution with mean $1/\lambda_s$.
- The total number of eMBMS sessions is S, and all sessions are ranked by popularity. Let β_i be the conditional probability that the ith most popular session is requested ($i = 1, 2, ..., S$), given that a request arrives. β_i is drawn from a

[6]If there is not any size or shape constraint for planning MBSFN areas and LMAs, "increase N_Z/N_L" means an increment by one.

Table. 2.1. Notation for MBSFN performance analysis.

λ_c	cell crossing rate
$\lambda_z(\lambda_l)$	MBSFN area (LMA) crossing rate
λ_s	eMBMS session service rate
S	total number of eMBMS sessions
m	number of sessions that can be transmitted simultaneously over a wireless link
α	Zipf-like distribution exponent
ρ^*	average number of users per unit area
ρ_i	average number of users per unit area of session i
$A_z(A_l)$	area of an MBSFN (an LMA)
$Z_{h,i}$	number of MBSFN area handovers of session i ($h = 1$: inter-MBSFN area handover moving to inactive areas, $h = 2$: inter-MBSFN area handover moving to active areas, $h = 3$: intra-MBSFN area handover)
$L_{h,i}$	number of LMA handovers of session i ($h = 1$: inter-LMA handover moving to inactive LMAs, $h = 2$: inter-LMA handover moving to active LMAs, $h = 3$: intra-LMA handover)
D_{Zh}	delay of an MBSFN area handover ($h = 1$: inter-MBSFN area handover moving to inactive areas, $h = 2$: inter-MBSFN area handover moving to active areas, $h = 3$: intra-MBSFN area handover)
D_{Lh}	delay of an LMA handover ($h = 1$: inter-LMA handover moving to inactive LMAs, $h = 2$: inter-LMA handover moving to active LMAs, $h = 3$: intra-LMA handover)

cut-off Zipf-like distribution [26], and is given by

$$\beta_i = \frac{\Omega}{i^{\alpha}}, \qquad \text{where } \Omega = \left(\sum_{i=1}^{S} \frac{1}{i^{\alpha}} \right)^{-1}, \quad 0 < \alpha \leq 1. \tag{2.1}$$

- The spatial distribution of eMBMS users follows a two-dimensional Poisson distribution [27] with net rate ρ^*, which is defined as the average number of users per unit area: $\rho^* = \lambda^*/\mu^*$, where λ^* is the users' arrival rate and μ^* is the number of users leaving per second. Therefore, the probability that x users appear in an area A is $(\rho^* A)^x e^{-\rho^* A}/x!$. From (2.1), the average number of users of the ith most popular session per unit area is $\rho_i = \beta_i \rho^*$.

- The sizes of cells are identical in the network. For each session, the sizes of MBSFN areas (and LMAs) are independent and identically distributed (i.i.d.). Let Z, L, and C be random variables representing the numbers of MBSFN area crossings (i.e., inter-MBSFN area handovers) per session, LMA crossings (i.e., inter-LMA handovers) per session, and cell crossings (i.e., the total number of handovers) per session, respectively.

- The residence times in an MBSFN area, an LMA, and a cell follow Gamma distributions with the mean $1/\lambda_z$ (variance V_z), $1/\lambda_l$ (variance V_l), and $1/\lambda_c$ (variance V_c), respectively[7]. The Gamma distribution is widely employed to model UE movement in many studies [28], [29], [30]. For each session, the residence times in a MBSFN area, LMA, and cell are independent and identically distributed.

[7]Since the values of $1/\lambda_z$ and $1/\lambda_l$ may differ for different sessions, $1/\lambda_{z,i}$ and $1/\lambda_{l,i}$ will be exact expressions. For the sake of simplicity, however, we skip the subscript i.

2.4.1 Disruption Time

The service disruption time for an eMBMS user is defined as the sum of all handover delays during the service time of an eMBMS session. For MBSFN areas, there are three types of handover: inter-MBSFN area handovers in which a UE moves to an inactive MBSFN area, and inter-MBSFN area handovers in which a UE moves to an active MBSFN area, and intra-MBSFN area handovers. Let Z_1 and Z_2 respectively be the numbers of inter-MBSFN area handovers to inactive and active MBS zones, and let Z_3 be the number of intra-MBSFN area handovers. Then $E[Z] = E[Z_1] + E[Z_2]$ and $E[C] = E[Z] + E[Z_3]$. The average service disruption time for the ith most popular session can be expressed as

$$\text{Average Disruption Time(i)} = E[Z_{1,i}] \cdot D_{Z1} + E[Z_{2,i}] \cdot D_{Z2} + E[Z_{3,i}] \cdot D_{Z3}, \quad (2.2)$$

where D_{Z1}, D_{Z2} and D_{Z3} are the unit delays for an inter-MBSFN area handover to an inactive MBSFN area, an inter-MBSFN area handover to an active MBSFN area, and an intra-MBSFN area handover, respectively.

Since the LMS partitions an MBSFN area into LMAs, the intra-MBSFN area handovers can be subclassified into three types of LMA handover: inter-LMA handovers in which a UE moves to an inactive LMA, inter-LMA handovers in which a UE moves to an active LMA, and intra-LMA handovers. Let L_1 and L_2 respectively be the numbers of inter-LMA handovers to inactive LMAs and to active LMAs, and let L_3 be the number of intra-LMA handovers. Then, we have $E[Z_3] = E[L_1] + E[L_2] + E[L_3]$. From (2.2), the average service disruption time for the ith most pop-

ular session in LMS, $T_{LMS,i}$, can be expressed as

$$\begin{aligned} T_{LMS,i} &= E[Z_{1,i}] \cdot D_{Z1} + E[Z_{2,i}] \cdot D_{Z2} + E[L_{1,i}] \cdot D_{L1} \\ &+ E[L_{2,i}] \cdot D_{L2} + E[L_{3,i}] \cdot D_{L3} \end{aligned} \tag{2.3}$$

where D_{L1}, D_{L2} and D_{L3} respectively are the unit delays for an inter-LMA handover to an inactive LMA, an inter-LMA handover to an active LMA, and an intra-LMA handover.

Let $p(x, i, A)$ denote the probability that there are x users subscribing the ith most popular session in an area A, so that $p(x, i, A) = (\rho_i A)^x e^{-\rho_i A}/x!$. The probability that there is no user subscribing to the ith most popular session in an MBSFN area with area A_z is given by $p\,(0, i, A_z) = e^{-\rho_i A_z}$. Let $\Pr(Z_1 = j|Z = n)$ be the conditional probability that there are n inter-MBSFN area handovers, among which j handovers are to inactive MBSFN areas. It follows a Bernoulli distribution and can be expressed as $\Pr(Z_1 = j|Z = n) = \binom{n}{j} [p\,(0, i, A_z)]^j [1 - p\,(0, i, A_z)]^{n-j}$.

The probability $\Pr(Z = n)$ can be obtained by using the results in [30], that is

$$\Pr(Z = n) = \begin{cases} 1 - \frac{\lambda_z(1 - f_z^*(\lambda_s))}{\lambda_s}, & n = 0 \\ \frac{\lambda_z}{\lambda_s} [1 - f_z^*(\lambda_s)]^2 [f_z^*(\lambda_s)]^{n-1}, & n > 0 \end{cases} \tag{2.4}$$

where $f_z^*(s) = [\lambda_z \gamma/(s + \lambda_z \gamma)]^\gamma$ is the Laplace-Stieltjes transform of a Gamma random variable with a parameter $\gamma = 1/(V_z \lambda_z^2)$. The average number of MBSFN area crossings can be computed as $E[Z] = \sum_{n=0}^{\infty} n\Pr(Z = n) = \lambda_z/\lambda_s$.

Then, the average number of inter-MBSFN area handovers to inactive MBSFN

areas for session i (i.e., $E[Z_{1,i}]$) can be expressed as

$$\begin{aligned} E[Z_{1,i}] &= \sum_{n=0}^{\infty}\sum_{j=0}^{n} j \cdot \Pr(Z_1 = j | Z = n) \cdot \Pr(Z = n) \\ &= \sum_{n=1}^{\infty}\sum_{j=1}^{n} \frac{j\lambda_z}{\lambda_s} \Pr(Z_1 = j | Z = n) \\ &\qquad \times [1 - f_z^*(\lambda_s)]^2 [f_z^*(\lambda_s)]^{n-1} \\ &= \frac{\lambda_z}{\lambda_s} p\left(0, i, A_z\right). \end{aligned} \tag{2.5}$$

From $\Pr(Z_2 = j | Z = n) = \binom{n}{j} \left[1 - p\left(0, i, A_z\right)\right]^j \left[p\left(0, i, A_z\right)\right]^{n-j}$, the average number of inter-MBSFN area handovers to active MBSFN areas (i.e., $E[Z_{2,i}]$) can be expressed as

$$\begin{aligned} E[Z_{2,i}] &= \sum_{n=0}^{\infty}\sum_{j=0}^{n} j \cdot \Pr(Z_2 = j | Z = n) \cdot \Pr(Z = n) \\ &= \frac{\lambda_z}{\lambda_s} \left(1 - p\left(0, i, A_z\right)\right). \end{aligned} \tag{2.6}$$

Recall that L is the random variable representing the number of LMA crossings, and then $E[L] = E[Z] + E[L_1] + E[L_2]$. By a similar derivation from (2.4), the LMA crossing probability $\Pr(L = n)$ can also be expressed as

$$\Pr(L = n) = \begin{cases} 1 - \frac{\lambda_l (1 - f_l^*(\lambda_s))}{\lambda_s}, & n = 0 \\ \frac{\lambda_l}{\lambda_s} \left[1 - f_l^*(\lambda_s)\right]^2 \left[f_l^*(\lambda_s)\right]^{n-1}, & n > 0 \end{cases}$$

where $f_l^*(s) = [\lambda_l \gamma / (s + \lambda_l \gamma)]^{\gamma}$ and $\gamma = 1/(V_l \lambda_l^2)$. Then, the average number of LMA crossings can be computed as $E[L] = \sum_{n=0}^{\infty} n \Pr(L = n) = \lambda_l / \lambda_s$. Since $E[L_1] + E[L_2] = E[L] - E[Z]$, the average number of inter-LMA handovers that do

not involve changing MBSFN areas is given by $E[L] - E[Z] = (\lambda_l - \lambda_z)/\lambda_s$. The probability that there is no user for session i in an LMA with area A_l is $p\left(0, i, A_l\right) = e^{-\rho_i A_l}$. So, $E[L_{1,i}]$ and $E[L_{2,i}]$ can be derived as

$$E[L_{1,i}] = \frac{\lambda_l - \lambda_z}{\lambda_s} \cdot p\left(0, i, A_l\right), \tag{2.7}$$

$$E[L_{2,i}] = \frac{\lambda_l - \lambda_z}{\lambda_s} \left(1 - p\left(0, i, A_l\right)\right). \tag{2.8}$$

We can also derive the average number of cell crossings, $E[C] = \sum_{n=0}^{\infty} n\text{Pr}(C = n) = \sum_{n=1}^{\infty}(n\lambda_c/\lambda_s)[1 - f_c^*(\lambda_s)]^2[f_c^*(\lambda_s)]^{n-1} = \lambda_c/\lambda_s$. Since $E[L_3] = E[C] - E[L]$, the average number of intra-LMA handovers can be written as

$$E[L_{3,i}] = \frac{\lambda_c - \lambda_l}{\lambda_s}. \tag{2.9}$$

From (2.3), (2.5), (2.6), (2.7), (2.8) and (2.9), we can write

$$\begin{aligned} T_{LMS,i} &= \frac{\lambda_z}{\lambda_s}\left[e^{-\rho_i A_z} \cdot D_{Z1} + \left(1 - e^{-\rho_i A_z}\right) \cdot D_{Z2}\right] \\ &+ \frac{(\lambda_l - \lambda_z)}{\lambda_s}\left[e^{-\rho_i A_l} \cdot D_{L1} + \left(1 - e^{-\rho_i A_l}\right) \cdot D_{L2}\right] \\ &+ \frac{(\lambda_c - \lambda_l)}{\lambda_s} \cdot D_{L3}. \end{aligned} \tag{2.10}$$

2.4.2 Bandwidth Usage

The bandwidth usage for a particular session is defined as the ratio of the number of cells transmitting multicast packets of the session to the total number of cells in the network. We define $U_{LMS,i}$ be the bandwidth usages in LMS, which can be expressed as

$$U_{LMS,i} = (1 - p(0, i, A_l)) = 1 - e^{-\rho_i A_l}. \tag{2.11}$$

2.4.3 Blocking Probability

When a UE requesting a particular session i cannot receive its content, we say the session i is blocked. The blocking probability $B_{LMS,i}$ is defined to be the probability that an attempt to request session i fails due to the lack of capacity in a cell. We assume that the blocking only occurs on a wireless link with finite capacity of the cell; that is the maximum number of sessions that can be transmitted simultaneously over a wireless link is denoted by m. However, requesting an already ongoing session is always accepted. Therefore, $B_{LMS,i}$ can be written

$$B_{LMS,i} = \pi_{0,i} \cdot B^m_{LMS,i}, \tag{2.12}$$

where $\pi_{0,i}$ is the steady state probability for session i to be not on air (or be in inactive state) and $B^m_{LMS,i}$ is the probability that the wireless link is consumed by m sessions other than requested session i [31].

To calculate $\pi_{0,i}$, we consider a Markov chain that alternates between two states, *on* and *off*. Note that $p(0, i, A_l)$ means the ratio of *off*-state cells in the network for session i, where A_l is the area of an LMA. If we call *off* state 0 and *on* state 1, the transition probability matrix is

$$\left(\begin{array}{c|c} B^m_{LMS,i}(1 - p(0, i, A_l)) & (1 - B^m_{LMS,i}) \\ +p(0, i, A_l) & \times(1 - p(0, i, A_l)) \\ \hline p(0, i, A_l) & 1 - p(0, i, A_l) \end{array}\right).$$

Therefore, the steady state probability to be in inactive state can be expressed as

$$\pi_{0,i} = \frac{p(0,i,A_l)}{1 - B^m_{LMS,i}(1 - p(0,i,A_l))}. \tag{2.13}$$

By combining (2.12) and (2.13), we obtain $B_{LMS,i}$ which is the blocking probability of session i in LMS:

$$B_{LMS,i} = \frac{B^m_{LMS,i} \cdot e^{-\rho_i A_l}}{1 - B^m_{LMS,i}(1 - e^{-\rho_i A_l})}. \tag{2.14}$$

Since our MBS system has S available sessions and m admitted sessions, $B^m_{LMS,i}$ can be modeled by an $M/M/m/m/S$ system which is referred to as the Engset system [32]. The sessions in LMS represent the users in an Engset system. In a generalized Engset system, the users are not identical; their arrival rates (λ_i) and departure rates (μ_i) as well as the requested resources (c_i) can be different. For an Engset system with capacity C, the user blocking probability of user i is $B^C_i = (\sum_{j=C-c_i+1}^{C} \pi^{(i)}_j)/(\sum_{j=0}^{C} \pi^{(i)}_j)$ where $\pi^{(i)}_j$ is the probability that j capacity units are occupied in an infinite system with user i removed. The probability $\pi^{(i)}_j$ can be calculated from the probability generating function: $P_i(z) = \sum_{j=0}^{\infty} \pi^{(i)}_j z^j = \prod_{k \in S-\{i\}}(q_k + p_k z^{c_k})$ where $q_k = e^{-\lambda_k/\mu_k} = 1 - p_k$.

In LMS, each session has different ρ_i that represents the average number of UEs staying in the system for session i. And every session requests the same amount of resources (i.e., $c_i = 1$ and $C = m$). Then, the blocking probability can be expressed as $B^m_i = \pi^{(i)}_m / \sum_{j=0}^{m} \pi^{(i)}_j$, where $\pi^{(i)}_j$ is the probability that the capacity is occupied by j sessions in an infinite system with session i removed. And, the probability generating function is $P_i(z) = \sum_{j=0}^{\infty} \pi^{(i)}_j z^j = 1/(q_i + p_i z) \prod_{k=1}^{S}(q_k + p_k z)$ where

Table. 2.2. eMBMS Handover Delays.

D_{Z3}	D_{L3}	(link-level handover delay)
	D_{L2}	(link-level handover delay)
	D_{L1}	(link-level handover delay) + (multicast distribution updating)
D_{Z2}		(link-level handover delay) + (eMBMS session restarting)
D_{Z1}		(link-level handover delay) + (multicast distribution updating) + (eMBMS session restarting)

$q_k = e^{-\rho_k A_l}$ in LMS. Since we have $\frac{d^j}{dz^j} \sum_{j=0}^{\infty} \pi_j^{(i)} z^j|_{z=0} = (j!)\pi_j^{(i)}$, the probability $\pi_j^{(i)}$ can be expressed as $\pi_j^{(i)} = \frac{1}{j!} \frac{d^j}{dz^j} P_i(z)|_{z=0}$. Therefore, $B_{LMS,i}^m$ can be computed by

$$B_{LMS,i}^m = \frac{\left(\frac{1}{m!}\right) \frac{d^m}{dz^m} \left[\frac{\prod_{k=1}^{S} \left\{ e^{-\rho_k A_l + z(1-e^{-\rho_k A_l})} \right\}}{e^{-\rho_i A_l + z(1-e^{-\rho_i A_l})}} \right]_{|z=0}}{\sum_{j=0}^{m} \left(\frac{1}{j!}\right) \frac{d^j}{dz^j} \left[\frac{\prod_{k=1}^{S} \left\{ e^{-\rho_k A_l + z(1-e^{-\rho_k A_l})} \right\}}{e^{-\rho_i A_l + z(1-e^{-\rho_i A_l})}} \right]_{|z=0}}.$$

2.5 Numerical Results

We will now evaluate the performance of LMS in terms of service disruption time, wireless link bandwidth usage, and blocking probability. To analyze disruption time, we need to quantify each kind of handover delay by identifying the major handover steps: link-level handover delay, eMBMS signaling delay and multicast distribution update. We assume that the link-level delay is 100 msec, the IP multicast distribution updating takes 200 msec, and the eMBMS session restarting is 300 msec[8]. Table 2.2 lists all eMBMS handover delays which are decomposed into those steps.

[8]Since MBSFN area infomation is delivered every 320 msec via System Infomation Block type 13, we assume that eMBMS session restarting takes at least 300 msec including eMBMS signaling exchanges.

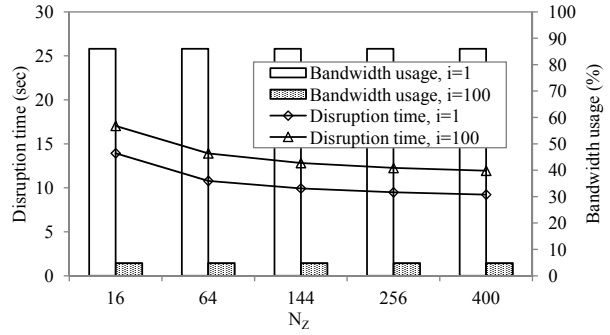
For the sake of simplicity, it is assumed that MBSFN areas, LMAs, and cells are circular or square-shaped, and that there are N_Z cells in an MBSFN area and N_L cells in an LMA. We also assume each cell has an area of 1 km^2. We can then use a fluid flow mobility model to express the cell boundary crossing rate as $\lambda_c = (16v)/(\pi l)$, where v is the average velocity of UEs and l is the length of the perimeter of a cell. This allows us to approximate λ_z and λ_l by $\lambda_c/\sqrt{N_Z}$ and $\lambda_c/\sqrt{N_L}$ respectively [33]. The average duration of a session, $1/\lambda_s$, is set to 60 minutes, and the total number of eMBMS sessions, S, is set to 100. Comparing to the original MBSFN scheme without applying LMAs (denoted by OMS), we will use the following notations to identify each area planning scheme:

- OMS(N_Z): the original method of MBSFN area planning with N_Z-cell MBSFN areas.
- LMS(N_Z, N_L): LMA-based MBSFN area planning with N_Z-cell MBSFN areas which are partitioned into N_L-cell LMAs.

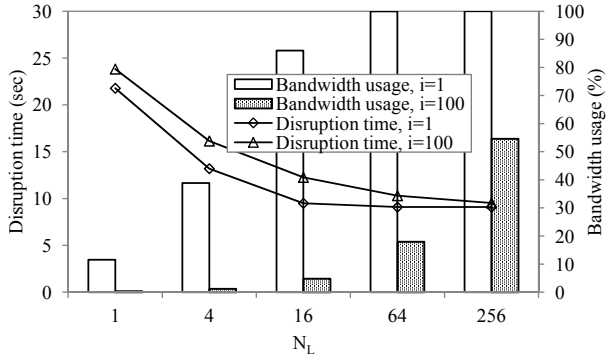
Note that OMS(k) can be modeled by LMS(k, k) for any value of k. Additionally, we will use the notations $T_{OMS,i}$, $U_{OMS,i}$, and $B_{OMS,i}$ for OMS(k) which can be derived from $T_{LMS,i}$, $U_{LMS,i}$, and $B_{LMS,i}$ for LMS(k, k), respectively.

2.5.1 Effect of N_Z and N_L

Fig. 2.3 shows the disruption time for LMS ($T_{LMS,1}$ and $T_{LMS,100}$) and its bandwidth usage ($U_{LMS,1}$ and $U_{LMS,100}$) as a function of N_Z or N_L when $\alpha = 0.8$, $\rho^* = 1$ user/cell, $v = 60$ km/h. In Fig. 2.3(a), the bandwidth usage is shown to be independent of the value of N_Z which is because eMBMS packets are selectively



(a) LMS(N_Z, 16): effect of N_Z



(b) LMS(256, N_L): effect of N_L

Fig. 2.3. Disruption time and bandwidth usage.

transmitted to active LMAs. But, the disruption time decreases as N_Z increases. As N_Z rises from 16 to 400, $T_{LMS,1}$ and $T_{LMS,100}$ can be reduced by 34% and 30%, respectively. Fig. 2.3(b) demonstrates the disruption time and bandwidth usage against N_L with $N_Z = 256$. Recall that N_L is a single unit of transmission for LMS while N_Z is the unit of transmission for OMS. Therefore the $U_{LMS,1}$ and $U_{LMS,100}$ curves for LMS(256,k) are exactly same as the $U_{OMS,1}$ and $U_{OMS,100}$ curves for OMS(k)

for any k. Since $U_{LMS,i}$ is independent of the MBSFN area size, we have

$$U_{LMS,i} \text{ of } LMS(n,k) = U_{OMS,i} \text{ of } OMS(k)$$
$$\leq U_{OMS,i} \text{ of } OMS(n), \quad n \geq k.$$

Unlike the bandwidth usage, the disruption time decreases as N_L increases because the number of UEs crossing LMAs during a session is reduced as the LMAs become larger. When N_L is 1 in Fig. 2.3(b) (i.e., LMS(256,1)), $T_{LMS,1}$ and $T_{LMS,100}$ have their highest values of 21.8 and 23.8, respectively. However, LMS(256,k) significantly outperforms OMS(k) for a small k in terms of disruption time while both of them use the same amount of bandwidth[9]. When k is large, LMS(256,k) and OMS(k) have a similar performance. In addition, the impact of session popularity (i.e., the distribution of users) on the disruption time becomes less significant, since more cells need to transmit the packets of an unpopular session.

With the same mobility, the number of inter-MBSFN area handovers in OMS(k) and the number of inter-LMA handovers in LMS(256,k) are expected to be equal, and the inter-LMA handover delay is much shorter than the inter-MBSFN area handover delay. Consequently, LMS(n,k) shows less disruption time than OMS(k) for all values of n and k ($n > k$). However, the disruption time for LMS(n,k) is worse than that for OMS(n) since it may include additional inter-LMA handover delays as well as the same level of inter-MBSFN area handover delays that occur with OMS(n). For example, LMS(256,16) causes 4.4% more delay than OMS(256) when $i = 1$. Based

[9] $T_{LMS,1} = 44.2$ and $T_{LMS,100} = 45.9$ for OMS(1). $T_{LMS,1} = 13.9$ and $T_{LMS,100} = 17.0$ for OMS(16).

on this observation, we can bound the range of $T_{LMS,i}$ for LMS(n,k) as

$$\begin{aligned} T_{OMS,i} \text{ of } OMS(n) &\leq T_{LMS,i} \text{ of } LMS(n,k) \\ &\leq T_{OMS,i} \text{ of } OMS(k), \quad n \geq k. \end{aligned}$$

The blocking probability in LMS is independent of the value of N_Z like the bandwidth usage since N_L is a single unit of transmission for LMS. For example, $B_{LMS,1}$ for LMS(n,16) is exactly the same as the $B_{OMS,1}$ for OMS(16) for all values of n ($n \geq 16$). Therefore, we have

$$B_{LMS,i} \text{ of } LMS(n,k) = B_{OMS,i} \text{ of } OMS(k), \quad n \geq k.$$

LMS decouples the requirement for wireless link bandwidth from the size of an MBSFN area to balance a tradeoff between the bandwidth usage and handover delay, compared to OMS. If we enforce the constraint that keeps the average handover delay below a specified value, LMS can provide a bandwidth-efficient solution. On the other hand, if we wish to ensure that the total bandwidth usage cannot exceed a specified value, it yields a less disruptive solution.

2.5.2 Deciding N_Z and N_L

Recall that the problem of MBSFN area planning for LMS is to search for the values of N_Z and N_L which satisfy $Pr[Y > D_{th}]_i \leq \delta$, where D_{th} is the handover delay threshold and δ is the tolerable disruption ratio of session i. The probability

Table. 2.3. Disruption probabilities for LMS ($N_Z = 16$ and $N_Z = 64$).

N_L	D_{th}	$Pr[Y > D_{th}]_i$			
		$N_Z = 16$		$N_Z = 64$	
		$i = 1$	$i = 100$	$i = 1$	$i = 100$
1	$D_{L3} \leq D_{th} < D_{L1}$	0.9132	0.9977	0.8988	0.9973
	$D_{L1} \leq D_{th} < D_{Z2}$	0.2500	0.2500	0.1250	0.1250
	$D_{Z2} \leq D_{th} < D_{Z1}$	0.0350	0.2379	0.0000	0.1026
4	$D_{L3} \leq D_{th} < D_{L1}$	0.4029	0.4969	0.3543	0.4954
	$D_{L1} \leq D_{th} < D_{Z2}$	0.2500	0.2500	0.1250	0.1250
	$D_{Z2} \leq D_{th} < D_{Z1}$	0.0350	0.2379	0.0000	0.1026
9	$D_{L3} \leq D_{th} < D_{L1}$	0.2776	0.3310	0.1939	0.3276
	$D_{L1} \leq D_{th} < D_{Z2}$	0.2500	0.2500	0.1250	0.1250
	$D_{Z2} \leq D_{th} < D_{Z1}$	0.0350	0.2379	0.0000	0.1026
16	$D_{L3} \leq D_{th} < D_{L1}$	0.2500	0.2500	0.1425	0.2440
	$D_{L1} \leq D_{th} < D_{Z2}$	0.2500	0.2500	0.1250	0.1250
	$D_{Z2} \leq D_{th} < D_{Z1}$	0.0350	0.2379	0.0000	0.1026
25	$D_{L3} \leq D_{th} < D_{L1}$	N/A	N/A	0.1285	0.1944
	$D_{L1} \leq D_{th} < D_{Z2}$	N/A	N/A	0.1250	0.1250
	$D_{Z2} \leq D_{th} < D_{Z1}$	N/A	N/A	0.0000	0.1026

that receiving session i is disrupted can be written as

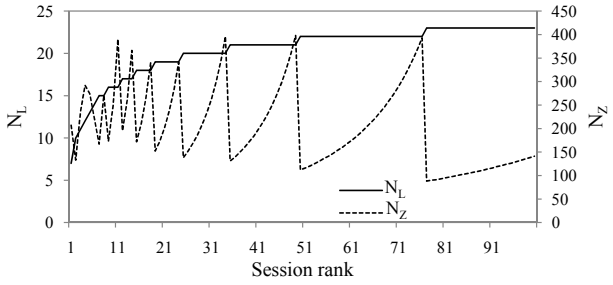
$$Pr[Y > D_{th}]_i = \begin{cases} 1, & 0 \leq D_{th} < D_{L3} \\ \frac{E[Z_{1,i}]+E[Z_{2,i}]+E[L_{1,i}]}{E[C]}, & D_{L3} \leq D_{th} < D_{L1} \\ \frac{E[Z_{1,i}]+E[Z_{2,i}]}{E[C]}, & D_{L1} \leq D_{th} < D_{Z2} \\ \frac{E[Z_{1,i}]}{E[C]}, & D_{Z2} \leq D_{th} < D_{Z1} \\ 0, & D_{Z1} \leq D_{th}. \end{cases}$$

Table. 2.4. Disruption probabilities for LMS ($N_Z = 256$ and $N_Z = 400$).

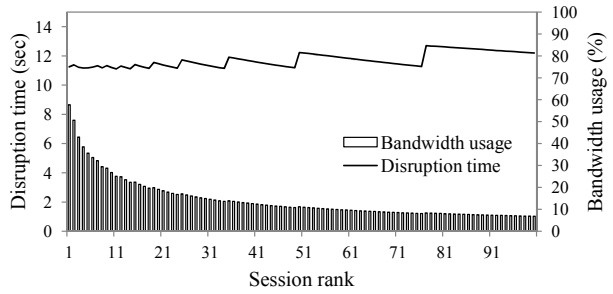
N_L	D_{th}	$Pr[Y > D_{th}]_i$			
		$N_Z = 256$		$N_Z = 400$	
		$i = 1$	$i = 100$	$i = 1$	$i = 100$
1	$D_{L3} \leq D_{th} < D_{L1}$	0.8916	0.9971	0.8901	0.9971
	$D_{L1} \leq D_{th} < D_{Z2}$	0.0625	0.0625	0.0500	0.0500
	$D_{Z2} \leq D_{th} < D_{Z1}$	0.0000	0.0284	0.0000	0.0145
4	$D_{L3} \leq D_{th} < D_{L1}$	0.3301	0.4946	0.3252	0.4945
	$D_{L1} \leq D_{th} < D_{Z2}$	0.0625	0.0625	0.0500	0.0500
	$D_{Z2} \leq D_{th} < D_{Z1}$	0.0000	0.0284	0.0000	0.0145
9	$D_{L3} \leq D_{th} < D_{L1}$	0.1521	0.3259	0.1437	0.3256
	$D_{L1} \leq D_{th} < D_{Z2}$	0.0625	0.0625	0.0500	0.0500
	$D_{Z2} \leq D_{th} < D_{Z1}$	0.0000	0.0284	0.0000	0.0145
16	$D_{L3} \leq D_{th} < D_{L1}$	0.0887	0.2410	0.0780	0.2404
	$D_{L1} \leq D_{th} < D_{Z2}$	0.0625	0.0625	0.0500	0.0500
	$D_{Z2} \leq D_{th} < D_{Z1}$	0.0000	0.0284	0.0000	0.0145
25	$D_{L3} \leq D_{th} < D_{L1}$	0.0689	0.1898	0.0569	0.1889
	$D_{L1} \leq D_{th} < D_{Z2}$	0.0625	0.0625	0.0500	0.0500
	$D_{Z2} \leq D_{th} < D_{Z1}$	0.0000	0.0284	0.0000	0.0145

The resulting disruption probabilities are summarized in Tables 2.3 and 2.4, where $D_{L3} \leq D_{th} < D_{L1}$, $D_{L1} \leq D_{th} < D_{Z2}$, and $D_{Z2} \leq D_{th} < D_{Z1}$ under α=0.8, ρ^*=1 user/cell, and v=60 km/h.

Fig. 2.4(a) shows the values of N_Z and N_L determined by the LMA-based MBSFN area planning algorithm given in Algorithm 2.1 where $\delta = 0.2$ and $D_{L3} \leq D_{th} < D_{L1}$ for all eMBMS sessions ($N_{L,min} = 1$, $N_{Z,min} = 1$, $N_{Z,max} = 400$, $\rho^* = 1$ user/cell, and $v = 60$ km/h). Then, the determination of N_L is affected by the popularity of the session. As shown in Fig. 2.4(a), a less popular session has the higher values of N_L. Intuitively, it is expected that the less popular session has fewer receivers so the bandwidth usage will be lower. In addition, an MS receiving an un-



(a) Deciding N_L and N_Z



(b) Disruption time and bandwidth usage

Fig. 2.4. LMA and MBSFN area planning results.

popular session has a less chance to encounter LMAs with the session on air which results in a larger disruption time. Therefore, to mitigate this effect, N_L should be increased as the session popularity decreases. The determination of N_Z, however, is hardly affected by session popularity. Note that the N_Z curve shown in Fig. 2.4(a) represents minimum values of N_Z. For LMS, the bandwidth usage does not rely on the value of N_Z, so a larger N_Z is always better. For example, a minimum value of N_Z between $i = 50$ and 76 increases as the session popularity decreases because N_L is fixed. At $i = 76$ in particular, the minimum value of N_Z approaches $N_{Z,max}$. So when i reaches 77, a smaller N_Z becomes feasible because of the increase in N_L. Additionally, Algorithm 2.1 usually selects a large N_Z (> 100) in order to avoid inter-

Table. 2.5. Blocking probabilities for the max concurrent session, $m = 20$.

Popularity	$i = 1$	$i = 4$	$i = 7$	$i = 10$	$i = 20$	$i = 30$	$i = 50$	$i = 100$
LMS	0.0%	0.0%	0.4%	0.7%	2.8%	4.2%	8.5%	9.6%
OMS	0.3%	4.5%	7.2%	8.6%	10.7%	11.6%	12.5%	13.3%

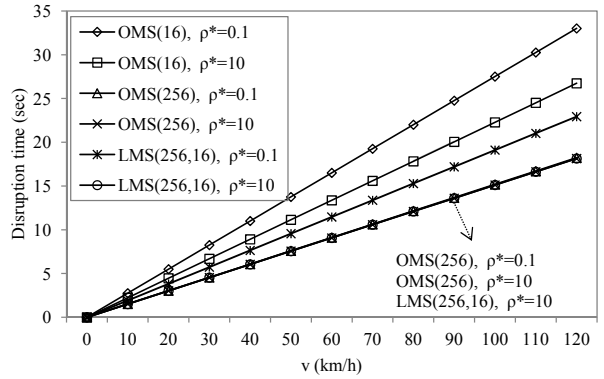
MBSFN area handovers. Fig. 2.4(b) depicts the disruption time and bandwidth usage based on the planning results, and Table 2.5 shows blocking probabilities for $m = 20$ (α=0.8, and ρ^*=1 user/cell). Compared with OMS(25) which satisfies $\delta = 0.2$ minimizing its bandwidth usage[10], the disruption time for LMS is reduced by 15% on average. In addition, the bandwidth usage is slightly reduced so the blocking probabilities are also improved.

2.5.3 Effects of v and ρ^*

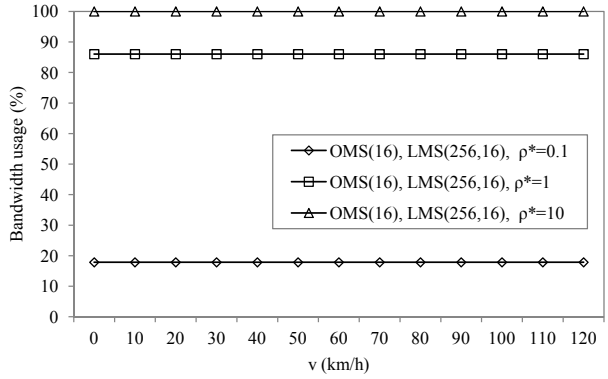
Fig. 2.5(a) shows the average service disruption time of the most popular session ($i = 1$), plotted against v, the average speed of UEs ($\alpha = 0.8$). The disruption time linearly increases with v, because a higher average velocity implies that more frequent handovers are performed during an eMBMS session. Moreover, $T_{OMS,1}$ and $T_{LMS,1}$ are lower for a large population ($\rho^* = 10$) than they are for a small population ($\rho^* = 0.1$). If a large number of users are receiving the same eMBMS packets, each user is more likely to move to active MBSFN areas or active LMAs, which reduces the handover delay.

From Fig. 2.5(b), it can be seen that the bandwidth usage with a large population is higher than it is with a small population. However, note that bandwidth usage is

[10]By putting $N_L = N_Z$ in Algorithm 2.1, the OMS planning result can similarly be obtained.



(a) Disruption time



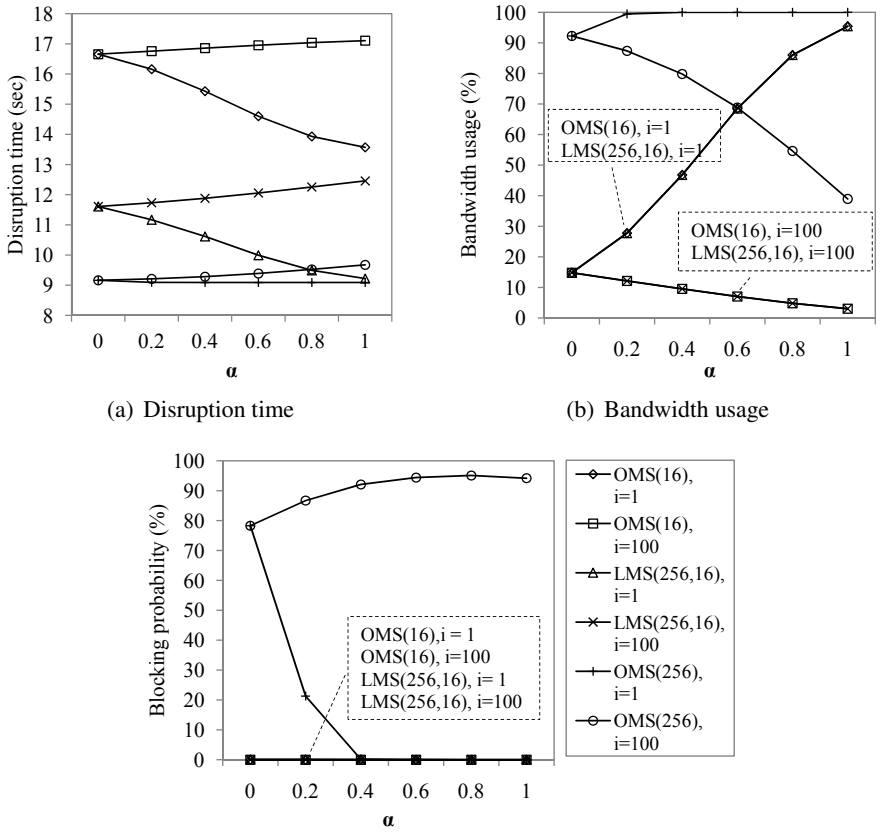
(b) Bandwidth usage

Fig. 2.5. Effects of v and ρ^*.

not dependent on v, but rather on the density of the user distribution, ρ^* (users/cell). If ρ^* is increased from 0.1 to 1, then $U_{LMS,1}$ for LMS(256, 16) (which is the same as $U_{OMS,1}$ for OMS(16)) increases from 18% to 86%.

2.5.4 Effect of α

The average disruption time and bandwidth usage as functions of α are shown in Fig. 2.6 ($m = 20$, $\rho^* = 1$ user/cell, and $v = 60$ km/h). Values of α in the range



(a) Disruption time

(b) Bandwidth usage

(c) Blocking probability

Fig. 2.6. Effect of α.

of 0.64–0.98 have been reported [26]. When $\alpha = 0$, Eq. (2.1) can be simplified to $\beta_i = 1/S$ for all i, and the request rates for all eMBMS sessions become equivalent. Fig. 2.6(a) indicates that $T_{OMS,1}$ for OMS(16), and $T_{LMS,1}$ for LMS(256,16) substantially decrease as α increases. Meanwhile, $T_{OMS,100}$ for OMS(16) and $T_{LMS,100}$ of LMS(256,16) increase slightly with α. In the case of large MBSFN areas (e.g., OMS(256)), the effect of α is insignificant. However, a change in α has a significant impact on the bandwidth usage as shown in Fig. 2.6(b). As α increases, $U_{LMS,1}$

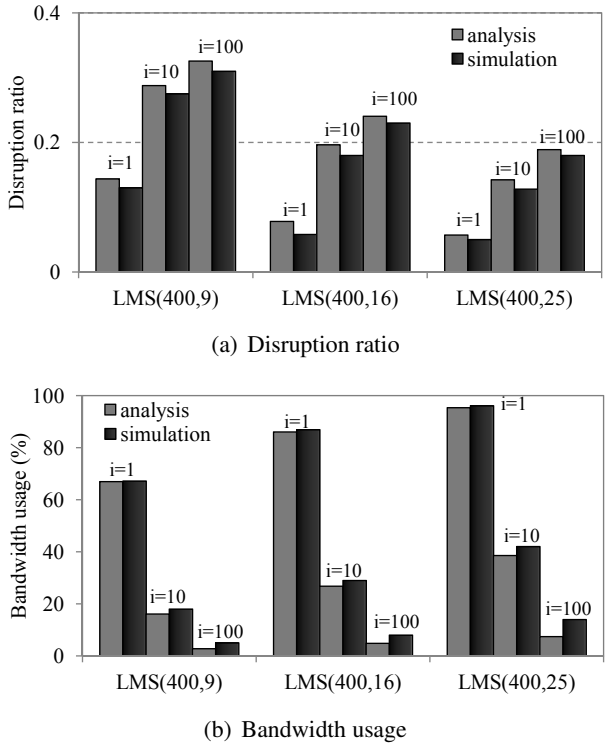
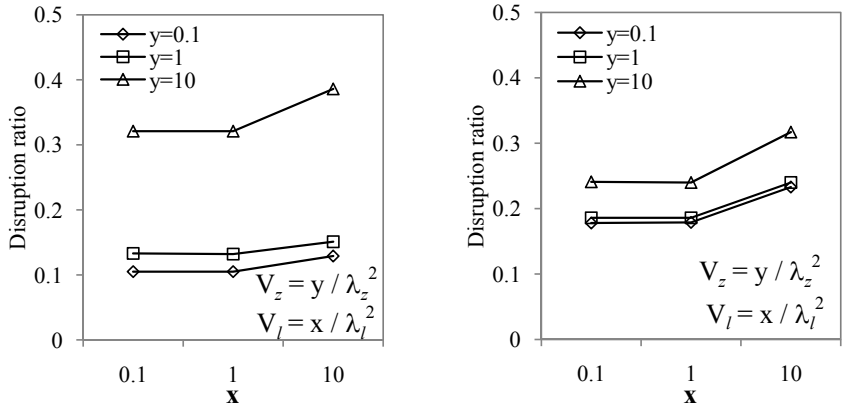


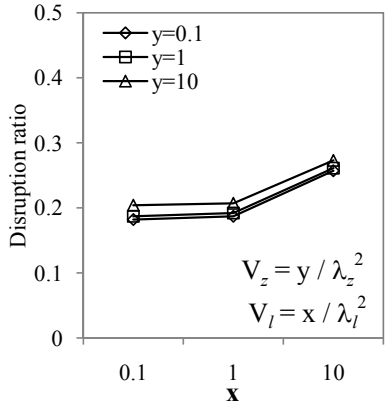
Fig. 2.7. A comparison of the simulation results with the analytical results.

for LMS(256,16) (which is the same as $U_{OMS,1}$ for OMS(16)) increases sharply while $U_{OMS,100}$ for OMS(256) decreases sharply. Also, the more popular session has a higher probability to be present among the already transmitted sessions as α increases. Accordingly, the blocking probability of a popular session decreases. However, unpopular sessions are hardly accepted in a large MBSFN area, as shown in Fig. 2.6(c).



(a) $i = 1$ at LMS(400,9)

(b) $i = 10$ at LMS(400,16)



(c) $i = 100$ at LMS(400,25)

Fig. 2.8. Effects of V_l and V_z on disruption ratios.

2.6 Simulation Results

Our simulations have two goals: one is to verify previously discussed results in Section 2.5.2, and the other is to show the impact of variance of the MBSFN area and LMA residence times. First, we assumed an eMBMS service area (or an LTE network) comprised of 400 cells with 400 randomly located users. Each user

moves according to a 2-D random walk mobility model of which the speed range is uniformly distributed between 0 and 120 km/h, and the service area is actually wrapping-around to remove the boundary effect. To quantify handover delays, our simulation assume that delay values are normally distributed: the link-level handover delay is normally distributed with mean $100ms$ and variance 10^2 (i.e., $N(100, 10^2)$); a multicast distribution updating follows $N(200, 10^2)$; an eMBMS session restarting follows $N(300, 10^2)$. Then, the mean values of D_{Z1}, D_{Z2}, D_{Z3}, D_{L1}, D_{L2} and D_{L3} are given by 600, 400, 100, 300, 100 and 100 msec, respectively. The handover threshold value is set to 130 msec. The simulation duration of each run is 120 minutes.

Fig. 2.7(a) shows the ratio for a noticeable session disruption that an eMBMS user experiences while Fig. 2.7(b) shows the bandwidth usage. Recall that the MBSFN area and LMA planning results for $\delta = 0.2$ are illustrated in Fig. 2.4(a). In the simulations, three sizes of LMAs are examined; each LMA consists of 3 $\times$ 3, 4 $\times$ 4, or 5 $\times$ 5 cells. According to the the analytical results in Table 2.4 and Fig. 2.4(a), LMS(400,25) is feasible (satisfies $\delta = 0.2$) for all sessions, but LMS(400,16) is not feasible for $i = 100$. And LMS(400,9) is only feasible when $i = 1$. These match well with the simulation results shown in Fig. 2.7(a), though the disruption ratios are a little different between the simulations and analytical results. It is because of the mobility of each UE, which may not guarantee that eMBMS users are uniformly distributed over the network. Since our algorithm finds values of N_Z and N_L that minimize the use of bandwidth[11] while keeping a delay requirement, LMS(400,9) is

[11]Fig. 2.7(b) shows that the bandwidth usage is overestimated in the simulations when $i = 100$. It is because the simulations use an integer (2 for $i = 100$) as the expected number of session users, rather than a real number (1.24 for $i = 100$, by analysis).

the best for $i = 1$, whereas LMS(400,16) is the best for $i = 10$. For $i = 100$, only LMS(400,25) is feasible.

Second, we modified our simulations for each UE to have the residence times that follow a specific Gamma distribution so the impact of variance of the residence times can be studied. Fig. 2.8 shows the average disruption ratios of UEs for Gamma residence time distributions with different variance values: the variance of MBSFN area residence times $V_z = \{0.1/{\lambda_z}^2, 1/{\lambda_z}^2, 10/{\lambda_z}^2\}$ and the variance of LMA residence times $V_l = \{0.1/{\lambda_l}^2, 1/{\lambda_l}^2, 10/{\lambda_l}^2\}$. Note that the exponential distribution is a special case of Gamma distributions with mean $1/\lambda$ and variance $1/\lambda^2$.

Overall, the figure indicates that the disruption ratios are substantially affected when the variance of the residence times of MBSFN areas and LMAs are high. For a high variance of LMA residence times, our analysis may underestimate the ratio for a session disruption. In Fig. 2.8, most of the cases for $V_l = 10/{\lambda_l}^2$ do not satisfy $\delta = 0.2$, since the disruption ratios for $V_l = 10/{\lambda_l}^2$ are increased by 14-37% compared to those for $V_l = 1/{\lambda_l}^2$. On the other hand, the variance of the MBSFN area residence times only affects the disruption ratios of popular sessions[12]. As V_z is increased from $1/{\lambda_z}^2$ to $10/{\lambda_z}^2$, the disruption ratios are more than doubled in Fig. 2.8(a) ($i = 1$), whereas they are hardly affected by V_z in Fig. 2.8(c) ($i = 100$).

2.7 Conclusion

LTE Multimedia Broadcast and multicast service over a Single Frequency Network (MBSFN) reduces the eMBMS service disruption caused by handovers; but this requires all Evolved Node Bs (eNBs) to send the same packets in the MBSFN

[12]It may depend on the value of D_{th}. In the simulations, we assumed $D_{L3} \leq D_{th} = 130\text{ms} < D_{L1}$.

area. This has motivated us to propose an MBSFN area planning scheme based on location management areas (LMAs) in order to save wireless link bandwidth while keeping the service disruption time at an acceptable level. We have presented a novel mathematical model of the service disruption time, bandwidth usage and blocking probability which consider user mobility, distribution and eMBMS session popularity. We have evaluated the performance of our scheme (LMS) and compared it with the original MBSFN scheme (OMS) with the following results:

- The inter-LMA handover delay is shorter than the inter-MBSFN area handover delay while the intra-LMA and intra-MBSFN area handover delays are the same. As a result, LMS outperforms OMS in terms of the average disruption time when they use the same amount of bandwidth; or in terms of the bandwidth usage when their average disruption times are the same.

- eMBMS user distribution and session popularity have significant effects on the bandwidth usage and blocking probability while the service disruption time is mainly affected by mobility (e.g., average user speed). Moreover, the disruption ratio can significantly be affected by the variance of the LMA and/or MBSFN residence times.

We demonstrated how to determine the MBSFN area and LMA sizes, which can make the best use of bandwidth while maintaining the quality of eMBMS services. Our results suggest that LMA-based MBSFN area planning would deliver more efficient multicast and broadcast services over LTE systems.

Chapter 3

Proactive Approach for LMA-based MBSFN

3.1 Introduction

The need for bandwidth-efficient data distribution to a large number of users has led to the definition of Enhanced Multimedia Broadcast and Multicast Service (eMBMS) in LTE systems [3]. With the eMBMS support, an LTE network can offer more multimedia streaming services by utilizing its bandwidth resources efficiently. However, as users move across cell boundaries, handovers should be performed, which may incur some disruption in eMBMS session continuity. In this chapter, we focus on how to bound the degree of service degradation when LTE systems provide delay-sensitive multimedia streaming contents using eMBMS.

In LTE, a user equipment (UE) normally performs hard handovers, in which all connections to the serving Evolved Node B (eNB) are broken before new connections are made to the target eNB. As a result, some packets sent through the serving eNB during the handover may be lost. One approach to mitigate this problem is deploying an MBSFN area, defined by the 3GPP standard [2]. The MBSFN area is a group of (typically adjacent) eNBs transmitting the same content, which allows all UEs of the same eMBMS session in their cells to use the same multicast bearer connection and security key during handovers within the same MBSFN area. In this way, the

UE can perform handover with the minimum delay, which in turn minimizes the disruption (or the lost eMBMS packets) as long as it moves within the MBSFN area. This is called an intra-MBSFN area handover. However, since every eNB in the same MBSFN area broadcasts the packets of the eMBMS session regardless of the presence of a user, the scarce wireless link bandwidth can be wasted.

In Chapter 2, we leveraged the location awareness of UEs for MBSFN area planning, balancing the tradeoff between the bandwidth usage and service disruption. But, since this approach relies on the network keeping track of the locations of UEs, the granularity of location management affects the average handover delay, which means that the disruption in eMBMSs may not be readily configurable by LTE operators.

In this chapter, we propose to transmit eMBMS packets proactively and probabilistically to stochastically bound the average service disruption time for an eMBMS user. The central idea is to broadcast eMBMS packets with a certain probability in a test cell[1] despite no current eMBMS user if there is an eMBMS user in the adjacent cell(s). In this way, when the eMBMS user in the adjacent cell hands over to the test cell, he/she can continue receiving the packets of the eMBMS session with minimum discontinuity. The performance of the proposed scheme is highly dependent on how to determine the probability of proactive transmissions. To provide stochastic QoS provisioning for eMBMSs, the probability of proactive transmissions is determined by considering user mobility, spatial user distribution, and session popularity.

Table. 3.1. Notation for proactive MBSFN performance analysis.

μ_z	mean residence time in an MBSFN area
μ_l	mean residence time in an LMA
μ_c	mean residence time in a cell
μ_s	mean eMBMS session duration
ρ_i	average number of users per unit area of ith most popular session
A_z	area of an MBSFN area
A_l	area of an LMA
N	total number of LMAs in a network
S	total number of eMBMS sessions
Z_k	number of MBSFN area handovers ($k = 1$: inter-MBSFN area handover moving to inactive MBSFN areas, $k = 2$: inter-MBSFN area handover moving to active MBSFN areas)
L_k	number of LMA handovers ($k = 1$: inter-LMA handover moving to inactive LMAs, $k = 2$: inter-LMA handover moving to active LMAs, $k = 3$: intra-LMA handover)
D_{Zk}	delay of the corresponding MBSFN area handover Z_k
D_{Lk}	delay of the corresponding LMA handover L_k

3.2 Network and MBSFN Modeling

We consider an LTE network in which there are total N cells; we assume that the LMA-based MBSFN handover is supported in the network, as described in Chapter 2. That is, the LTE network consists of multiple MBSFN areas, each of which is divided into multiple LMAs. For the sake of exposition, we assume that cells, LMAs, and MBSFN areas are square-shaped, and the residence times of an eMBMS user in an

[1] Actually, we will consider a group of cells as a unit area of broadcasting eMBMS packets, to be detailed later.

MBSFN area, an LMA, and a cell follow exponential distributions with means μ_z, μ_l and μ_c, respectively, such that $\mu_z \geq \mu_l \geq \mu_c$. The eMBMS session duration time follows an exponential distribution with mean μ_s.

The total number of eMBMS sessions on air is S, and all sessions are ranked by popularity, which indicates how many users currently receive the packets of a particular eMBMS session. Let β_i be the conditional probability that a request arrives for the ith most popular eMBMS session ($i = 1, 2, ..., S$). β_i is drawn from a cut-off Zipf-like distribution [26], and is given by

$$\beta_i = \frac{\Omega}{i^\alpha}, \qquad \text{where } \Omega = \left(\sum_{i=1}^{S} \frac{1}{i^\alpha} \right)^{-1}, \quad 0 < \alpha \leq 1.$$

The spatial distribution of users in the LTE network follows a two-dimensional Poisson distribution with net rate ρ^*, which is defined as the average number of users per unit area: $\rho^* = \lambda^*/\mu^*$, where λ^* is the users' arrival rate into the network and μ^* is the rate at which users leave the network. Then, the average number of users of the ith most popular session per unit area is $\rho_i = \beta_i \rho^*$.

From a perspective of UEs of a particular session, there are two states of MBSFN areas and LMAs: active and inactive. MBSFN areas which currently contain users of the current eMBMS session are called *active* MBSFN areas, whereas those without such users are called *inactive* MBSFN areas (Similarly, there are active and inactive LMAs.). Depending on the states of target MBSFN areas/LMAs, the handover delay becomes different since multicast distribution updating and/or eMBMS session restarting in handovers to inactive ones are skipped in handovers to active ones;for details, refer to Table 2.2. Let the delay of an inter-MBSFN area handover to

an inactive MBSFN area be D_{Z1}, and let the delay of an inter-MBSFN area handover to an active MBSFN area be D_{Z2}. Since an MBSFN area is partitioned into multiple LMAs, intra-MBSFN area handovers are classified into three cases, which are inter-LMA handovers to an inactive LMA, inter-LMA handovers to an active LMA, and intra-LMA handovers. Let each delay be D_{L1}, D_{L2}, and D_{L3}, respectively.

Table 3.1 summarizes the notation to analyze the disruption of an eMBMS session. Note that each of the following random variables indicates how many times an eMBMS user will experience each kind of handover delay during his/her session. Z_1 and Z_2 are the random variables for the numbers of inter-MBSFN area handovers to inactive and active MBSFN areas, respectively. L_1 and L_2 are the random variables for the numbers of inter-LMA handovers to inactive LMAs and to active LMAs, respectively. L_3 is the random variable for the number of intra-LMA handovers. Following the approach in Chapter 2, the service disruption time for an eMBMS user is defined as the sum of all the expected handover delays during the service time of an eMBMS session, which can be expressed as

$$\begin{aligned}\text{Disruption Time} = {} & E[Z_1] \cdot D_{Z1} + E[Z_2] \cdot D_{Z2} \\ & + E[L_1] \cdot D_{L1} + E[L_2] \cdot D_{L2} + E[L_3] \cdot D_{L3}, \quad (3.1)\end{aligned}$$

where

$$\begin{cases} E[Z_1] = \frac{\mu_s}{\mu_z} e^{-\rho_i A_z} \quad \text{and} \quad E[Z_2] = \frac{\mu_s}{\mu_z}(1 - e^{-\rho_i A_z}), \\ E[L_1] = \frac{\mu_s(\mu_z - \mu_l)}{\mu_l \mu_z} \cdot \Pr\{\textit{inactive LMA}\}, \\ E[L_2] = \frac{\mu_s(\mu_z - \mu_l)}{\mu_l \mu_z}(1 - \Pr\{\textit{inactive LMA}\}), \\ E[L_3] = \frac{\mu_s(\mu_l - \mu_c)}{\mu_c \mu_l}. \end{cases}$$

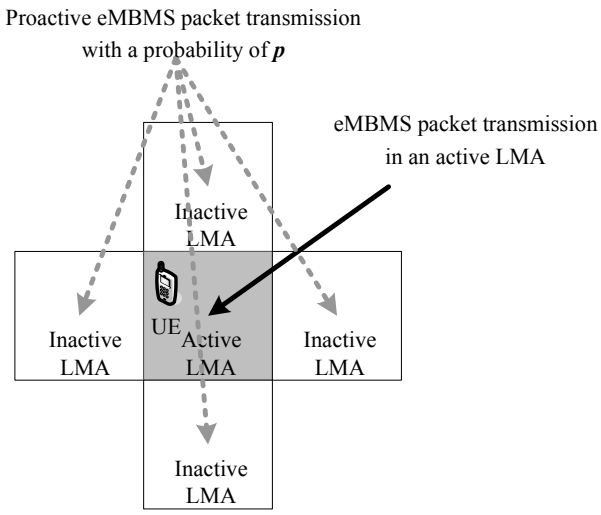


Fig. 3.1. An illustration of the popularity-based proactive MBSFN.

Note that Pr{*inactive LMA*} denotes the probability that an LMA is inactive.

3.3 Proactive LMA-based MBSFN

In this chapter, we propose a proactive transmission-based (and LMA-based) MBSFN scheme that not only transmits the eMBMS packets of session i to active LMAs, but also probabilistically transmits the eMBMS packets to inactive LMAs adjacent to active LMAs, as shown in Fig. 3.1. The probability of proactively transmitting the packets of session i in inactive LMAs is denoted by p_i, and the proposed scheme is called *proactive* MBSFN. Intuitively, when a UE moves from an active LMA to an inactive LMA, the UE should trigger rejoining the corresponding multicast session during its handover process. This results in reconstructing the multicast distribution tree, which incurs substantial delay due to the graft latency [34]. In other

words, the proactive MBSFN scheme removes the multicast distribution updating delay when a UE hands over from an active LMA to an inactive LMA if the inactive LMA already has joined the multicast tree and proactively broadcast eMBMS packets. From now on, an inactive LMA where eMBMS packets are proactively transmitted is called a "proactive" LMA.

Determining the value of p_i relies on popularity and delay requirements of eMBMS session i. In general, however, as the value of p_i increases, the number of LMAs transmitting data increases. Then, the average number of inter-LMA handovers to an inactive LMA will be decreased, and thus the average handover delay is reduced.

3.3.1 Problem Formulation

To determine the probability p_i, we establish a constraint that the average handover delay for a session i user is kept to some specified value. We call this value *handover delay threshold*. At first, the probability that an LMA has x users of session i is $(\rho_i A_l)^x e^{-\rho_i A_l}/x!$ due to the two-dimensional Poisson process. By making x 0, the probability that a given LMA has no user is calculated. In the proposed scheme, each inactive LMA (that is adjacent to at least one active LMA of session i) will be changed to a proactive LMA with a probability of p_i. Therefore, Pr{*inactive LMA*} is expressed as

$$\Pr\{\textit{inactive LMA}\} = (1 - p_i)e^{-\rho_i A_l}. \tag{3.2}$$

Then, the average handover delay can be calculated by dividing the service disruption time by the total number of handovers (μ_s/μ_c) during a session. From (3.1) and (3.2), the average handover delay for session i with proactive transmission probability p_i,

$D(p_i)$, is expressed as

$$\begin{aligned} D(p_i) &= \frac{\mu_c}{\mu_z}\left\{e^{-\rho_i A_z} D_{Z1} + \left(1 - e^{-\rho_i A_z}\right) D_{Z2}\right\} \\ &+ \frac{\mu_c(\mu_z - \mu_l)}{\mu_l \mu_z}\left\{q_i e^{-\rho_i A_l} D_{L1} + \left(1 - q_i e^{-\rho_i A_l}\right) D_{L2}\right\} \\ &+ \frac{(\mu_l - \mu_c)}{\mu_l} D_{L3}, \end{aligned} \tag{3.3}$$

where $q_i = 1 - p_i$ and $0 \leq p_i \leq 1$.

We now proceed to state how the proactive MBSFN scheme can satisfy the QoS requirement of session i. Let the handover delay threshold of ith most popular session be γ_i. Using (3.3), the problem is formulated as:

Given: μ_c, μ_l, μ_z, ρ_i, A_l, A_z, D_{Z1}, D_{Z2}, D_{L1}, D_{L2}, D_{L3}, γ_i $(i = 1, 2, ..., S)$.

To find: For each session i, find the minimum p_i that satisfies $D(p_i) \leq \gamma_i$.

Minimum

$$p_i = 1 - \left[\frac{\left(\frac{\mu_l \mu_z}{\mu_z - \mu_l}\right)\left\{\frac{\gamma_i}{\mu_c} - \frac{D_{Z1}}{\mu_z} e^{-\rho_i A_z} - \frac{D_{Z2}}{\mu_z}\left(1 - e^{-\rho_i A_z}\right) - \frac{D_{L3}(\mu_l - \mu_c)}{\mu_c \mu_l}\right\} - D_{L2}}{e^{-\rho_i A_l}(D_{L1} - D_{L2})}\right]. \tag{3.4}$$

Increasing p_i will improve QoS for a given level of mobility, but the handover delay cutback comes at the cost of amplified traffic. Therefore, the bandwidth cost of an eMBMS session due to proactive transmissions should be analyzed. The bandwidth cost function for session i is denoted by $C(p_i)$, which is defined as the ratio of the number of active and proactive LMAs transmitting eMBMS packets to the total number of LMAs in the network. Let θ be the probability that an LMA which has no user is being considered as a proactive LMA candidate. Then, $C(p_i)$ can be expressed

as

$$C(p_i) = \frac{n + p_i\theta(N-n)}{N}, \tag{3.5}$$

where n is the expected number of active LMAs and is given by $n = N(1 - e^{-\rho_i A_l})$. An LMA with no user can be proactive only if there should be at least one neighbor LMA with eMBMS user(s). Since the users of session i can be assumed to be evenly distributed with the net rate ρ_i, each user is located on any LMA with probability $1/N$. This yields

$$\begin{aligned} \theta &= Pr\{\text{at least one adjacent LMA is active} \mid \text{a given LMA has no user}\} \\ &= \frac{\sum_{j=1}^{u} \binom{u}{j} [4/N]^j [(N-5)/N]^{u-j}}{[(N-1)/N]^u}, \end{aligned}$$

where u is the total number of users receiving session i in the network.

3.3.2 Overall procedure

The overall procedure to determine the p_i for QoS requirements of eMBMS session i with the constraint on the maximum bandwidth cost is described in Algorithm 3.1. As the real network and eMBMS parameters vary over time, we may need to run this algorithm at the multi-cell/multicast coordination entity (MCE) periodically or when the perceived QoS deviation exceeds a certain threshold.

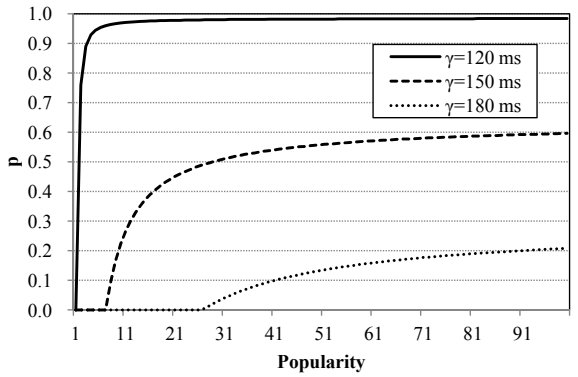
Algorithm 3.1 Popularity-based Proactive MBSFN Scheme.

// C_{max} is the maximum cost allowed in the network.
// M_i is the set of active LMAs for session i.
Update μ_c, μ_l and μ_z.
for all i such that $1 \leq i \leq S$ **do**
    Update ρ_i.
    Find minimum p_i^{new} such that $D(p_i^{new}) \leq \gamma_i$.
    if $\sum_{j=1}^{S} C(p_j) - C(p_i) + C(p_i^{new}) \leq C_{max}$ **then**
        $p_i \leftarrow p_i^{new}$
    end if
    for all $m \in M_i$ **do**
        for all inactive neighbor LMAs of m **do**
            Transmit the packets of session i with a probability of p_i.
        end for
    end for
end for

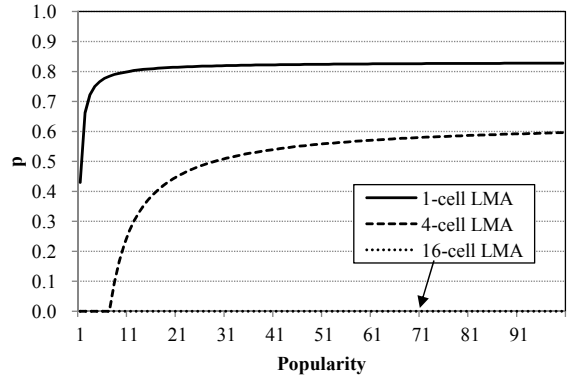
3.4 Performance Evaluation

3.4.1 Simulation Setup

We simulate an MBSFN service area (or an LTE network) comprising of 1024 cells, which is partitioned into four MBSFN areas. Each cell is an 1 km $\times$ 1 km square. Three sizes of LMAs are examined; each LMA consists of 1 $\times$ 1, 2 $\times$ 2, or 4 $\times$ 4 cells. For instance, if each LMA consists of 2 $\times$ 2 cells, it means that the location of each UE is tracked at a granularity of 4 km^2. There are total 100 eMBMS sessions which are ranked by the popularity (i.e., session 1 is the most popular). The total number of users (across all the sessions) is 1024 $\times$ 10, which means ρ^* is 10. For each session, the number of receivers is assigned by a Zipf-like distribution with $\alpha = 0.8$. Our simulation uses a 2-D random walk mobility model of which speed range is $[0, 120]$ km/h, and the service area is actually wrapping-around to



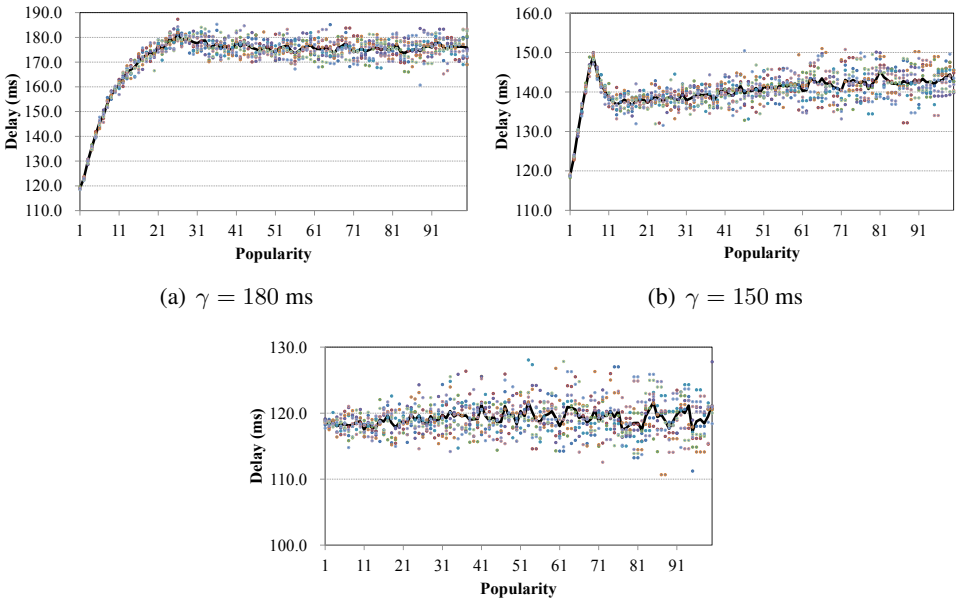
(a) For different handover thresholds when 4-cell LMA is used



(b) For different LMA sizes when $\gamma = 150$ msec

Fig. 3.2. Computation of p_i.

remove the boundary effect. To quantify handover delays, our simulation follows the handover delay values used in Chapter 2. Since we try to reduce not the LTE link-layer handover process but the multicast distribution updating process, the link-layer handover delay is ignored and no link error is assumed. The simulation duration of each run is 30 minutes.



(a) $\gamma = 180$ ms

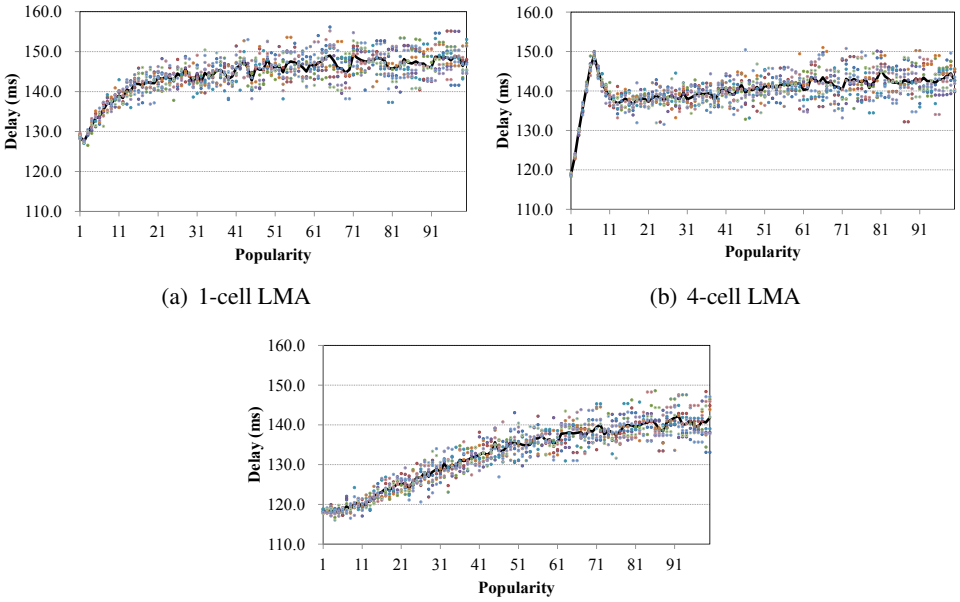
(b) $\gamma = 150$ ms

(c) $\gamma = 120$ ms

Fig. 3.3. Handover delays for different γ values.

3.4.2 Computation of p_i

Given the above values we can get $D(p_i)$, and subsequently compute the probability p_i from Eq. (3.4). Fig. 3.2 plots the probability p_i as a function of the session rank for different handover thresholds and different LMA sizes. Notice that the less popular session has the higher values of p_i. An MS joining an unpopular session has less chance to encounter active LMAs, which results in larger average handover delay. Therefore, to reduce disruption, the number of proactive LMAs should be increased as the session popularity decreases. Also, the higher value of p_i is needed for the tighter delay requirement in Fig 3.2(a), while Fig 3.2(b) shows that the smaller LMA size requires the higher values of p_i.



(a) 1-cell LMA

(b) 4-cell LMA

(c) 16-cell LMA

Fig. 3.4. Handover delays for different LMA sizes.

3.4.3 Simulation Results

We simulate handover delays of UEs using the value of p_i obtained from the analytic modeling (Fig. 3.2). For each session rank, we carry out 10 simulation runs (plotted as dots) and the average handover delay of UEs is drawn as bold lines. Figs. 3.3(a), 3.3(b), and 3.3(c) show the average handover delays for different handover thresholds with the 4-cell LMA size. In most cases, the handover delays are below the threshold. When $\gamma = 180$ ms in Fig. 3.2(a), our model suggests that p_i is zero for the session index $i \leq 26$, since the handover delay is expected to be less than 180 ms without proactive transmissions due to popularity of sessions. In Fig. 3.3(a), the delays are rapidly increased up to 180 ms as i goes to 26. As i exceeds 26, p_i

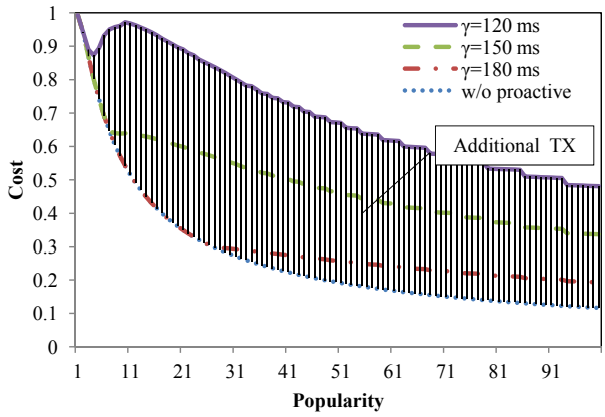
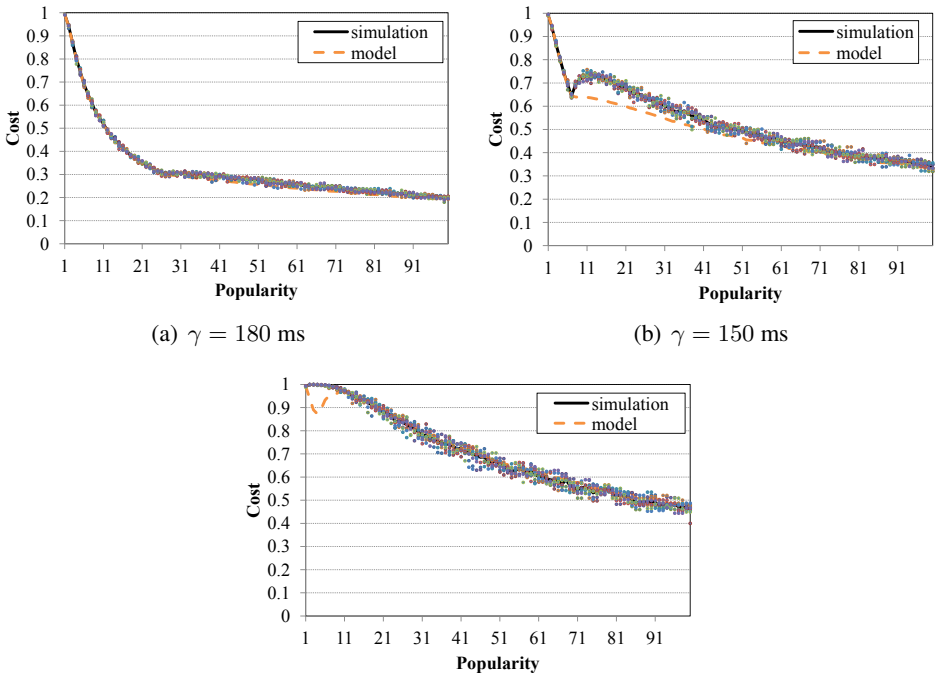


Fig. 3.5. Bandwidth cost.

is increased, so that the delays are reduced by proactive transmissions. Figs. 3.3(b) and 3.3(c) also confirm the results obtained from Fig. 3.2(a).

Figs. 3.4(a), 3.4(b), and 3.4(c) show the average handover delays for different LMA sizes with $\gamma = 150$ ms. As the LMA size increases, the number of crossing LMAs during a session is reduced; thus, p_i becomes smaller as shown in Fig. 3.2(b). By adjusting p_i depending on LMA sizes, the delays are shown to be readily controlled. Therefore, our proposed scheme can adjust the probability of proactive transmissions to satisfy QoS requirements, irrespective of the LMA sizes.

Fig. 3.5 depicts the bandwidth cost from Eq. (3.5) with 4-cell LMA sizes. The dotted curve (which is the minimum) represents the cost for transmitting eMBMS packets only to active LMAs, and therefore the shaded area between the dotted curve and the cost curve for $\gamma = 120$ shows how much cost is raised by the proposed scheme. The bandwidth cost significantly increases as the delay requirement becomes tighter. Figs. 3.6(a), 3.6(b), and 3.6(c) show the bandwidth costs measured by the



(a) $\gamma = 180$ ms

(b) $\gamma = 150$ ms

(c) $\gamma = 120$ ms

Fig. 3.6. Bandwidth cost simulations with 4-cell LMA configuration.

simulations, compared to the dashed lines computed by Eq. (3.5). The dots are the bandwidth cost of each simulation run; the average cost of 10 runs is drawn as the solid line.

3.5 Conclusions

One crucial issue for real-time multimedia streaming services in the LTE Enhanced Multimedia Broadcast and Multicast Service (eMBMS) framework is how to provide QoS for recipients despite their mobility. The location management area (LMA)-based MBSFN scheme partitions an MBSFN area into multiple LMAs to bal-

ance the average handover delay and the bandwidth usage overhead. In this chapter, assuming the same LMA-based MBSFN framework, we seek to bound the disruption of an eMBMS session stochastically. To this end, the proposed scheme transmits eMBMS packets not only to the LMAs with eMBMS users, but also to their neighbor LMAs without eMBMS users probabilistically and proactively. By considering session popularity, user distribution, and user mobility, the probability of proactive transmissions in the neighbor LMAs without users is determined for each eMBMS session. Through extensive simulations, the analytic model to determine the probability of proactive transmissions is verified to satisfy the QoS requirements.

Chapter 4

Performance Improvements on HTTP Adaptive Video Streaming

4.1 Introduction

HTTP adaptive streaming (HAS) has become one of the most cost-effective solutions in delivering video content due to the abundance of Web platforms, and received great attention from both industry and research communities [4–7]. Major video service providers including Netflix or YouTube usually offer video streaming services based on the HAS techniques.

The basic idea of the HAS techniques is to divide the whole video into multiple segments in time dimension, and to have multiple versions for each segment corresponding to multiple bitrates (e.g., 720p or 1080p). Depending on time-varying link conditions and various device capabilities, a suitable version of a video segment is dynamically decided and delivered over HTTP [4]. Thus, HAS can enhance users' QoS of video due to its dynamically adjusting capability. The current state-of-the-art HAS techniques can be categorized into two approaches: (i) client-side and (ii) network-side. In the client-side approach (e.g., FESTIVE [35]), the bitrate (or the corresponding version) of each video segment is decided in a user based on her measured throughput. However, the decision of video bitrate in a user can affect other users in the same cell since wireless resources in each cell are shared by all the users,

which often results in the under-utilization of resources in the cell. The network-side approaches (e.g., AVIS [36] and AGBR [37]) aim at optimizing the overall radio resource utilization considering the QoS of the users, but they often suffer from the stability of video quality.

This chapter first reveals the root causes of the problems of the HAS techniques (both client-side and network-side approaches) in cellular networks based on a simulation study. As to the client-side approach, we find that a video-user[1] tends to be assigned less bandwidth compared to a data-user due to (i) the conservative bitrate selection for maintaining the video quality stable, (ii) the unawareness of the status of wireless link resources in a cell, and (iii) the absence of a mechanism of prioritizing the video traffic. In the network-side approach, we find the stability of video quality is difficult to achieve due to the indirect control mechanism. That is, the central controller in a network sets guaranteed bit rate (GBR) / maximum bit rate (MBR) and lets the rate controller in a user equipment (UE) select the bitrate.

To provide a fair, efficient, and stable video streaming service by addressing the above problems, we propose a network-side HAS solution that optimizes the total utility of all users in a cell including video- and data-users, while maintaining the stable video quality. The key characteristics are: (i) unification of video- and data-users into a single utility framework, (ii) direct rate control by conveying the assigned rates to the video client through overwritten HTTP Response messages, and (iii) rate allocation for stability by a stateful approach. After addressing discrete bitrate models, we further provide a continuous optimization algorithm for the bitrate assignment to reduce the computational complexity of the proposed solution. By conducting an ex-

[1] *Video-users* indicate users who download video content while *data-users* refer to users who download delay-insensitive content.

tensive simulation study using the ns-3 simulator, we show that the proposed solution significantly enhances the average video bitrates, stability of video quality, and balance among video- and data-users, compared to other client-side and network-side solutions.

The rest of this chapter is organized as follows. After reviewing the related work in Section 4.2, we discuss the problems and their root causes of prior HAS solutions in Section 4.3. Sections 4.4 and 4.5 describe the overview and main algorithms of the proposed solution, respectively. Section 4.6 compares the proposed solution with other HAS solutions via simulations. In Section 4.7, we implement a prototype of our solution using an LTE femtocell. We conclude this chapter in Section 4.8.

4.2 Related Work

Dynamic Adaptive Streaming over HTTP (DASH) [4] is a standard for adaptive video streaming solutions, which is standardized by Moving Picture Expert Group (MPEG). There are also well-known proprietary solutions such as Adobe Systems HTTP Dynamic Streaming [5], Apple HTTP Live Streaming [6], and Microsoft Smooth Streaming [7]. The basic idea of these adaptive video streaming solutions is to divide the whole video into multiple segments in time dimension; each segment usually corresponds to a time interval, say a few seconds. Each segment again has multiple versions corresponding to multiple bitrates. For each time interval, one version of a segment is decided (and requested) depending on the network status and the device capacity. The information about segments such as sequence, timing, bitrates, and URLs is described in the Media Presentation Description (MPD) file. Prior to down-

loading video segments, a client downloads the MPD file, parses it, and requests each segment by the HTTP GET method.

Several papers have proposed the bitrate assignment algorithms for adaptive streaming [35–41]. Jiang *et al.* [35] developed bitrate adaptation techniques (called FESTIVE) that guide the trade-offs among stability, fairness, and efficiency. Tian *et al.* [38] proposed video rate control algorithms that balance the smoothness of video rate and high bandwidth utilization. Some studies have focused on the bitrate assignment algorithms in cellular networks [36,37]. Chen *et al.* [36] proposed an in-network resource management framework (AVIS) that schedules HTTP-based video flows in cellular networks. They reported that the proposed framework achieves a balance between optimal resource allocation and user QoE. Vleeschauwer *et al.* [37] proposed a utility maximization framework that takes into account different utility functions for video and data flows, called adaptive guaranteed bit rate (AGBR). In AGBR, the target bitrates are calculated and passed onto a scheduler in a base station. Mok *et al.* [40] suggested a QoE-aware DASH system, where its (measured) available bandwidth information facilitates the selection of video quality level at the client-side. Recently, Li *et al.* [41] revealed that the discrete nature of the video bitrates makes clients difficult in perceiving its fair share bandwidth, and proposed an adaptation algorithm akin to TCP congestion control. Our solution is differentiated from the above network-side approaches in terms of strict rate enforcement, stateful rate selection for the stability, and a unified optimization framework for video and data flows.

There also have been efforts in investigating the performance of adaptation algorithms that are implemented on various adaptive streaming players (e.g., Smooth Streaming, Netflix, and OSMF) [42–45], revealing valuable insights into the perfor-

mance aspects of adaptive streaming such as the reaction to the changes of bandwidth or fairness among players. We evaluate the performance of the representative client-side and network-side algorithms over LTE networks from the perspective of balance between data and video flows, and stability of video quality.

4.3 Problem Definition

The current state-of-the-art DASH solutions can be categorized into two approaches: (i) client-side and (ii) network-side. In client-side approaches, the video bitrate is controlled by the adaptation algorithm which is located in user devices. The adaptation algorithm performs the bandwidth estimation by using the history of previously downloaded segments, and decides the bitrate according to the video information, bandwidth estimation, and history of previous bitrate selections. However, in client-side approaches, video-users suffer from lower bitrates than data-users, which results in lower utility to video-users. That is, while data-users can fully utilize the bandwidth as long as the TCP mechanism permits, video-users cannot fully utilize the available bandwidth.

The reason is that the conservative client-side algorithms (i) select the bitrate under the average link throughput to avoid frequent bitrate changes, and (ii) wait without downloading segments if the buffer is full. In addition, this imbalance problem results from user's being unaware of the wireless channel status, and the absence of mechanism for prioritizing video traffic, in contrast to network-side approaches. To illustrate this problem, we conducted a simple simulation experiment using the settings in Table 4.3. We run the simulation 20 times using the client-side algorithm, FES-

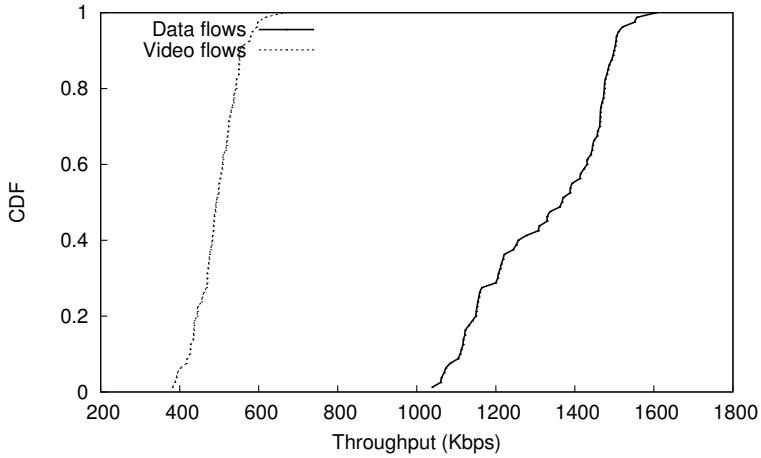


Fig. 4.1. Throughput comparison of video and data users in FESTIVE.

TIVE [35], and Fig. 4.1 compares the average throughput of video and data users. It shows that the video-users achieve lower throughput than the data-users, even though the video-users require more traffic than the data-users.

The existing network-side approaches consider the each user's channel status, utility, and decide the bitrates for all video users so that the total utility is maximized. Then the decided bitrates are applied by setting the guaranteed bit rate (GBR) or maximum bit rate (MBR) for each flow using the base station scheduler or resource slicing technique. These approaches assume that client-side adaptation modules in user devices are performed separately. Prior network-side approaches such as AGBR [37] or AVIS [36], however, suffer from the instability of video quality. These approaches allocate the bandwidth using GBR / MBR of video-users, and expect for the bitrates requested by the adaptation modules in user devices to converge to the

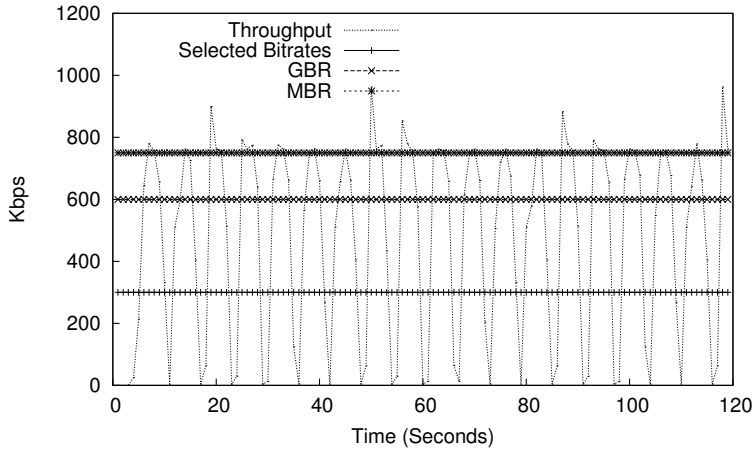


Fig. 4.2. Bitrate selection under constant GBR / MBR settings.

allocated bandwidth. This expectation does not always come true since available bitrates are discrete. To demonstrate this, we conduct a simple simulation by setting the GBR to 600 Kbps, MBR to 750 Kbps, and available bitrates to 300, 600, 900 Kbps, and running one video flow with a simple client-side algorithm that chooses the highest rate that can be supported by the estimated throughput. In this setting, the AVIS algorithm expects that the 600 Kbps bitrate is selected. Figure 4.2 shows that the selected bitrate does not converge to the allocated bandwidth. This is due to (i) the absence of consideration on the effect of the slow start mechanism of TCP, and/or (ii) the conservative nature of the client-side adaptation algorithms in selecting the bitrates.

We are motivated by the above problems of state-of-the-art DASH solutions. To this end, our proposal seeks to achieve the following goals:

- *Total utility*: To maximize the total utility of all the users in a cell, high bitrates are targeted as much as possible. The total utility is calculated by the sum of utilities of video-users and data-users according to their achieved bitrates as follows: $\sum_{u \in U} \beta_u (1 - \frac{\theta_u}{T_u}) + \sum_{u \in D} \log \frac{T_u}{\theta_u}$ where T_u is the average bitrate for user u. The other notation is explained in Table 4.2.

- *Stability*: While the utility value reflects the quality of user experience for the achieved birates, it does not capture the degradation of user experience due to the frequent changes of bitrates. The stability is defined as $\sum_{u \in U} \sum_{i \in I} XOR(L_u^i - L_u^{i-1})$ where I is the set of all the bitrate assignment interval indices, XOR is the function that returns 1 if two arguments are the same, and 0 otherwise. L_u^i is the bitrate index of i-th segment of user u.

- *Fairness*: We also try to achieve fair throughput allocation among users. Note that this is different from *"resource fairness"* discussed in [36]. That is, we focus on the fairness of actual throughput, which is related with real user experiences, instead of the fairness of allocated resources. The fairness is defined as Jain's fairness index (in terms of the average throughput of each user) as follows: $\frac{(\sum_{u \in U} T_u)^2}{N \sum_{u \in U} T_u^2}$.

4.4 Utility-aware Network-side Streaming Approach

In this section, we suggest a new network-side HTTP streaming scheme for LTE networks to achieve the three goals described in the previous section. Fig. 4.3 illustrates the overall architecture of the proposed scheme. As shown in Fig. 4.3, a new entity, streaming proxy (SP), is added between a media server and Evolved

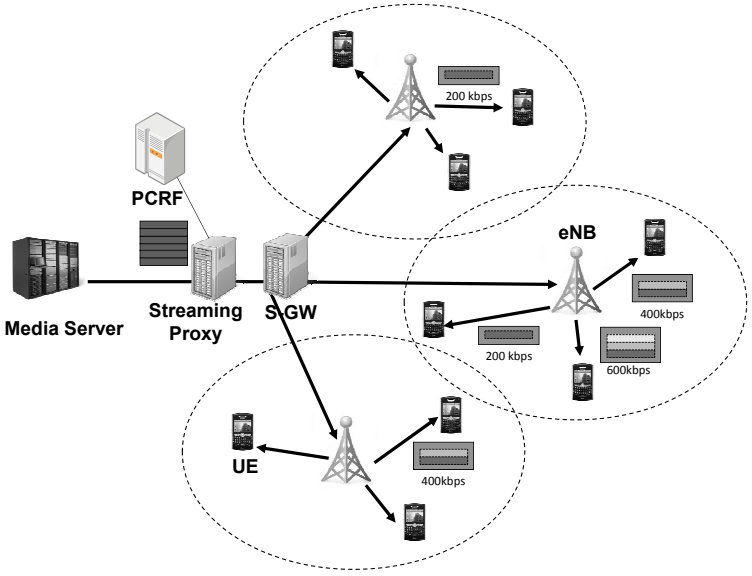


Fig. 4.3. Overall network architecture for the proposed scheme.

Node Bs (eNBs) in an LTE network.

4.4.1 Streaming Proxy (SP)

The proposed scheme introduces the SP which calculates the bitrate for each DASH client in a User Equipment (UE) to maximize the total utility of all users in each cell (to be detailed in the next section). Then the SP replaces the bitrates decided by the UEs with the ones chosen by SP via overwriting HTTP requests. A single SP can manage multiple UEs that belong to multiple eNBs. Since each bitrate decision problem of each cell is independent of each other, the bitrate calculation can be performed at cell-level in parallel.

We assume that there can be two types of SPs: (i) an SP with a cache and (ii) an SP without a cache. Note that an SP with a cache has a large storage to store the video

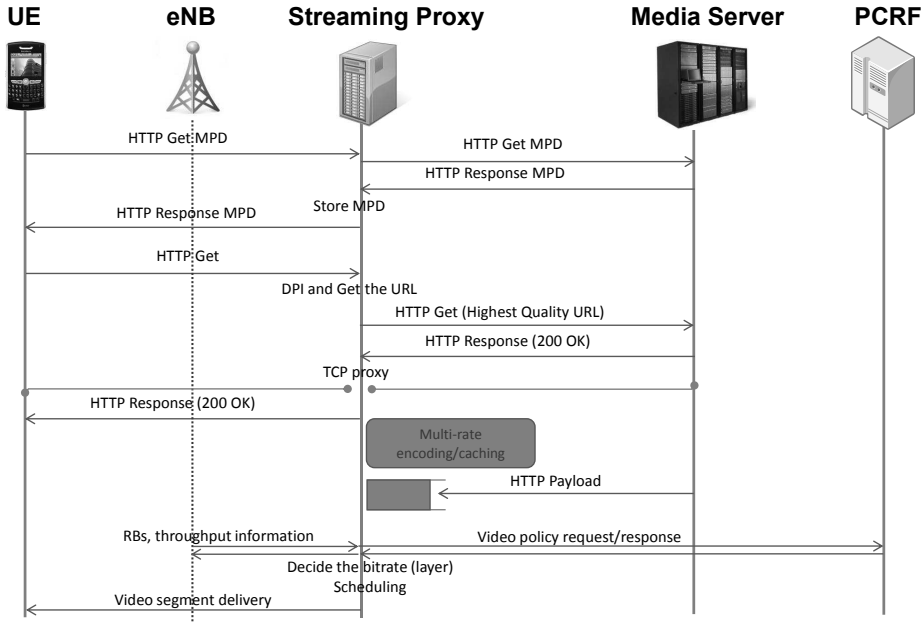


Fig. 4.4. A message flow of SP with a cache.

to provide stable services. The SP with a cache: (1) stores (or caches) the highest bitrate segments and encodes them into various bitrates according to the policy of mobile operators, (2) selects bitrates based on the various information including the resource blocks (RBs) availability, RB assignment history, UE throughput history, video traffic policy from the PCRF (Policy and Charging Rules Function), and (3) sends a video segment of the selected bitrate in response to each requests from the UE. On the other hand, the SP without a cache: (1) selects the bitrate for each video-user based on the information mentioned above, (2) replaces the URLs (of the bitrates selected by UEs) in HTTP request messages by URLs (of selected bitrates by the

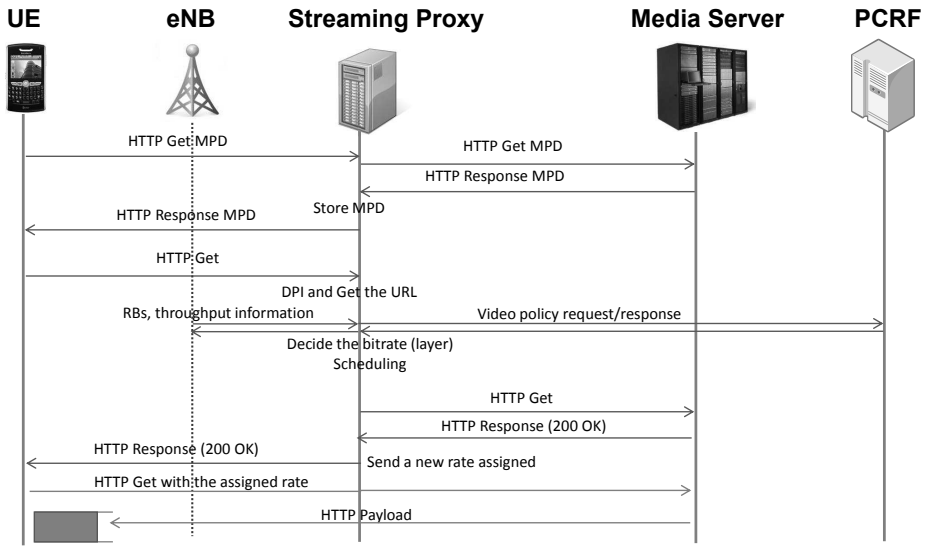


Fig. 4.5. A message flow of SP without cache.

SP)[2].

4.4.2 Message Flows

The overall procedures of the proposed scheme for the SPs with and without a cache are illustrated in Figs. 4.4 and 4.5, respectively. In both cases, when a UE sends a HTTP message to request an Media Presentation Description (MPD) file, the SP detects and parses it, and stores the MPD from the media server while forwarding the corresponding packets. The MPD is used to extract the available bitrates for each video.

Once the UE requests a segment after receiving the MPD, the SPs with and with-

[2]The widely used deep packet inspection (DPI) facilities can be used here. Since the replacement happens per video segment (not per packet), we believe the overhead can be acceptable.

Table. 4.1. Comparison of DASH techniques.

	Proposed Scheme	AVIS [36]	AGBR [37]	Client-side approaches
Rate enforcement	Modify HTTP requests & Set GBR values	CellSlice [46]	Set GBR values	Not applicable
Rate decision maker	Gateway	Gateway	eNB	Client
Stability	Stateful rate selection	Penalty function	Not validated	Depends on algorithm

out a cache have different operations. In the case of the SP with a cache (Fig. 4.4), the SP performs the DPI to get the URL, and requests the video segments with the highest quality to the media server. Note that the SP splits the TCP connection between the UE and the media server. For each segment delivered from the media server, the SP encodes the video segment into multiple versions (of multiple bitrates), which are stored locally. For each UE, after selecting the bitrate based on the RB, throughput, and PCRF policy information, the SP sends video segments of the selected bitrate. As to the SP without a cache (Fig. 4.5), after selecting the bitrate, the SP performs the DPI for each HTTP Response packet, and writes the assigned bitrate on a non-necessary field of the packet. When the UE receives the HTTP Response packet, it updates its bandwidth estimation according to the received bitrate information, and selects the video segment of the assigned bitrate the next time.

4.4.3 Characteristics

We summarize the characteristics of the proposed scheme by comparing with the state-of-the-art HTTP adaptive streaming techniques in Table 4.1. Note that AVIS [36] and AGBR [37] are the network-side approaches including our scheme, and "Client-side approaches" represents the common characteristics of the client-side approaches such as FESTIVE [35]. In general, the network-side approaches achieve higher performance in terms of throughput than the client-side ones since they can utilize the wireless link information directly, but require an additional function or an entity on the existing cellular systems. We compare the proposed scheme and the other techniques from three perspectives: (i) the rate enforcement method, (ii) the position of the rate decision-maker, and (iii) the algorithm for stability.

Among the HTTP adaptive streaming techniques, AVIS is somewhat similar to the proposed scheme, but it differs from AVIS in the following points. First, unlike AVIS, our scheme incorporates the resource allocation for video and data flows into a single optimization framework. Since the absence of consideration on the data users' utility results in inefficient utilization of resources, the proposed scheme can optimize the total utility of all the users. Second, the proposed scheme can avoid the instability caused by players who select bitrates differently by AVIS. Our scheme enforces the bitrates at SP by modifying the HTTP Response messages sent to UEs.

AGBR enforces the bitrates by setting the GBR value at every bitrate decision interval. For this, a new functionality that decides the bitrates needs to be added at each eNB in AGBR. On the other hand, the SP in our scheme can manage multiple eNBs since it will be located behind a Serving Gateway. To achieve the stability, our scheme selects the bitrate by providing the stateful rate selection mechanism

Table. 4.2. Notation for calculating the bitrates and allocating resources.

U	the set of all video users (flows)
D	the set of all data users (flows)
R_u^i	the bitrate for ith segment of user u
L_u^i	the bitrate index for ith segment of user u
β_u	the video-dependent utility parameter for user u
θ_u	the screen size-dependent utility parameter for user u
α	the balancing parameter between video and data traffic
n	the number of data flows
r	the portion of RBs for video flows
r_k	the kth bitrate for the video
T	the length of the bitrate assignment interval
b_u^i	the number of bytes transmitted to user u in interval i
n_u^i	the number of RBs assigned to user u in interval i
N_{total}	the total number of RBs in one interval

proposed by Jiang *et al.* [35]. The stateful rate selection is a mechanism that the next bitrate is decided by considering the current bitrate [35]. For example, we can assign an UE with the lower/higher bitrate to ramp up/down aggressively or gently. AGBR constantly changes the GBR parameters so that the UEs can adapt to fluctuating link bandwidth, but the stability of AGBR has not been investigated.

Client-side algorithms also take into account the stability in the bitrate selection process, but their mechanisms perform poorly in LTE wireless networks, which will be detailed in Section 4.6.

4.5 Bitrate Assignment

In this section, we explain how to calculate the bitrates for video users and to allocate resources to video and data flows in the proposed scheme. The following

Algorithm 4.1 Algorithm for calculating the video bitrates and allocating resources.

Output : R_u^i, r
maximize : $\sum_{u \in U} \beta_u (1 - \frac{\theta_u}{R_u^i}) + \alpha n \log(1 - r)$
subject to:
$[C1] R_u^i \in \{r_1, ..., r_u\}\ for\ \forall u \in U$
$[C2] \sum_{u \in U} \frac{T \cdot R_u^i}{\frac{b_u^{i-1}}{n_u^{i-1}}} \leq r N_{total}$

assumptions are made for simplicity: (1) all video flows are based on DASH, (2) all data flows use TCP, and (3) the number of data flows are predictable. Due to the recent popularity of DASH such as Netflix and YouTube, we believe the first assumption is acceptable. Also, Gember *et al.* [47] and Zhang *et al.* [48] showed that the amount of UDP traffic is marginal compared to that of TCP traffic.

4.5.1 Bitrate Calculation

Table 4.2 shows the notation of the algorithms in this chapter. The bitrate calculation algorithm is formulated as a mixed discrete non-linear problem [49], which (periodically) runs on every bitrate assignment interval (BAI). We want to decide the bitrate for each DASH video-user, and the portion of the resource blocks (RBs) (that will be used for video users) to the total RBs available for each BAI. Our algorithm for bitrate calculation is described in Algorithm 4.1.

The aim of Algorithm 4.1 is to maximize the total utility of video- and data-users. The utility function for video- and data-users is defined as $\sum_{u \in U} \beta_u (1 - \frac{\theta_u}{R_u^i}) + \alpha n \log(1 - r)$, as defined in [37]. The bitrate for each user is limited to one of the bitrates defined in the MPD for the video that the user is downloading (constraint [C1]). Constraint [C2] indicates that the sum of RBs allocated to video-users is limited to

the total number of RBs for all the video-users, where N_{total} is the total number of RBs in BAI. In [C2], we calculate the number of bytes that have been transmitted per RB in the previous BAI ($\frac{b_u^{i-1}}{n_u^{i-1}}$) so that we can estimate the number of RBs needed for each user.

Rather than deciding the rate for each data flow, we simply use one term for all the data flows, $\alpha n \log(1-r)$. We assume that (i) the sum of the throughput of all the data flows is proportional to the ratio of resources for data flows (i.e., $1-r$), and (ii) the ratio among the throughput values of individual flows is maintained throughout the lifetime of the flow. Let T_u^i is the throughput of the data flow of user u for bitrate assignment interval i when there is no video flow. Then, the sum of the utility of all the data flows can be represented as $\sum_{u \in D} \log(T_u^i \cdot (1-r))$ where D is the set of all the data users (or flows). T_u^i can be deemed as a constant, and be omitted; thus, the utility can be simply represented as $n \log(1-r)$. We multiply α to balance the resource allocation between video and data flows.

4.5.2 Enhancing Stability

Cranley *et al.* [50] and Balachandran *et al.* [51] pointed out that the frequent changes of bitrates can adversely affect the experience of video-users. Therefore, we devise a technique for enhancing the stability in bitrate selection. First, we limit the maximum bitrate that each user can select in interval i to the next higher bitrate index to the one in the previous interval $i-1$ (i.e., $L_u^{i-1} \leq L_u^{i-1}+1$). Then, if Algorithm 4.1 selects the bitrate index for interval i equal to L_u^{i-1}, the selected one applied. If the algorithm decides to increment the bitrate index for interval i to be $L_u^{i-1}+1$, the selected one is held until $L_u^{i-1}+1$ is repeatedly selected for β times. In this way,

we conservatively increase the bitrate, so that the negative effect of frequent bitrate changes is reduced, which also helps to achieve the fair rate allocations among users. This approach is similar to Jiang *et al.* [35], but the detailed realization is somewhat different. The rate decrease in the proposed scheme is determined by the calculated bitrate by the optimization algorithm, whereas FESTIVE drops the bitrate to the next lower index even if the estimated bitrate is lower than p times the one in the previous interval ($p < 1$). Note that there is a tradeoff between the efficiency of resource utilization and the stability of video quality in this technique.

4.5.3 Algorithm for Continuous Bitrates

In order to reduce the computational complexity (NP-hard) of Algorithm 4.1, we devise another algorithm for the continuous bitrate assignment. To this end, we convert the first constraint [C1] into a continuous version, i.e., $r_1 \leq R_u^i \leq bitrate(L_u^{i-1}+1)\ for\ \forall u \in U$ where the $bitrate(l)$ indicates the bitrate for the bitrate index l. After calculating R_u^i, the final bitrate is selected as the largest $r_k \leq R_u^i$. We do not further optimize this solution, but the performance degradation is marginal in terms of the average bitrate compared to the discrete algorithm to be shown in Subsection 4.6.3.

4.5.4 Handling the Bottleneck of Wired Networks

So far, we have focused on the wireless links of LTE networks. However, the bandwidth of wired links can be a bottleneck for HTTP video streaming. This situation can be addressed by adding the following functions to the SP in the proposed scheme. The first function measures the throughput of each video flow periodically; i.e., T_u^i is the throughput observed in ith bitrate assignment interval for user u. The

second function allocates the maximum bitrate that can be selected in interval i as the $min(L_u^{i-1}+1, floor(T_u^{i-1}))$ where $floor$ operation means the maximum bitrate index which is equal to or less than T_u^i. As time goes on, the bottleneck point/bandwidth can vary, but the proposed scheme can handle either case since it considers the minimum of the wired link bandwidth (by using $floor(T_u^{i-1})$) and the wireless link bandwidth (by Algorithm 4.1). It retains its stability since our stateful rate selection mechanism still holds for both cases (by limiting with $L_u^{i-1}+1$).

Table. 4.3. Simulation settings

Simulator	ns-3 3.18.1
Simulation time	1200 seconds
Territory	2000m x 2000m
Number of video UEs	8
Placement of UEs	Random
Fading model	Trace-based model
Video segment duration	10 seconds
Video bitrates	100, 250, 500, 1000, 2000, 3000 Kbps
TCP	Westwood
Scheduler	Priority set scheduler

Table. 4.4. Default values of parameters for the three schemes.

Proposed Scheme	α	1.0
	β	4
	θ_u	0.2 Mbps
	β_u	10
FESTIVE	k	4
	p	0.85
	α	12
AVIS	$alpha$	0.01
	W	150

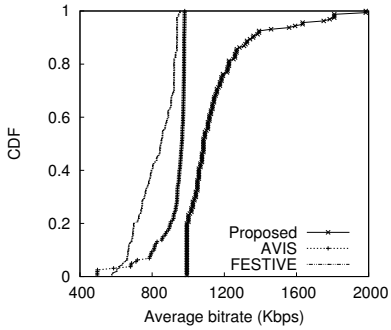
4.6 Simulation

We conduct a comprehensive simulation study to evaluate the proposed scheme on ns-3 simulator [52] with the LTE module. For comparison purposes, we also evaluate a client-side algorithm (FESTIVE [35]) and a network-side algorithm (AVIS [36]). For AVIS, we run a simple rate control algorithm on a UE that requests the highest possible rate based on the estimated throughput. We also set the GBR / MBR using the scheduler in the base station for AVIS instead of using the resource slicing techniques.

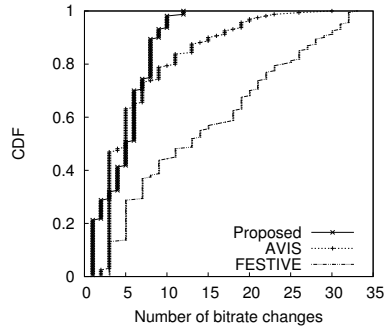
Table 4.3 summarizes the simulation settings. We use TCP Westwood since it is appropriate for wireless environments. For the base station scheduler, we use the Priority Set Scheduler [53] that is QoS-aware. We modified the Priority Set Scheduler module in the ns-3 simulator to add the MBR assignment, and to retrieve the information about the assigned RBs and transmitted bytes for each user. We set the parameters for each scheme as shown in Table 4.4. Refer to references for parameters for FESTIVE and AVIS. For each plot, we run the simulation 20 times.

4.6.1 Static Scenario

We first consider a static scenario where all the users have no mobility. Figs 4.6(a) and 4.6(b) show the CDFs of the throughput values and numbers of bitrate changes for each scheme. The average throughput of the proposed scheme increases by 24% and 39% compared to those of AVIS and FESTIVE, respectively. On the other hand, the average number of bitrate changes of the proposed scheme decreases by 26% and 66% compared to those of AVIS and FESTIVE, respectively. Recall that AVIS

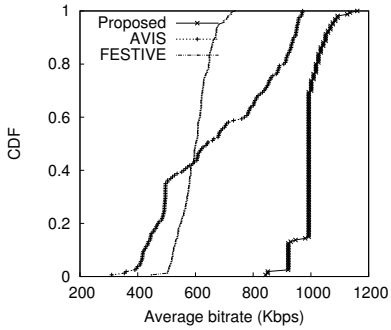


(a) CDF of the throughput values.

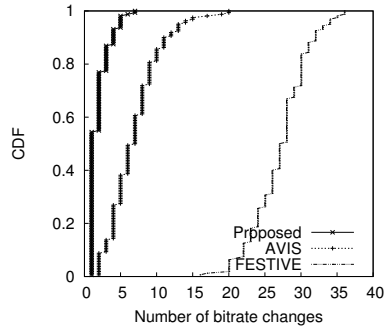


(b) CDF of the numbers of bitrate changes.

Fig. 4.6. Simulation results in static scenario.



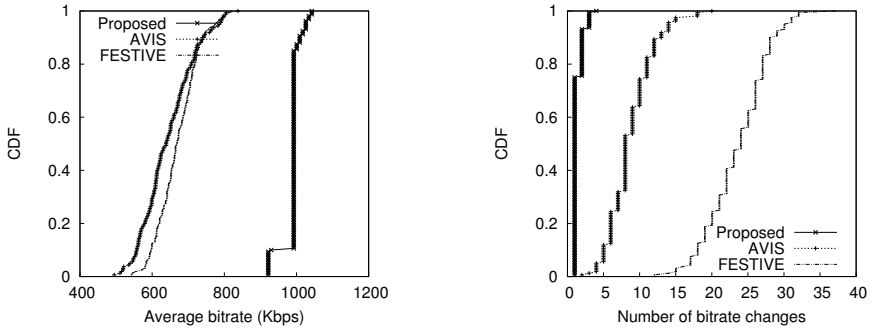
(a) CDF of the throughput values.



(b) CDF of the numbers of bitrate changes.

Fig. 4.7. Simulation results in a pedestrian scenario.

sets only GBR / MBR and lets the rate controller in a UE select the actual video bitrate, which results in lower throughput and instability. We can observe a mismatch between the bitrates set by AVIS and bitrates selected by UEs. FESTIVE performs worse than the others due to unawareness of the link conditions in a cell. The average Jain's fairness index is 0.989, 0.989, and 0.986 for the proposed scheme, AVIS, and FESTIVE, respectively, meaning that the fairness is good and comparable across the schemes in the static scenarios.



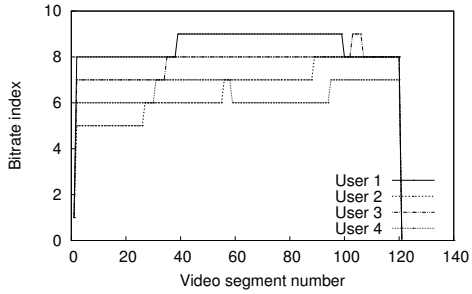
(a) CDF of the throughput values.

(b) CDF of the numbers of bitrate changes.

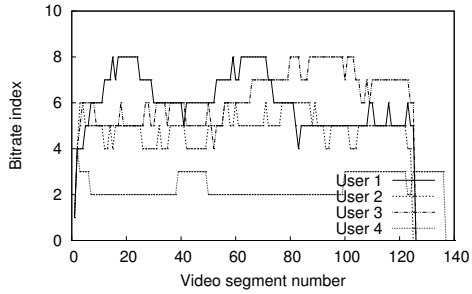
Fig. 4.8. Simulation results in vehicular scenario.

4.6.2 Mobile Scenarios

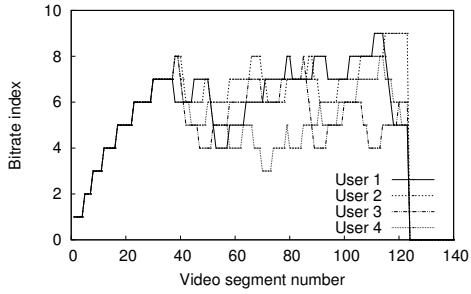
We next consider two mobile scenarios: (i) pedestrian and (ii) vehicular. Figs. 4.7(a) and 4.7(b) show the simulation results in the pedestrian scenario. The average throughput of the proposed scheme is enhanced by 48% and 66% compared to those of AVIS and FESTIVE, respectively. On the other hand, the average number of bitrate changes decreases by 73% and 93% compared to those of AVIS and FESTIVE, respectively. The throughput of the proposed scheme in the pedestrian scenario decreases compared to the one in the static scenario, but the stability of our scheme in the pedestrian scenario is much more enhanced. In the proposed scheme, the rate increase occurs less (than FESTIVE and AVIS) in presence of the high variation of link bandwidth in mobile scenarios, which leads to the conservative use of the wireless resources. The average Jain's fairness index is 0.998, 0.923, and 0.992 for the proposed scheme, AVIS, FESTIVE, respectively, which means our scheme and FESTIVE achieve the almost perfect fairness, but AVIS shows the less fairness. We can observe similar results in the vehicular scenario as shown in Figs. 4.8(a) and 4.8(b). Our scheme



(a) The bitrate variation of the proposed scheme.



(b) The bitrate variation of AVIS.

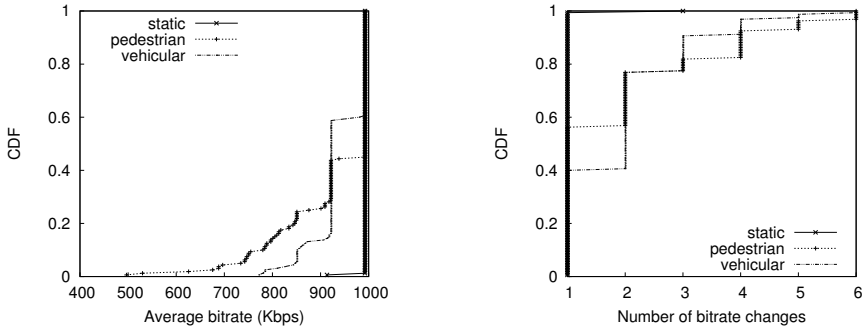


(c) The bitrate variation of FESTIVE.

Fig. 4.9. An illustrative example of bitrate variations.

shows 53% and 47% improvements in the average throughput compared to AVIS and FESTIVE. The average number of bitrate changes decreases by 85% and 95% compared to those of AVIS and FESTIVE, respectively. The average Jain's fairness index shows a similar pattern: 0.999, 0.988, and 0.993 for the proposed scheme, AVIS, and FESTIVE, respectively.

To see how the bitrate changes as time goes on for each algorithm, we take an illustrative example for the pedestrian scenario. Figs. 4.9(a), 4.9(b), and 4.9(c) show the example of bitrate variations for the proposed scheme, AVIS, and FESTIVE, re-



(a) CDF of the throughput values.

(b) CDF of the numbers of bitrate changes.

Fig. 4.10. Analysis for continuous bitrates in the proposed scheme.

spectively. Our scheme tends to choose higher bitrates, and change the bitrates less compared to the other algorithms. In addition, it takes the full advantage of the bandwidth for all the segments (during current playing) rather than buffering many segments for future video playing as in AVIS. FESTIVE gradually increases the bitrate from the lowest one since it cannot aware of the overall channel state.

4.6.3 Algorithm for Continuous Bitrates

To reduce the computational complexity of the original algorithm, we suggested a continuous version of the proposed scheme for the bitrate assignment in Subsection 4.5.3. We evaluate the continuous bitrate algorithm in Figs. 4.10(a) and 4.10(b). The average throughput slightly drops by 14%, 8%, and 6% for static, pedestrian, and vehicular scenarios compared to those in the original proposed algorithm. The stability varies depending on scenarios. The number of bitrate changes decreases by 80% in the static scenario, but increases by 2% and 48% in the pedestrian and vehicular scenarios, respectively. The average Jain's fairness index is 0.999, 0.989, and 0.997

Fig. 4.11. Testing environment.

for the static, pedestrian, and vehicular scenarios, respectively.

4.7 Experiments

In this section, we implement a prototype of the proposed solution in our testbed as shown in Fig. 4.11. We have 1 Streaming Proxy, 1 media server, 1 LTE femtocell eNB, 1 Core Network Emulator (CNE) [3], and 4 laptops as UEs. The overall composition of our testbed is illustrated in Fig. 4.12.

[3] A commercial Evolved Packet Core (EPC) network emulator is used in the experiments to provide MME/SGW/PGW operations including LTE bearer services and their QoS supports. [54]

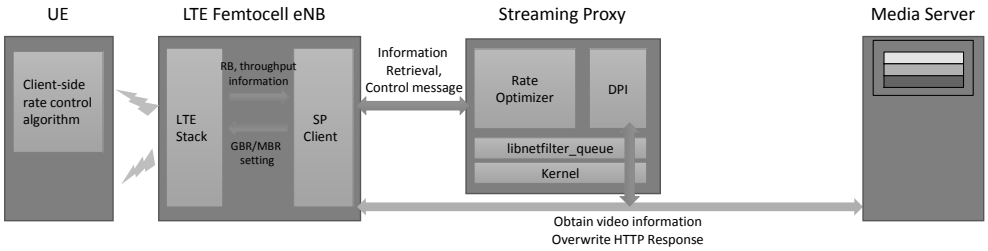


Fig. 4.12. Logical network entity diagram of the testbed.

4.7.1 Implementation of DASH Player

In order to implement the DASH player in UE, we modified the MPEG-DASH /Media Source demo player [55] which is implemented using Javascript and HTML. MPEG-DASH/Media Source demo player provides a rate control algorithm, which we call GOOGLE algorithm. GOOGLE maintains the bandwidth estimation based on the history of previously received segments, and selects the largest available video rate which is less than 0.85 times the bandwidth estimation. It keeps both the slow bandwidth estimation which reflects long-term bandwidth trend, and fast bandwidth estimation which reflects short-term bandwidth variation. It uses the minimum between the two when it selects the video rate. We also implemented FESTIVE algorithm on the DASH player. For the proposed solution, we added a simple function that parses and applies the decided rate by SP which is recorded in the HTTP Response packets. We also implemented functions for recording and displaying the information about the segment transmission, bitrates, and buffer status.

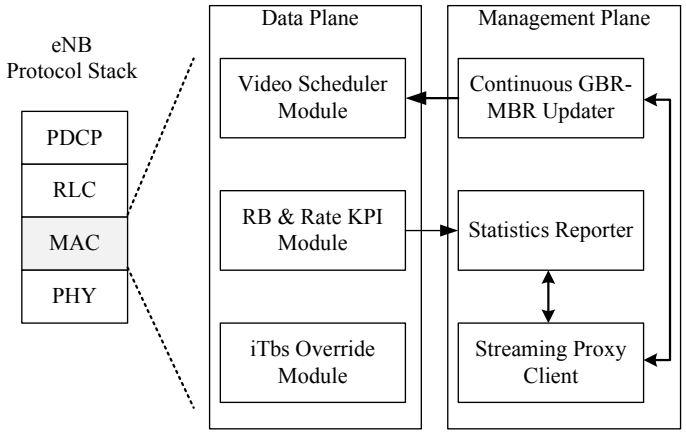


Fig. 4.13. New modules introduced in the eNB software architecture.

4.7.2 Implementation of eNB

We use a commercially deployed LTE femtocell eNodeB (eNB) [56]. It supports 10Mhz-bandwidth FDD operations on E-UTRA Band 7 [57], and 50 RBs are available per transmission time interval (TTI). We developed a number of new modules in the LTE medium access control (MAC) layer of an eNB protocol stack, as shown in Fig. 4.13.

- *Video Scheduler Module*: It performs a GBR-based per-TTI scheduling for video traffics. Fig 4.14 shows that how overall scheduling works for video and data flows. Our scheduling consists of two phases. Video Scheduler Module performs GBR-based scheduling of video flows in the first phase. It assigns the guaranteed RBs for videos. If there is any remaining RB after completing the first phase, it is assigned for data and video flows by an existing proportional fair scheduler. In our test eNB, a PF scheduler is implemented considering three

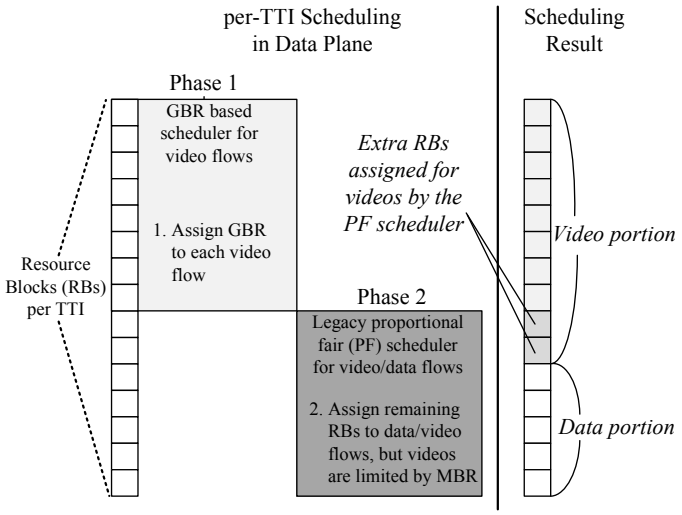


Fig. 4.14. Video and data scheduling in the eNB data plane.

factors: channel quality information (CQI) of UEs, data served rate, and QoS class identifier (QCI). CQI aims at throughput maximization, data served rate aims at fairness, and QCI aims at providing QoS. An MBR setting is also applied in the PF scheduler to balance RB assignments of video flows and those of data flows.

- *RB & Rate Trace Module*: It includes RB and rate counters for each video traffic flow. RB counters are accumulated by the Video Scheduler Module. Depending on the wireless channel reports from a UE, transmission rate (i.e., bytes per RB) can be changed. With RB & rate traces, we can also estimate a UE's channel status.

- *iTbs Override Module*: Due to the nature of time-varying wireless channels, the channel quality reports from a UE are not always the same, although the UE is

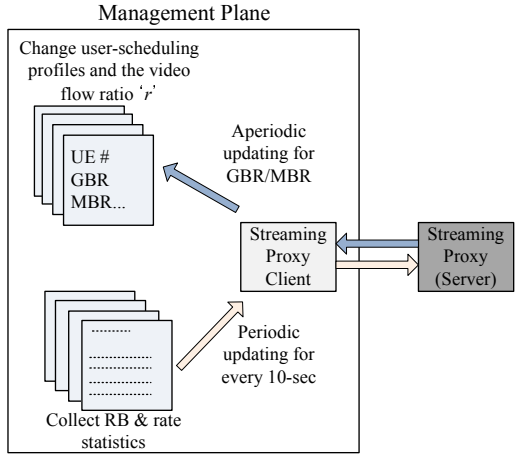


Fig. 4.15. A work flow in the eNB management plane.

not moving. If the contents of channel quality reports are different, the rate will be changed by selecting a different transport block size (TBS). Each TBS index (iTbs) defines its modulation and coding scheme [58]. To be a fair comparison, we need to emulate a fixed wireless channel by overriding the index of TBS.

- *Continuous GBR-MBR Updater*: Originally GBR and MBR are assigned when the traffic bearer is set up. However, we need to change the values of GBR and MBR dynamically and continuous whenever the Streaming Proxy requests.

- *Statistics Reporter*: It collects counters from the RB & Rate trace module, and make a statistical report whenever a report timer is expired. In our experiments, we set a value of the report timer to 10 seconds.

- *Streaming Proxy Client*: Fig. 4.15 illustrates a work flow in the management plane of the eNB. First, the Streaming Proxy Client connects to the Streaming Proxy Server located in the LTE core network. Then, a periodic statistics re-

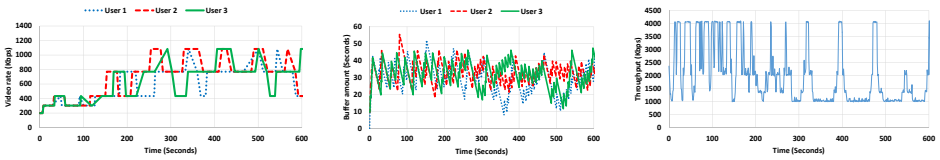
port made by Statistics Reporter is transmitted to the server. The server (i.e., Streaming Proxy) calculates GBR/MBR values according to its bitrate assignment algorithm, and send new GBR/MBR settings if needed. Therefore the process of setting GBR/MBR is aperiodic.

4.7.3 Implementation of Streaming Proxy

The Streaming Proxy (SP) consists of two components: Rate optimizer and DPI. The SP receives the video information (from DPI), as well as the RB and throughput information (from the SP client). Based on these information, it runs Rate optimizer and decides the bitrate for each video-user. Then, the chosen bitrates are sent to the SP client and the DPI, which enforces them by setting the GBR values and overwriting HTTP Response packets, respectively. We implement DPI using *libnetfilter_queue* [59] which is a user-space library providing an API to packets that have been queued by the kernel packet filter. DPI conducts the roles of obtaining the video information, and applying the decided bitrates to the video segments.

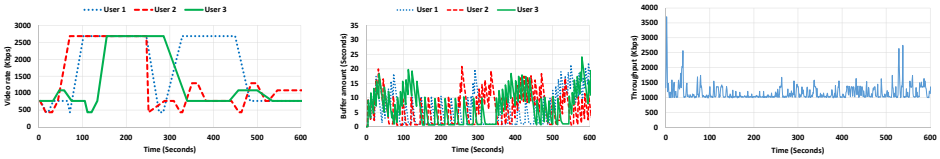
4.7.4 Experimental Results

We compare our solution with two client-side adaptation algorithms, FESTIVE and GOOGLE. We encoded a video into 200, 310, 450, 790, 1100, 1320, 2280, 2750 Kbps for the experiments. We run 3 video flows on 3 notebooks and set one data flow on a laptop by executing the Iperf [60]. We test the algorithms in static scenario in which the wireless channel status does not change, and also in dynamic scenario in which it changes dynamically. The same PF scheduler is used for all test cases. For the dynamic scenario, we assume a 4-min cycle that increases the index of TBS



(a) Video rates (b) Buffer usage (c) Data throughput

Fig. 4.16. Performance of the FESTIVE in a static scenario

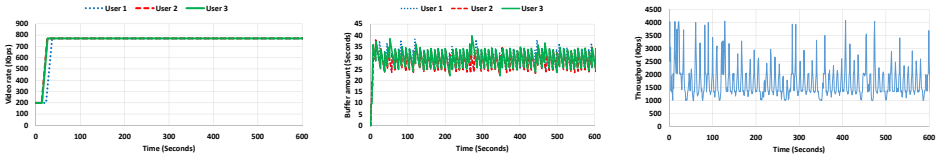


(a) Video rates (b) Buffer usage (c) Data throughput

Fig. 4.17. Performance of the GOOGLE in a static scenario

gradually from 1 to 12 during 2 minutes, and then decreases from 12 to 1 for the next 2 minutes. Each UE has its own 4-min wireless channel cycle and it starts at a different time. Therefore UEs experience different wireless channels at a time. We conduct each experiment for 10 minutes.

Figs. 4.16, 4.17, and 4.18 illustrate the performance of FESTIVE, GOOGLE, and our scheme in static scenario, respectively. In this static scenario, the legacy PF scheduler serves for the fairness between users, because all CQIs of UEs are fixed



(a) Video rates (b) Buffer usage (c) Data throughput

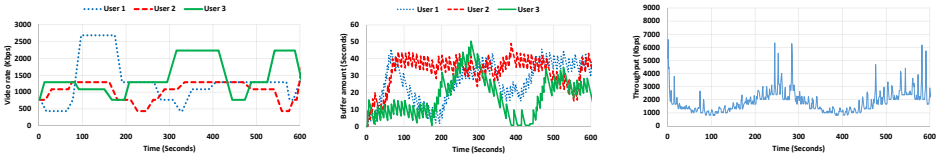
Fig. 4.18. Performance of the proposed scheme in a static scenario

at the same level, and we used the same QCI values for video and data flows. Two different client-side algorithms, FESTIVE and GOOGLE, show opposite behaviors. FESTIVE selects the video rate conservatively resulting in slow convergence to a stable states, and the relatively large data throughput. However, due to its ignorance of the overall channel status, it changes the bitrate frequently around the optimal bitrates degrading users' QoEs. GOOGLE selects the video rate overly aggressively. It sometimes chooses the highest video rate, but this causes frequent pauses of playing a video due to the buffer underflow. Note that during the time periods in which the buffer amount is less than 1 second, the video play is paused. GOOGLE assigns the radio resources almost evenly among the video flows and the data flow. Our proposed algorithm selects the video rate extremely stably. It chooses a video rate of 790 Kbps constantly except for the initial periods in which it selects the basic minimum rate. The number of bitrate changes in our scheme is 1; it is smaller than FESTIVE (20.3) and GOOGLE (9.7). The average video rate of our scheme is 726 Kbps, while those of FESTIVE and GOOGLE are 638 Kbps and 1151 Kbps, respectively. The average Jain's fairness index is 0.999, 0.998, 0.990 for our scheme, FESTIVE, and GOOGLE, respectively. The average data-user's throughput of our scheme is 1800 Kbps, while those of FESTIVE and GOOGLE are 2512 Kbps and 1140 Kbps, respectively. Our scheme also maintains a stable amount of buffers during the whole of the experimental period.

Figs. 4.19, 4.20, and 4.21 show the performance of FESTIVE, GOOGLE, and our scheme in a dynamic scenario. In this dynamic scenario, the legacy PF scheduler chooses those candidates with good channel conditions (CQIs) if served rates of users are similar. If there are someone with lower CQIs to starve, the PF sched-

(a) Video rates (b) Buffer usage (c) Data throughput

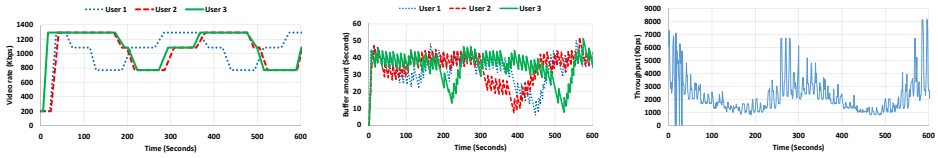
Fig. 4.19. Performance of the FESTIVE in a dynamic scenario



(a) Video rates (b) Buffer usage (c) Data throughput

Fig. 4.20. Performance of the GOOGLE in a dynamic scenario

uler tries to be fair, hence overall throughput is compromised. FESTIVE exhibits a similar behavior in selecting the bitrates and the throughput of a data flow, as shown in Fig. 4.16. The bitrates oscillate around the optimal rates with a large difference between the peak and the lowest values. The GOOGLE's aggressive behaviors in selecting the bitrates cause the frequent pauses of playing a video as in the static scenario. The proposed scheme changes the bitrates according to the variation of wireless channel status, but the differences between the peak and the lowest values



(a) Video rates (b) Buffer usage (c) Data throughput

Fig. 4.21. Performance of the proposed scheme in a dynamic scenario

are much smaller than those of FESTIVE and GOOGLE on the average. The number of bitrate changes in our scheme is 11.3; it is smaller than FESTIVE (22.7) and GOOGLE (14). The average video rate of our scheme is 1025 Kbps, while those of FESTIVE and GOOGLE are 839 Kbps and 1297 Kbps, respectively. The average Jain's fairness is 0.998, 0.998, and 0.997 for our scheme, FESTIVE, and GOOGLE, respectively. The average data-user's throughput of our scheme is 2300 Kbps, while those of FESTIVE and GOOGLE are 3870 Kbps and 1870 Kbps, respectively. Although our scheme does not directly touch the buffers in a DASH client, it does not cause a buffer underflow even in the worst channel status.

4.8 Conclusion

In this chapter, we investigate the problems of the existing HAS techniques in LTE networks, which are divided into client-side and network-side approaches. For the client-side approach, we find that a video-user tends to be assigned less bandwidth compared to a data-user. In the network-side approach, we find the stability of video quality is difficult to achieve. To provide a fair, efficient, and stable video streaming services by addressing the above problems, we propose a network-side HAS solution that optimizes the total utility of all users in a cell, while maintaining the stable video quality. Our simulation results show that the proposed scheme significantly enhances the average throughput (or bitrate) and stability of video quality. Our scheme is also designed to achieve the balance between video- and data-users, compared to current state-of-the-art client-side and network-side approaches. Experiments conducted on our end-to-end HAS testbed using real video traffic also show the performance en-

hancement and practicability of the proposed scheme.

Chapter 5

Summary & Future Work

This dissertation proposed the performance enhancement techniques of real-time and non-real-time video delivery services in LTE networks.

First, we investigated how to reduce the enhanced Multimedia Broadcast and Multicast Service (eMBMS) disruption for delivering a real-time video to multiple users in a LTE network. LTE has defined Multimedia Broadcast and multicast service over a Single Frequency Network (MBSFN) area which can reduce eMBMS service disruption due to handovers. However, reducing handover delay is achieved by making all Evolved Node Bs (eNBs) send the same packets in the MBSFN area. This has motivated us to propose an MBSFN area planning scheme based on location management areas (LMAs) in order to save wireless link bandwidth while keeping the service disruption time at an acceptable level. Our scheme partitions an MBSFN area into multiple LMAs to balance the average handover delay and the bandwidth usage overhead. We have presented a novel mathematical model of the service disruption time, bandwidth usage and blocking probability which consider user mobility, distribution and eMBMS session popularity. We also studied how to decide MBSFN area and LMA sizes, which can make the best use of bandwidth in maintaining the quality of eMBMS services. Our results suggested that LMA-based MBSFN scheme would delivery more efficient multicast and broadcast services over LTE networks.

Second, we studied a proactive transmission-based approach assuming the same

LMA-based MBSFN framework. The proposed scheme transmits eMBMS packets not only to the LMAs with eMBMS users, but also to their neighbor LMAs without eMBMS users probabilistically and proactively. By considering session popularity, user distribution, and user mobility, the probability of proactive transmissions in the neighbor LMAs without users is determined for each eMBMS session. Through extensive simulations we also revealed that our proposed scheme can reduce the average handover delay effectively.

Third, we investigated the problems of the existing HTTP adaptive streaming (HAS) techniques in LTE networks which has become one of the most popular solutions especially for delivering non-real-time video content (i.e., video-on-demand). In the existing approaches, we found throughput unfairness between video users and data users, and we also found the unstability of video quality. To provide a fair, efficient, and stable video streaming services by addressing the above problems, we propose a network-side HAS solution that optimizes the total utility of all users in a cell, while maintaining the stable video quality.

This dissertation assumes LTE network-based operations, but our work can be applied to other existing and/or future access network technologies. For example, the WiMAX network is also including Multicast and Broadcast Service with 'MBS Zone' operations [61], which are similar to the LTE MBSFN. Our LMA-based approaches can easily be integated to the WiMAX MBS Zone, utilizing the location information of users provided by the definition of WiMAX 'Paging Group'. Any future cellular network will have a similar concept of multicast/broadcast service area like an MBSFN, and identifying the location of users is also getting more important in any network. Our bit-rate assignment algorithm of the proposed HAS solution

is basically independent to the access technology, but how to apply it to a specific technology may differ.

Our work opens up several research avenues for future work. Investigating how the MBSFN area and LMA can be organized automatically is an interesting topic (i.e., self-constructing and self-optimizing MBSFN). Since a single eNB can belong to multiple MBSFN areas (and LMAs), the MBSFN area / LMA re-selection is an imminent topic for User Equipments (UEs). The HAS techniques will evolve continuously, and they will bring up a challenging issue, how to be deeply integrated with LTE eMBMS framework for further optimization.

Bibliography

[1] "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 11)," *3GPP TS 36.300*, v11.10.0, http://www.3gpp.org/, Jun. 2014.

[2] "Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description (Release 11)," *3GPP TS 23.246*, v11.1.0, http://www.3gpp.org/, Mar. 2012.

[3] D. Lecompte and F. Gabin, "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: overview and Rel-11 enhancements," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 68–74, Nov. 2012.

[4] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet," *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, Apr. 2011.

[5] "Adobe http dynamic streaming," http://www.adobe.com/products/hds-dynamic-streaming.html.

[6] "Apple http live streaming," http://developer.apple.com/streaming.

[7] "Microsoft smooth streaming," http://www.iis.net/downloads/microsoft/smooth-streaming.

[8] A. Dutta, J. Chennikara, W. Chen, O. Altintas, and H. Schulzrinne, "Multicasting Streaming Media to Mobile Users," *IEEE Communications Magazine*, vol. 41, no. 10, pp. 81–89, Oct. 2003.

[9] R. Koodli, *Fast Handovers for Mobile IPv6*, IETF RFC 4068, Jul. 2005.

[10] R. Ramjee, K. Varadhan, L. Salgarelli, S. Thuel, S. Y. Wang, and T. L. Porta, "HAWAII: A Domain-based Approach for Supporting Mobility in Wide-area Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 396–410, Jun. 2002.

[11] M. Hauge and O. Kure, “Multicast in 3G Networks: Employment of Existing IP Multicast Protocols in UMTS,” *Proc. 5th ACM International Workshop Wireless Mobile Multimedia*, pp. 96–103, 2002.

[12] R. Rummler, Y. W. Chung, and A. H. Aghvami, “Modeling and Analysis of an Efficient Multicast Mechanism for UMTS,” *IEEE Transactions on Vehicular Technology*, vol. 54, no. 1, pp. 350–365, Jan. 2005.

[13] A. Alexiou and C. Bouras, “Multicast in UMTS: Evaluation and Recommendations,” *Wiley Wireless Communications and Mobile Computing*, vol. 8, no. 4, pp. 463–481, May 2008.

[14] A. Alexiou, D. Antonellis, and C. Bouras, “An Efficient Mechanism for Multicast Data Transmission in UMTS,” *Wireless Personal Communications*, vol. 44, no. 4, pp. 455–471, Mar. 2008.

[15] *Mobile WiMAX - Part I: A Technical Overview and Performance Evaluation*, http://www.wimaxforum.org/documents/, Aug. 2006.

[16] S. Sengupta, M. Chatterjee, and S. Ganguly, “Improving Quality of VoIP Streams over WiMax,” *IEEE Transactions on Computers*, vol. 57, no. 2, pp. 145–156, Feb. 2008.

[17] J. She, F. Hou, P.-H. Ho, and L.-L. Xie, “IPTV over WiMAX: Key Success Factors, Challenges, and Solutions,” *IEEE Communications Magazine*, vol. 45, no. 8, pp. 87–93, Aug. 2007.

[18] W. Jiao, P. Jiang and Y. Ma, “Fast Handover Scheme for Real-Time Applications in Mobile WiMAX,” *Proc. IEEE International Conference on Communications*, pp. 6038–6042, 2007.

[19] J. Wang, M. Venkatachalam, and Y. Fang, “System Architecture and Cross-Layer Optimization of Video Broadcast over WiMAX,” *IEEE Journals on Selected Areas in Communications*, vol. 25, no. 4, pp. 712–721, May 2007.

[20] S. Parkvall, E. Englund, M. Lundevall, and J. Torsner, "Evolving 3G Mobile Systems: Broadband and Broadcast Services in WCDMA," *IEEE Communications Magazine*, vol. 44, no. 2, pp. 30–36, Feb. 2006.

[21] A. Jiang, C. Feng and T. Zhang, "Research on Resource Allocation in Multi-cell MBMS Single Frequency Networks," *Proc. IEEE International Conference on Wireless and Optical Communications Networks*, pp. 1–5, 2010.

[22] M. F. Azman, N. F. Tuban, K. A. Noordin and M. F. Ismail, "Genetic Algorithm Approach for Solving Radio Resource Allocation for Overlapping MBSFN Area," *Proc. IEEE International Conference on Information Technology and Multimedia*, pp. 1–5, 2011.

[23] A. Alexiou, C. Bouras, V. Kokkinos and G. Tsichritzis, "Communication Cost Analysis of MBSFN in LTE," *Proc. IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, pp. 1366–1371, 2010.

[24] D.-N. Yang and M.-S. Chen, "Efficient Resource Allocation for Wireless Multicast," *IEEE Transactions on Mobile Computing*, vol. 7, no. 4, pp. 387–400, Apr. 2008.

[25] Y.-B. Lin, "A Multicast Mechanism for Mobile Networks," *IEEE Communications Letters*, vol. 5, no. 11, pp. 450–452, Nov. 2001.

[26] L. Berslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications," *Proc. IEEE INFOCOM*, pp. 126–134, Mar. 1999.

[27] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaa, and L. E. Meester, *A Modern Introduction to Probability and Statistics. Understanding Why and How*, Springer, 2005.

[28] Y. Fang and I. Chlamtac, "Teletraffic Analysis and Mobility Modeling of PCS Networks," *IEEE Transactions on Communications*, vol. 47, no. 7, pp. 1062–1072, Jul. 1999.

[29] S.-R. Yang and Y.-B. Lin, "Performance Evaluation of Location Management in UMTS," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 6, pp. 1603–1615, Nov. 2003.

[30] Y.-B. Lin, "Reducing Location Update Cost in a PCS Network," *IEEE Transactions on Networking*, vol. 5, no. 1, pp. 25–33, Feb. 1997.

[31] Y. Lu, F. Kuipers, M. Janic, and P. V. Mieghem, "E2E Blocking Probability of IPTV and P2PTV," *Lecture Notes in Computer Science*, vol. 4982, pp. 445–456, 2008.

[32] J. Karvo, J. Virtamo, S. Aalto, and O. Martikainen, "Blocking of Dynamic Multicast Connections in a Single Link," *Proc. International Broadband Communications*, pp. 473–483, Apr. 1998.

[33] X. Zhang, J. G. Castellanos, and A. T. Campbell, "P-MIP: Paging Extensions for Mobile IP," *ACM Mobile Networks and Applications*, vol. 7, no. 2, pp. 127–141, Mar. 2002.

[34] A. Acharya, A. Bakre, and B. R. Badrinath. "IP Multicast Extentions for Mobile Internetworking," *Proc. INFOCOM*, pp. 67–74, March 1996.

[35] J. Jiang, V. Sekar, and H. Zhang, "Improving Fairness, Efficiency, and Stability in HTTP-based Adaptive Video Streaming with FESTIVE," *Proc. ACM CoNEXT*, pp. 97–108, Dec. 2012.

[36] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A Scheduling Framework for Adaptive Video Delivery over Cellular Networks," *Proc. ACM Mobicom*, pp. 389–400, Sep. 2013.

[37] D. D. Vleeschauwer, H. Viswanathan, A. Beck, S. Benno, G. Li, and R. Miller, "Optimization of HTTP Adaptive Streaming over Mobile Cellular Networks," *Proc. IEEE INFOCOM*, pp. 898–997, Apr. 2013.

[38] G. Tian and Y. Liu, "Towards Agile and Smooth Video Adaptation in Dynamic HTTP Streaming," *Proc. ACM CoNEXT*, pp. 109–120, Dec. 2012.

[39] C. Liu, I. Bouazizi, and M. Gabbouj, "Rate Adaptation for Adaptive HTTP Streaming," *Proc. ACM MMSys*, pp. 169–174, Feb. 2011.

[40] R. K. P. Mok, X. Luo, E. W. W. Chan, and R. K. C. Chang, "QDASH: A QoE-aware DASH System," *Proc. ACM MMSys*, pp. 11–22, Feb. 2012.

[41] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, and D. Oran, "Probe and Adapt: Rate Adaptation for HTTP Video Streaming at Scale," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 4, pp. 719–733, Apr. 2014.

[42] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An Experimental Evaluation of Rate-adaptation Algorithms in Adaptive Streaming over HTTP," *Proc. ACM MMSys*, pp. 157–168, Feb. 2011.

[43] S. Akhshabi, L. Anantakrishnan, A. C. Begen, and C. Dovrolis, "What Happens When HTTP Adaptive Streaming Players Compete for Bandwidth?," *Proc. ACM NOSSDAV*, pp. 9–14, Jun. 2012.

[44] C. Müller, S. Lederer, and C. Timmerer, "An Evaluation of Dynamic Adaptive Streaming over HTTP in Vehicular Environments," *Proc. ACM Workshop on Mobile Video*, pp. 37–42, Feb. 2012.

[45] T. C. Thang, H. T. Le, A. T. Pham, and Y. M. Ro, "An Evaluation of Bitrate Adaptation Methods for HTTP Live Streaming," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 4, pp. 693–705, Apr. 2014.

[46] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "CellSlice: Cellular Wireless Resource Slicing for Active RAN Sharing," *Proc. IEEE COMSNETS*, pp. 1–10, Jan. 2013.

[47] A. Gember, A. Anand, and A. Akella, "A Comparative Study of Handheld and Non-handheld Traffic in Campus Wi-Fi Networks," *Proc. PAM*, pp. 173–183, Mar. 2011.

[48] M. Zhang, M. Dusi, W. John, and C. Chen, "Analysis of UDP Traffic Usage on Internet Backbone Links," *Proc. IEEE SAINT*, pp. 280–281, Jul. 2009.

[49] J. S. Arora, M. W. Huang, and C. C. Hsieh, "Methods for Optimization of Nonlinear Problems with Discrete Variables: A Review," *Structural Optimization*, vol. 8, no. 2-3, pp. 69–85, Oct. 1994.

[50] N. Cranley, P. Perry, L. Murphy, "User Perception of Adapting Video Quality," *International Journal of Human-Computer Studies*, vol. 64, no. 8, pp. 637–647, Aug. 2006.

[51] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "A Quest for An Internet Video Quality-of-experience Metric," *Proc. ACM Workshop on Hot Topics in Networks*, pp. 97–102, Oct. 2012.

[52] "NS-3," http://www.nsnam.org/.

[53] G. Monghal, K. I. Pedersen, I. Z. Kovacs, P. E. Mogensen, "QoS Oriented Time and Frequency Domain Packet Schedulers for the UTRAN Long Term Evolution," *Proc. IEEE VTC*, pp. 2532–2536, May. 2008.

[54] "Accelerate Development of LTE Evolved Packet Core Product with Aricent Solutions," http://www.aricent.com/pdf/Aricent_Solution_Brief_LTE_EPCṗdf.

[55] "Mpeg-dash / Media Source Demo," http://dash-mse-test.appspot.com/.

[56] "JL-620 LTE Enterprise Indoor Small Cell," http://www.juniglobal.com/products/products_lte_jl620_d.asp.

[57] "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Transmission and Reception (Release 11)," *3GPP TS 36.101*, v11.10.0, http://www.3gpp.org/, Sep. 2014.

[58] "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures (Release 11)," *3GPP TS 36.213*, v11.8.0, http://www.3gpp.org/, Sep. 2014.

[59] "libnetfilter_queue," http://www.netfilter.org/projects/libnetfilter/_queue/.

[60] "Iperf," http://sourceforge.net/projects/iperf/.

[61] IEEE Std. 802.16e-2005 and IEEE Std. 802.16-2004/Cor 1-2005, http://www.ieee.org/, 2005.

초 록

LTE는 향상된 멀티미디어 브로드캐스트 및 멀티캐스트 서비스 (eMBMS)를 지원한다. 그러나 지연시간에 민감한 실시간 비디오 스트리밍 서비스는 무선 자원의 효율적인 사용과 함께 낮은 수준의 핸드오버 지연시간을 필요로 하는 어려움이 있다. 3GPP 표준에서는 단일 주파수 네트워크로 구성된 멀티미디어 멀티캐스트 서비스 (MBSFN) 지역을 도입하였는데, 이 지역에서는 복수의 기지국들이 동일한 멀티캐스트 패킷을 전송하게 된다. 이를 통하여 이 지역들 안에서는 핸드오버 지연시간을 줄여줄 수 있으나, LTE 망 내에 트래픽 부하를 올리는 결과를 낳는다.

본 학위 논문에서는 먼저 위치 관리 지역 (LMA)을 기반으로 한 MBSFN 구조를 제안하였다. 본 기법은 무선 대역폭의 큰 낭비없이 MBSFN 지역의 크기를 키워 평균 핸드오버 지연시간을 낮출 수 있다. 분석 모델로는 MBSFN과 LMA 크기, 사용자의 이동성, 분포, 시청하는 컨텐츠의 인기도에 따른 서비스의 지연 시간, 대역폭의 사용량, 서비스 차단 확률, 세가지를 제시하였다. 수학적 결과와 시뮬레이션을 통하여 제안한 LMA 기반 MBSFN 방법은 서비스 지연 시간을 적절히 유지하면서 대역폭을 효율적으로 사용할 수 있는 비디오 멀티캐스팅 방법임을 보였다.

다음으로, eMBMS 패킷을 확률적인 계산을 통하여 필요한 지역에 미리 전송하여 개별 사용자가 느끼는 평균 핸드오버 지연시간을 통계적으로 보장하고자 하는 방법을 제안하였다. 사전 전송에 따른 대역폭 소비와 핸드오버 지연시간의 감소를 분석하기 위한 수학적 모델을 제시하였고, 사용자의 이동성, 분포, 시청하는 비디오의 인기도를 고려하였다. 또한 시뮬레이션을 통하여 모델을 검증하였다.

한편, LTE망에서 HTTP 기반 스트리밍 (HAS)은 VoD와 같은 비 실시간 비디오 전송의 가장 주요한 기술이 될 것으로 기대된다. 본 학위 논문에서는 기존의 HAS

기술의 문제점을 분석하고, 그 통찰을 바탕으로 공정하고, 효율적이며, 안정적인 비디오 스트리밍이 가능한 새로운 네트워크 HAS 기법을 제시하였다. 주요한 특징으로는 비디오와 일반 데이터 사용자를 하나의 프레임워크 안에서 고려하였으며, HTTP Response를 통해 최적의 비디오 속도를 전달하는 방식으로 비디오 전송 속도의 직접적인 컨트롤이 가능하고, 전송 속도의 안정성이 높은 장점이 있다. 실제 LTE 펨토셀 네트워크에서의 실험을 통해 제안한 기법과 기존의 기법들을 비교하였으며, 제안한 기법이 비디오의 평균 전송속도를 향상시키고 질적인 안정성을 보장하며 비디오 사용자와 데이터 사용자 간의 균형을 맞출 수 있음을 보였다.

주요어: 망 계획, 비디오, 스트리밍, eMBMS, MBSFN, HAS, LTE
학번: 2006-21263