Universidade do Minho
U.PORTO
Universidade de Aveiro
2015
Departamento de
Eletrónica, Telecomunicações e Informática

Sérgio Miguel Calafate de Figueiredo

Distribuição de Video para Grupos de Utilizadores em Redes Móveis Heterogéneas

Mobile Video Multicasting in Heterogeneous Networks

Universidade de Aveiro
2015
Departamento de
Eletrónica, Telecomunicações e Informática

Sérgio Miguel Calafate de Figueiredo

Distribuição de Video para Grupos de Utilizadores em Redes Móveis Heterogéneas

Mobile Video Multicasting in Heterogeneous Networks

Tese a ser apresentada às Universidades de Minho, Aveiro e Porto para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Eletrotécnica / Telecomunicações no âmbito do programa doutoral MAP-Tele, realizada sob a orientação científica do Doutor Rui Luís Andrade Aguiar, Professor Catedrático do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro, e do Doutor Seil Jeon, investigador pós-doc do Instituto de Telecomunicações.

Apoio financeiro da Fundação para a Ciência e a Tecnologia – FCT através da bolsa SFRH/BD/68507/2010.

o júri / the jury

presidente / president

Doutor Carlos Manuel Martins da Costa
Professor Catedratico, Universidade de Aveiro

vogais / examiners committee

Doutor Alexandre Júlio Teixeira dos Santos
Professor Associado com Agregação, Escola de Engenharia, Universidade do Minho

Doutor Rui Luís Andrade Aguiar
Professor Catedrático, Universidade de Aveiro (orientador)

Doutora Marilia Pascoal Curado
Professora Auxiliar, Faculdade de Ciências e Tecnologia, Universidade de Coimbra

Doutor Rodolfo Alexandre Duarte Oliveira
Professor Auxiliar, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Doutor Diogo Nuno Pereira Gomes
Professor Auxiliar, Universidade de Aveiro

agradecimentos / acknowledgments

Gostaria de agradecer em primeiro lugar ao meu Orientador, Prof. Dr. Rui Luís Aguiar, e Co-Orientador, Doutor Seil Jeon, pelo valioso supervisionamento, e por me darem a oportunidade de aprender com alguns dos melhores profissionais da área.

A todos os meus colegas do grupo Advanced Telecommunications and Networks Group (ATNOG), em especial ao Daniel Corujo, por me aconselhar sempre que precisei, e ao Carlos Guimarães, pelo contributo importante para o meu trabalho em IP multicast. Também aos meus colegas do Programa Doutoral MAP-Tele, e àqueles com quem eu tanto aprendi no Projecto MEDIEVAL, pelos momentos partilhados em diversas ocasiões.

Ao Instituto de Telecomunicações pelo seu essencial contributo enquanto instituição de acolhimento, e à Fundação para a Ciência e a Tecnologia pelo suporte financeiro.

Gostaria também de agradecer aos meus colegas da Associação de Defesa Pessoal de Portugal, por me proporcionarem um ambiente que me permitiu crescer pessoalmente durante o desenrolar deste Doutoramento.

Por último, um agradecimento muito especial aos meus pais e ao meu irmão pelo apoio incondicional.

palavras-chave

IP multicast, Video, Gestão de Mobilidade Distribuida, Encaminhamento Multicast, Handover Undependente do Acesso, Offloading, Testbeds.

resumo

A evolução verificada nas características dos dispositivos moveis (capacidade de armazenamento, resolução do ecrã, processador, etc.) durante os últimos anos levou a uma alteração significativa nos comportamentos dos utilizadores, sendo agora comum o consumo e produção de conteúdos multimédia envolvendo terminais móveis, em particular o tráfego vídeo. Consequentemente, as redes de operador móvel, embora tendo também sido alvo constante de evoluções arquitecturais e melhorias em vários parâmetros (tais como capacidade, ritmo de transmissão/recepção, entre outros), vêem-se cada vez mais frequentemente desafiadas por aspectos de desempenho associados à natureza do tráfego de vídeo, seja pela exigência de requisitos associados a esse serviço, quer pelo aumento do volume do mesmo nesse tipo de redes.

Esta Tese propôe alterações à arquitetura móvel para a disseminação de vídeo mais eficiente, definindo e desenvolvendo mecanismos aplicáveis à rede, ou ao utilizador móvel. Em particular, são focados cenários suportados por IP multicast em redes móveis heterogéneas, isto é, com ênfase na aplicação destes mecanismos sobre diferentes tecnologias de acesso. As alterações sugeridas aplicam-se a cenários de utilizador estático ou móvel, sendo este a fonte ou receptor do tráfego vídeo. Da mesma forma, são propostas soluções tendo em vista operadores com diferentes objectivos de disseminação de vídeo, ou cujas redes têm diferentes características. A metodologia utilizada combinou a avaliação experimental em testbeds físicas com a avaliação matemática em simulações de redes, e permitiu verificar o impacto sobre a optimização da recepção de vídeo em terminais móveis.

keywords

IP multicast, Video, Distributed Mobility Management, Multicast Routing, IEEE 802.21, Media Independent Handover, Offloading, Testbeds.

abstract

The evolutions verified in mobile devices capabilities (storage capacity, screen resolution, processor, etc.) over the last years led to a significant change in mobile user behavior, with the consumption and creation of multimedia content becoming more common, in particular video traffic. Consequently, mobile operator networks, despite being the target of architectural evolutions and improvements over several parameters (such as capacity, transmission and reception performance, amongst others), also increasingly become more frequently challenged by performance aspects associated to the nature of video traffic, whether by the demanding requirements associated to that service, or by its volume increase in such networks.

This Thesis proposes modifications to the mobile architecture towards a more efficient video broadcasting, defining and developing mechanisms applicable to the network, or to the mobile terminal. Particularly, heterogeneous networks multicast IP mobility supported scenarios are focused, emphasizing their application over different access technologies. The suggested changes are applicable to mobile or static user scenarios, whether it performs the role of receiver or source of the video traffic. Similarly, the defined mechanisms propose solutions targeting operators with different video broadcasting goals, or whose networks have different characteristics. The pursued methodology combined an experimental evaluation executed over physical testbeds, with the mathematical evaluation using network simulation, allowing the verification of its impact on the optimization of video reception in mobile terminals.

“It is better to have enough ideas for some of them to be wrong,
than to be always right by having no ideas at all.”
— Edward De Bono

Aos meus pais, João e Eugénia. Ao meu irmão João Paulo. E a toda a minha família e amigos, que estiverem sempre presentes.

Contents

List of Contributions

Journal Articles

- Sérgio Figueiredo, Carlos Guimarães, Daniel Corujo, Rui L. Aguiar, “MI3M: A framework for media independent multicast mobility management”, Elsevier’s Computer Networks, September 2014

- Sérgio Figueiredo, Seil Jeon, Diogo Gomes, R. L. Aguiar, “D3M: Multicast listener mobility support mechanisms over distributed mobility anchoring architectures”, Elsevier’s Computer Network and Computer Applications, March 2015

Magazines

- Daniel Corujo, Sérgio Figueiredo, Rui L. Aguiar, "Sinalização Multiponto Independente do Meio para Melhoria do desempenho de Vídeo", Redes, Telecom e Instalações, Editora Aranda, August 2012

Conference Proceedings

- Daniel Corujo, Sérgio Figueiredo, Rui L. Aguiar, "Media-Independent Multicast Signalling for Enhanced Video Performance in the MEDIEVAL Project", Proc. 2011 Future Network & Mobile Summit, Warsaw, Poland, June 2011

- Fabio Giust, Carlos J. Bernardos, Sérgio Figueiredo, Pedro Neves, Telemaco Melia, "A Hybrid MIPv6 and PMIPv6 Distributed Mobility Management: the MEDIEVAL approach ", Proc. MediaWin 2011 IEEE Workshop on multiMedia Applications over Wireless Networks, Corfu, Greece, June 2011

- Sérgio Figueiredo, Michelle Wetterwald, Tien-Thinh Nguyen, Lucas Eznarriaga, Noam Amram, Rui L. Aguiar, "SVC Multicast Video Mobility Support in MEDIEVAL Project", Proc. 2012 Future Network & Mobile Summit, Berlin, Germany, July 2012

- Sérgio Figueiredo, Seil Jeon, Rui L. Aguiar, "Use-cases Analysis for Multicast Listener Support in Network-based Distributed Mobility Management", Proc. 23rd IEEE PIMRC, Sydney, Australia, September 2012

- Seil Jeon, Sérgio Figueiredo, Rui L. Aguiar, "A Channel-Manageable IP Multicast Support Framework for Distributed Mobility Management", Proc. IFIP Wireless Days 2012, Dublin, Ireland, November 2012

- Sérgio Figueiredo, Carlos Guimarães, Rui L. Aguiar, Tien-Thinh Nguyen, Liron Yadin, Nuno Carapeto, Carlos Parada, "Broadcasting User Content over Novel Mobile Networks", Proc. 2013 International Conference on Communications, Budapest, June 2013

- Seil Jeon, Sérgio Figueiredo, Rui L. Aguiar, "On the Impacts of Distributed and Dynamic Mobility Management Strategy - A Simulation Study", Proc. IFIP Wireless Days 2013, Valencia, Spain, November 2013

- Sérgio Figueiredo, Seil Jeon, Rui L. Aguiar, "Empowering IP multicast for multimedia delivery over heterogeneous mobile wireless networks," Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on, May 2014

Standardization Efforts

- Sérgio Figueiredo, Seil Jeon, and Rui L. Aguiar, “IP Multicast Use Case and Analysis over Distributed Mobility Management”, draft-sfigueiredo-multimob-use-case-dmm-03.txt, IETF Internet draft, October 2012. Presented in the 85th IETF (Atlanta, November 2012)

- H. Anthony Chan, *et al*, “Requirements for Distributed Mobility Management”, IETF RFC7333, August 2014

- Seil Jeon, Sérgio Figueiredo, and Younghan Kim, “Use Cases and API Extension for Source IP address selection draft”,-sijeon-dmm-use-cases-api-source-00.txt, IETF Internet draft, February 2015

Demonstrations

- “Personal Broadcasting Scenario”, Future Networks and Mobile Summit (FUNEMS), Lisbon, July 2013

- “Personal Broadcasting and Mobility Scenario”, MEDIEVAL Final Audit, Sophia-Antipolis, October 2013

Software Suites

- "Multicast Mobility Engine (MUME)": implemented in C++ on top of MRD6.

Nomenclature

3GPP	Third Generation Partnership Project
ABR	Adaptive Bitrate Streaming
AHS	Adaptive HTTP Streaming
AMazING	Advanced Mobile wIreless Network playGround
AN	Access Network
ANDSF	Access Network Discovery and Selection Function
APN	Access Point Names
AR	Access Router
ASM	Any-Source Multicast
ATM	Asynchronous Transfer Mode
AVC	Advanced Video Coding
BCE	Binding Cache Entry
BM-SC	Broadcast/Multicast Service Center
BSR	Bootstrapping Router
BU	Binding Update
BUL	Binding Update List
CAPEX	Capital Expenditure
CCN	Content Centric Networking
CDMA	Code Division Multiple Access

CDN	Content Delivery Network
CLO	Cross Layer Optimizer
CM	Connection Manager
CN	Correspondent Node
CoA	Care-of-Address
CSIPTO	Coordinated SIPTO: Change of PGW
CU	Coding Unit
DASH	Dynamic Adaptive Streaming over HTTP
DHCP	Dynamic Host Configuration Protocol
DMM	Distributed Mobility Management
DMRS	Direct Multicast Routing Scheme
DMS	Directed Multicast Service
DR	Designated Router
DVB-H	Digital Video Broadcasting for Handheld
EAP-AKA	Extensible Authentication Protocol - Authentication and Key Agreement
EAP-SIM	Extensible Authentication Protocol - Subscriber Identity Module
eNB	eNodeB
EPS	Evolved Packet System
FA	Foreign Agent
FMIP	Fast Mobile IP
GCR	Groupcast with Retries
GPRS	General Packet Radio Service
GTP	Generic Packet Routing System Tunneling Protocol
HA	Home Agent

HD	High Definition
HeNB	Home eNodeB
HEVC	High Efficiency Video Coding
HI	Handover Indication
HLS	HTTP Live Streaming
HMIPv6	Hierarchical Mobile IPv6
HNP	Home Network Prefix
HO	Handover
HoA	Home Address
HSPA	High Speed Packet Access
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
ICT	Information and Communication Technologies
IDR	Instantaneous Decoding Refresh
IEEE	Institute of Electrical and Electronics Engineers
IFOM	IP Flow Mobility
IGMP	Internet Group Membership Protocol
IIF	Incoming Interface
IMA	Inverse Multiplexing over Asynchronous Transfer Mode
IP	Internet Protocol
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission
ISP	Internet Service Provider
JVT	Joint Video Team
L2	Layer 2

L3	Layer 3
LCU	Largest Coding Unit
LGW	Local Gateway
LI	Logical Interface
LIMONET	LIPA mobility and SIPTO at the Local network
LIPA	Local IP Access
LMA	Local Mobility Anchor
LMD	Local Mobility Domain
LTE	Long Term Evolution
MAC	Medium Access Control
MAG	Mobilie Access Gateway
MAP	Mobility Anchor Point
MAPCON	Multi-Access PDN Connectivity
MAR	Mobility Access Router
MBMS	Multimedia Broadcast/Multicast Service
MBMS-GW	MBMS Gateway
MBS	MLD Base Solution
MBSFN	MBMS over Single Frequency Network
MCE	Multicast Coordination Entity
MEDIEVAL	MultiMEDia transport for mobIlE Video AppLications
MFIB	Multicast Forwarding Information Base
MFR	Multilink Frame Relay
MICS	Media Independent Control Service
MIES	Media Independent Event Service

MIF	Multiple Interfaces
MIH	Media Independent Handover
MIIS	Media Independent Information Service
MIP	Mobile IP
MIPv6	Mobile IPv6
ML-PPP	Multilink Point to Point Protocol
MLD	Multicast Listener Discovery
MME	Mobility Management Entity
MMT	MPEG Media Transport
MN	Mobile Node
MNG	Mobile Node Group
MPEG	Moving Picture Experts Group
MR	Multicast Router
MRD6	Multicast Routing Daemon for IPv6
MRIB	Multicast Routing Information Base
MUME	MUlticast Mobility Engine
NAL	Network Abstraction Layer
NDP	Network Decision Point
NGWN	Next Generation Wireless Networks
OPEX	Operational Expenditure
OPMIP	Open Proxy Mobile IP
OSI	Open Systems Interconnection
OTT	Over-the-top
P2P	Peer-to-peer

PBS	Personal Broadcasting Service
PCoA	Proxy Care-of-Address
PDN	Packet Data Network
PF	Proactive and Fully distributed scheme
PGW	PDN Gateway
PIM	Protocol Independent Multicast
PIM-SM	Protocol Independent Multicast Sparse Mode
PIM-SSM	Protocol Independent Multicast Source Specific Multicast
PMIPv6	Proxy Mobile IPv6
PoA	Point of Attachment
PP	Proactive and Partially distributed scheme
PSNR	Peak Signal-to-Noise Ratio
QoE	Quality of Experience
QoS	Quality of Service
RAM	Random Access Memory
RF	Reactive and Fully distributed scheme
RNC	Radio Network Controller
RP	Rendezvous Point / Reactive and Partially distributed scheme
RPF	Reverse Path Forwarding
RPT	Rendezvous Point Tree
RTCP	RTP Control Protocol
RTMP	Real Time Messaging Protocol
RTP	Real-time Transport Protocol
RTSP	Real Time Streaming Protocol

SA	System Architecture
SaMOG	S2a-based Mobility Over GTP
SAP	Service Access Point
SCS	Stream Classification Service
SEI	Supplemental Enhancement Information
SFN	Single Frequency Network
SGSN	Serving GPRS Support Node
SGW	Serving Gateway
SIAL	Subscription Information through the LMA
SIPTO	Selected IP Traffic Offload
SPT	Shortest Path Tree
SSH	Secure Shell
SSM	Source-Specific Multicast
SVC	Scalable Video Coding
TBIM	Tuning the Behavior of the IGMP and MLD protocols in mobile environments
TLV	Type-length value
TS	Transport Stream
TSG-SA	Technical Specification Group for Service and Architecture
TWAG	Trusted Wireless Access Gateway
UDP	User Datagram Protocol
UE	User Equipment
UGC	User Generated Content
UMTS	Universal Mobile Terrestrial System
UPCON	User Plane Congestion Management

UR	Unsolicited Retry
UTRAN	Universal Terrestrial Radio Access Network
VCEG	Video Coding Experts Group
VCL	Video Coding Layer
WLAN	Wireless Local Area Network

List of Figures

List of Tables

Chapter 1

Introduction

We begin this chapter by discussing the motivation and goals of this Thesis, in order to acquaint the readers with the general emerging challenges relating to video support in mobile networks, as well as with those specific to Internet Protocol (IP) multicast and broadcasting scenarios, necessary for a clear understanding of this Thesis. Then, the main contributions are introduced, as well as the implemented methodology, including computational infrastructures or testbeds. Finally, we present the Thesis outline, summarizing the contents of upcoming chapters.

1.1 Motivation

Internet is globally accepted as one of mankind's most significant technological achievements. It is a powerful enabler whose impact has revolutionized human society both by relieving the boredom from previously time-consuming tasks, and by providing us with a plethora of communication options, all in presential and - more disruptively - in non-presential scenarios.

Internet has evolved from a few house-sized machines devised for the transfer of short text files to a worldwide communications powerhouse allowing fast dissemination of heavy-sized information, e.g. High Definition (HD) video to a single or multiple destinations. Watching videos in mobile devices has become highly popular, being the most consumed traffic among all mobile traffic; furthermore, it is predicted to have a 10-fold growth in smartphone traffic from 2014 to 2020, where it will account for 55% of all mobile data traffic [1]. Such a pattern of data consumption has given a challenging homework to Internet Service Providers (ISPs), which are required to enhance their networks to meet the increasing demand of mobile video consumption while simultaneously reducing Capital Expenditure (CAPEX) and Operational Expenditure (OPEX), for maximizing their profit.

[1]Ericsson Mobility Report, February 2015

This “boom” in the consumption of mobile video is explained by multiple reasons, from technical to business-related perspectives. We now dwell in these reasons.

Powerful and cheaper mobile devices are now widely available: These are now capable of processing video, pushed by overall hardware improvements, reduction in manufacturing costs, etc. Complementing technological advances, “bigger became better” regarding mobile screens and their resolutions. Supported by larger batteries, the ad hoc consumption of video has turned into a more straightforward, attractive and convenient process; reflected in its daily consumption during queuing or spare times, to devour short entertainment or informative videos.

User viewing habits have significantly changed, with ubiquitous access, high-speed connections, and quickly accessible and reproducible content being requirements which have shuffled mobile networks. Furthermore, the advent of social video portals the likes of YouTube and Vimeo has now introduced new scenarios beyond the typical “client-server model”, where users share their either previously recorded or live transmitted personal videos - marking the User Generated Content (UGC) era. The cyclic nature between technology availability and the users’s ever demanding habits is expressed in Figure 1.1.

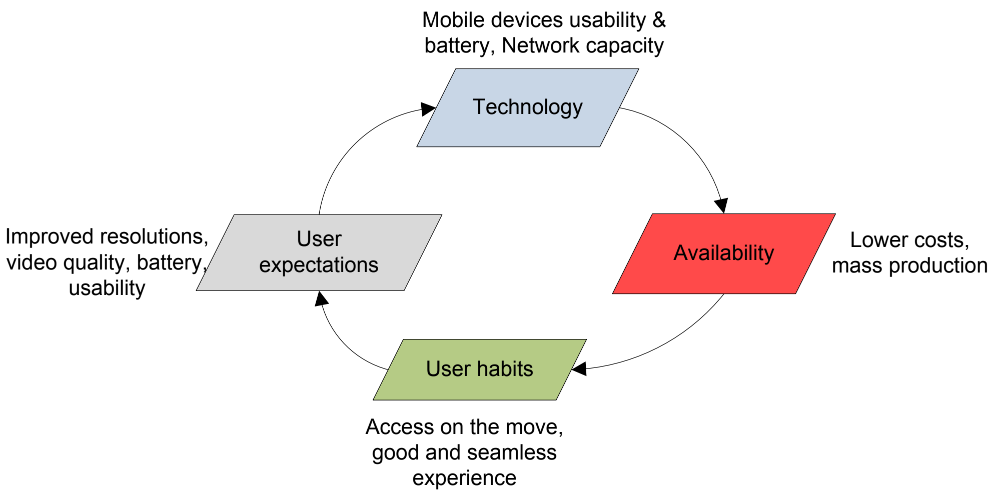


Figure 1.1: Technological cycle

Network performance improvements: The role undertaken by operators themselves has been crucial for paving the way to the mobile video era. The need to cater to increasing requirements of mobile data forced operators to modernize networks from access to core, an effort more evident with the deployment of High Speed Packet Access (HSPA), HSPA+ and Long Term Evolution (LTE) - as well as the ongoing effort on LTE Advanced. The increased capacity, improved mechanisms for delivering Quality of Service for services which need it, and the larger coverage, all benefited the realization of mobile multimedia access, with improved performance. Moreover, recently specified video compression standards are more efficient, and video transport protocols are more flexible and smarter: for instance, the possibility to transport video using Hypertext Transfer Protocol (HTTP) enables the virtual delivery to any IP connected device, including safety-concerned cellular networks.

1.2 Research Challenges

The continuous technological advances in mobile devices, content production and distribution, such as the increase in screen and produced video resolution (e.g. 4K), will continue to lead to higher per-device data consumption patterns. Coupled with an even greater number of connected and multimedia-capable devices per person (e.g. smart glasses, other connected sensors), the so-called “video explosion” mentioned above may still be beginning. Given the observed desire for multimedia content, novel and more demanding video services will surely emerge, representing one huge opportunity for operators, in particular considering mobile scenarios.

Motivated by the increasing relevance of UGC content and associated demanding scenarios, this Thesis focuses on live video delivery scenarios, where both the transmitter and the multiple service subscribers may be roaming between different geographic micro or macro locations. Some of these scenarios translate into previously unaddressed technical challenges or into mechanisms prone to optimizations. Moreover, while some partial solutions are already available for professional or semi-professional use, the contributions within this Thesis aim to facilitate the referred scenarios, as well as assure improved service throughout all session duration.

In the considered scenarios, the mobility of content transmitter adds difficulty to the process of assuring QoE for all associated receivers, while the mobility of receivers themselves represents distinct challenges depending on the initial and target access technologies properties and status, the service requirements, and others. In particular, the contributions developed within this Thesis fit into the following research challenges:

- **Delivery of IP multicast under heterogeneous scenarios**

Associated issues include the design of a solution specifically for a particular access technology, or service disruption in case of inter-technology handover (HO).

- **Symbiotic operation of multicast and mobility management mechanisms**

IP multicast mobility is mostly handled as an afterthought, adding unnecessary complexity to existing solutions, and preventing the straightforward resolution of some of the associated issues (e.g. tunnel convergence problem).

- **Converged source and receiver multicast mobility management**

Current solutions either tackle one of the multicast roles (source or receiver), leading to severe inefficiency in scenarios where both entities have mobile capabilities.

- **Video-aware multicast support and mobility**

The focus on QoS performance does not assure satisfactory user experience. Thus, the mutual awareness between mobility services and applications is required, enabling the application to perform necessary adaptations, and mobility decisions to be sensitive to application needs.

1.3 Thesis Objectives

This Thesis pursues the development of generic mechanisms for efficient provisioning of IP multicast video traffic scenarios under mobility. It covers several underlying technologies and concepts, such as video transport technologies, mobility management and HO modes, cross-layer optimizations, etc., all complying with the end-to-end Transport Control Protocol (TCP) / IP model. Concretely, the objectives addressed under the umbrella of this Thesis may be organized as follows:

- **Objective 1: Multicast mobility for both user roles**

Embed architectures with multicast mobility management solutions enabling the mobility scenarios of both multicast source or receivers under different types of HO - e.g. intra or inter-technology, smooth HOs.

- **Objective 2: Multicast mobility in Distributed Mobility Management (DMM)**

Design multicast mobility management solutions leveraging distributed and dynamic mobility paradigm, where traffic is not required to converge in single centralized transport hubs. Different standardized multicast mechanisms such as Protocol Independent-Multicast Sparse Mode (PIM-SM) or Multicast Listener Discovery (MLD) Proxies should be explored aiming at this objective.

- **Objective 3: Video-aware IP multicast delivery**

Design event-based video flow transport for IP multicast delivery. The possibility for dealing with variable events affecting video sessions (e.g. backhaul / access network congestion, user mobility, etc.) demands smart solutions taking into account multiple contexts, such as properties from applications, transport and mobility elements. This may involve the definition of transparent and generic cross-layer techniques sensitive to the requirements of video delivery over wireless access.

- **Objective 4: Adaptable multicast Mobility Management**

Design mobile architectures capable of multiple operation modes, whose selection takes the varying requirements of video services into account, both in terms of QoS and QoE parameters. Mobile video delivery is associated with significant heterogeneity, such as in terms of traffic flows requirements (e.g. Video on Demand (VoD) vs live video), devices or link channel conditions, which impose different constraints and have direct impact in QoE. This objective leverages Objective 2, i.e., the specification of isolated multicast mobility management solutions.

1.4 Contribution

The work developed throughout the duration of this Thesis addresses two of the main trends of Information and Communication Technologies (ICT): mobility and video communications. Briefly, it focuses on the global challenge of supporting video in mobile scenarios. The scenarios of interest are those where the underlying extreme conditions may be originated by variables such as the video quality - and proportional data rates demand - and the number of users - i.e. network population density -, or events like network congestion. While such challenges are already identifiable in present mobile video delivery scenarios, they are prone to repeat more frequently in the future, and with even stronger impact.

Reflecting the importance of efficient video delivery options, this Thesis places a strong emphasis in the support of mobility scenarios in group-based services, both for mobile producers or consumers. To achieve this, several optimizations to be embedded by mobile operators within their networks are described, aimed at maximizing the quality and efficiency of mobile video delivery, as well as the degree of robustness against events which are hazardous to - and may even disrupt - the communication or service. Such contribution consists of innovative architectures for the enhancement of mobile multicast video support on top of already standardized mobility management solutions [1] and the thorough research on future mobility management solutions overcoming scalability and routing limitations [2].

The symbiosis of the two topics resulted in the embryonic study of IP multicast support over Distributed Mobility Management paradigm, leading to the research of potential use cases and resulting limitations [3][4], as well as initial efforts to overcome these same limitations [5]. These works ultimately led to an important contribution to the recently created DMM Work Group (WG). Namely, it allowed to highlight of the importance of IP multicast for future mobile environments, and to point out which issues were to be avoided by future DMM solutions, materialized in RFC7333[6] through the inclusion of a new Requirement: (REQ8: "Multicast Considerations") and the modification of the Problem Statement section (by adding a new problem PS8 - "Duplicate Multicast Traffic", and adapting Problem Statement PS1 - "Non-optimal routes").

This paved the way to the development of more detailed solutions tackling the support of IP multicast under novel mobility management approaches, taking into account distinct requirement degrees [7][8], and suited to heterogeneous mobility scenarios [9].

Part of the studies on IP multicast mobility aimed at the preservation of the user-perceived quality of the video service throughout its session, by providing necessary mechanisms for improving the overall efficacy of HOs, such as steaming the adequate "network - host cooperation" [1][9]. The handling of events other than mobility, such as network congestion, or wireless conditions variation, is enabled by crossing the information o multiple layers, and realized through the definition of suitable architectures, leveraging on the extension of relevant protocols and the definition of new conceptual entities [10]. Original contributions that are further explored in this Thesis include the study and proposals on multicast source mobility support in DMM.

Finally, as a result of surveying the work on Distributed Mobility Management, the focus was placed in source IP address selection and configuration taking into account the requirement of the concept of "IP address type", where each type maps to different IP mobility management needs (e.g. reachability or session continuity). Through the identification of multiple use cases which differ according to the availability - or lack of - one or multiple IP address types, the necessary extensions to the socket API were proposed [11].

It is important to remark that a significant part of this work was integrated by the author into European Project MultiMEDia transport for mobIlE Video AppLications (MEDIEVAL [2]), that aimed to develop a full-fledged operator architecture tailored for improved video delivery. The participation in this Project enabled the consideration of additional aspects such as access layer enhancements -802.11aa, evolved Multicast/Broadcast Multimedia Service (eMBMS [10] -, or the facilitation of additional scenarios such as group mobility[12], whose results fueled IEEE 802.21d - though no direct contribution was realized. Such integration allowed the implementation, validation and performance evaluation of some of the solutions proposed within this Thesis, by means of prototyping and deployment in a testbed. Given the strong relationship between this Thesis and MEDIEVAL Project, at the beginning of each chapter the reader is informed about which contents are the result of the Project-wide work, and which were the result of the Thesis' author, apart from the Project.

1.5 Evaluation Tools

Most of the contributions delivered through this Thesis were subject to a system level performance evaluation through three main methods: 1) analytical approach / mathematical analysis, 2) computer simulations, and 3) experimental evaluation, supported by computa-

[2] http://www.ict-medieval.eu/

tional infrastructures / testbeds.

1.5.1 Mathematical Analysis

Mathematical analysis is necessary for the system modeling validation of simple concepts and technology improvements. Given the complex and variable characteristics of wireless media, it is only an option for the evaluation of network mechanisms where wireless is not an influencing factor, and where it is acceptable to build on assumptions and approximations of its behavior.

1.5.2 Computer Simulation

When the evaluation of an envisioned metric depends on multiple parameters whose behavior can't be forecasted, and when it is not possible to pursuit the evaluation through physical experimentation – e.g. due to unavailability -, it is both more practical and efficient to employ a computer simulation. Using this approach, care must be taken to apply realistic network characteristics. For instance, considering delivery latency, it is crucial to employ values for the specific considered media, processing and routing latencies, transmission latencies, among others. Moreover, considering discrete event simulation it is important that at each event the state of the system is fully updated. This is particularly relevant when designing customized simulations, which was the case in this Thesis. Specifically, MATLAB was used to design simple multicast mobility scenarios.

1.5.3 Infrastructures

One of the key methods allowing the validation of concepts, interfaces and ultimately the designed architectures operability, is experimental validation. Thus, besides the mathematical and simulation works, a significant part of the work was evaluated in two physical testbeds: a local one, the Advanced Mobile wireless playGrouNd (AMazING) testbed and a remote one, EURECOM's testbed.

1.5.3.1 AMazING testbed

AMazING was used for the validation of the multicast mobility mechanisms in centralized mobility management protocols, emulating user mobility by managing the available interfaces. AMazING is located at the rooftop of the Instituto de Telecomunicações building in Aveiro, Portugal[3]. The testbed was initially deployed for supporting research on next generation wireless networks (NGWN), and is characterized by two main advantages from

[3] http://amazing.atnog.av.it.pt

the experimenter perspective: increased controllability and high reproducibility of the experiments.

The testbed, shown in Figure 1.2, is composed by 25 nodes. Each node has a VIA Eden 1GHz processor with 1GB RAM, a wired Gigabit interface and two wireless interfaces: a 802.11a/b/g/n Atheros 9K and a 802.11a/b/g Atheros 5K. All nodes run the Linux OS (Debian distribution) with kernel version 3.6.7-686-pae. The testbed is described in more detail in [13].

This testbed was intensively used for the initial experiments with mobility management and multicast protocols, for trial and empiric observation of associated problems, as well as validation of designed architectures. Objectively, the experimental results from [1] were fully obtained in this testbed, as well as part of the concepts explored in [8].

Figure 1.2: AMazING testbed

1.5.3.2 EURECOM's testbed

In order to complement this experimental work, and integrated with MEDIEVAL's architecture evaluation, EURECOM's premises were used for evaluating solutions towards distributed mobility management, enriched by the diversity of available hardware and interface technologies (e.g. commercial LTE solution, 3G). The testbed is composed of seventeen physical machines (a subset of which can be depicted in Figure 1.3). Two machines were used for the management and the sites interconnection. The other fifteen machines are the core of the testbed. They all run Ubuntu Long Term Support distributions, either distribution 10.04 or 12.04. Among the fifteen machines, six are Mobile Nodes (MNs), five are Mobile

Access Routers (MARs), one is the central router and three are servers.

Different versions of Linux kernel are running on testbed machines. The machines acting like servers are running the generic kernel 2.6.32-21 coming by default with the Ubuntu distribution installation. The machines that carry the LTE technology require a specific real-time kernel.

The testbed is supported by two network topologies, one in IPv4 and another in IPv6. IPv4 addressing is used for the management and development, whereas IPv6 addressing is used for testing and operational purposes. The most powerful machines have been reserved to run the emulated LTE technology and laptops have been preferably used as Mobile Nodes. The remote access for the development phase was realized using IPv4, with the Secure Shell (SSH) server having two network cards. A static IPv4 public address has been assigned on the first one in order to connect it directly on Internet. The second one is connected to MEDIEVAL private network testbed. The SSH server is thus acting as an IPv4 gateway between the MEDIEVAL private testbed and Internet. All testbed machines are directly reachable from the SSH server in one hop. SSH'ing was an essential method for accessing, deploying and configuring developed modules. Besides, it enabled remote evaluation by repeated testing. IPv6 is used for the operational phase. The IPv6 core of the testbed is achieved by a Core Router, which acts as a IPv6 router and interconnects the different IPv6 networks of the testbed.

EURECOM's testbed was used for evaluating the performance of later mechanisms in the scope of multicast context transfer over DMM scenarios, namely [7] and [9].

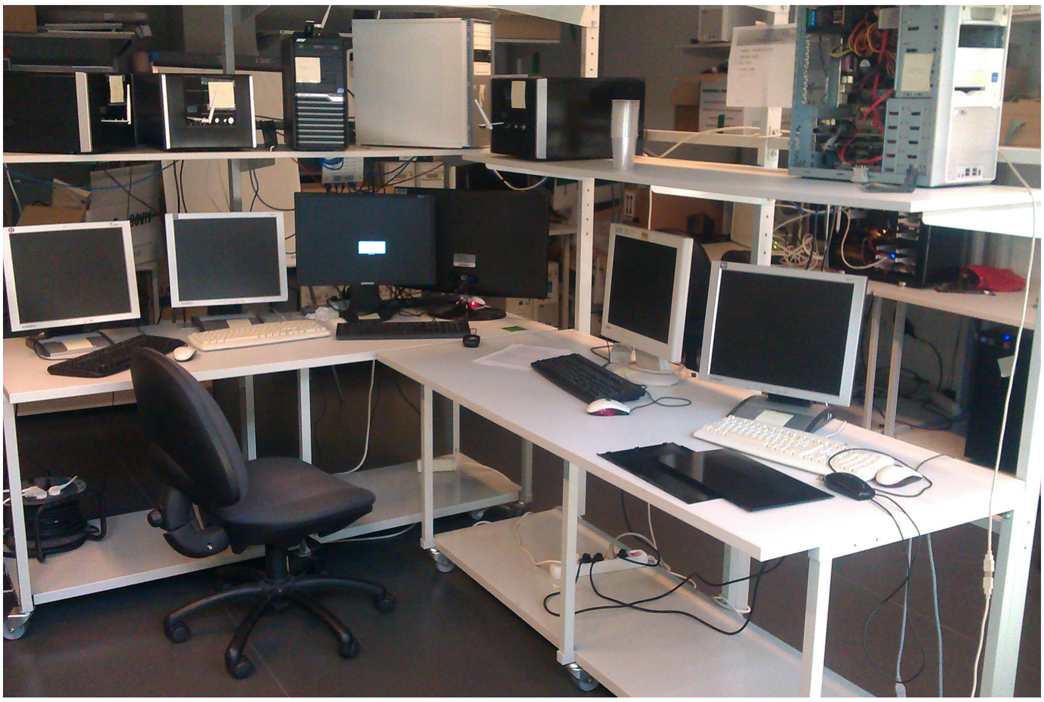


Figure 1.3: EURECOM's testbed

1.6 Thesis Outline

The remainder of this chapter presents the contribution of the Thesis and the computational resources which were used for the experimental results.

- **Chapter 2** introduces those concepts necessary for contextualizing the reader on the work presented throughout the Thesis. As such, the contents relate to three main topics: mobility management, IP multicast, and video. First, essential mobility concepts are shown, including the classification of different HO types, HO preparation and IP mobility management itself; the issues involved in the use of multiple interfaces are also addressed, as such issue directly relates to mobility within heterogeneous scenarios. Shifting to the cellular industry field, leveraging 3GPP (Third Generation Partnership Project) technology, the background on Mobile Access convergence is then considered. This work includes previous efforts on the integration between Cellular and non-3GPP networks, an item essential for providing operator networks with added capacity and path flexibility - and consequently resource efficiency. IP mobility support in cellular environments is described, as well as intelligent access network selection and its importance. Justified by the focus on the support of group-based services under mobile conditions, the core mechanisms for supporting these services are then re-visited; and the background on IP multicast mobility is presented, spanning both faced problems and available solutions. This section would not be complete without addressing video in mobile environments; and the clear listing of the involved challenges. Thus, a section is dedicated to present video playback and transport mechanisms. Finally, cross-layer design is presented as a key direction to enable improved video support in mobile environments.

- **Chapter 3** is focused on the enhancement of multicast mobility management over network-based solutions (PMIPv6), encompassing both mobile multicast receivers and sources. Based on the previously identified challenges, a set of goals to be achieved by potential solutions under the considered scenario is defined. Then, an architecture fulfilling these goals is specified, and clearly detailed. Finally, this architecture is extensively evaluated over an experimental testbed for the two relevant scenarios: receiver and source mobility.

- **Chapter 4** concerns the evolution from current mobility management from a centralized to a distributed paradigm. Thus, and given the multiple optimizations that may be achieved through DMM, a list of the main desired characteristics is defined. Motivated by the preliminary status on the definition of DMM protocols, existing design issues are identified and organized in different categories: data plane management, control plane management, and others. Finally, the performance of such DMM solution is compared against that of PMIPv6 over a specific network topology.

- **Chapter 5** studies the convergence of IP multicast with distributed mobility management, i.e., the support of IP multicast in mobility solutions applying DMM concept. As a first approach, multicast mobility leveraging on MLD Proxy is proposed and evaluated. As a result of the identified limitations when applying such method, the research focused on exploiting multicast routing deployment in mobility entities, as and alternative to MLD Proxies. A full-fledged architecture with these characteristics is thus presented, and extensively evaluated by means of mathematical and simple simulation scenarios.

- **Chapter 6** presents an operator architecture for mobile video delivery - the MEDIEVAL architecture, where a significant amount of this Thesis' work was applied. As part of this architecture, additional mechanisms for improving the delivery of mobile multicast video services are described. Namely, the following mechanisms and their integration in the system are described: H.264 SVC video compression standard, the enhancement of radio access technologies, and the signaling optimization of current HO procedures for group mobility. Moreover, two scenarios coupling the aforementioned mechanisms conclude the chapter.

- **Chapter 7** overviews the conclusions resulting from the developed work and associated achievements. It also expands beyond the reach of this Thesis, providing insights on upcoming challenges which may directly or indirectly benefit from the presented work.

Furthermore, the Thesis chapters are linked using inter-chapter information as follows. Each chapter is introduced by a short Foreword whose goal is two-fold: first, it overviews issues not yet addressed by previous chapter(s), similarly to a gap analysis. Secondly, it summarizes the motivation for the current chapter. Additionally, each chapter is finalized by means of Concluding Remarks, allowing the reader to review main achievements and conclusions.

Chapter 2

Preliminary concepts and Related Work

While previous chapter laid the supporting lines for the work developed within this Thesis, it is essential to provide the reader with the notions required for understanding the methodology followed, achievements and results. Thus, this chapter paves the way to the Thesis work by describing base concepts and relevant issues.

2.1 Introduction

The topic of this Thesis handles a diversity of subjects, as a consequence of crossing two previously independent "worlds": Mobile networks and Video delivery, which by themselves already represented interesting innovation opportunities. Moreover, a huge amount of research tackling video support under mobility is underway, including IP multicast-based video, from physical to application layer, and from the terminal to the network core. As presented, this Thesis focuses on scenarios which benefit from - or are only possible through - efficient network solutions.

This chapter presents relevant base concepts, technologies and protocols required to contextualize the developed work, which is shown in upcoming sections.

Chapter Contents

- **Section 2.2**: First, concepts related to IP connectivity are introduced, such as HO and mobility management. A part of this section is dedicated to IEEE 802.21, which employs a relevant position within this Thesis and is seen as a key intelligence provider in HO processes.
- **Section 2.3**: The second part of this section presents the background on cellular and Wireless Local Area Network (WLAN) convergence, necessary for understanding the

difficulties in providing mobility and seamless video support in heterogeneous environments.

- **Section 2.4**: The third part is dedicated to group-based services; first, base IP multicast protocols are covered, before presenting challenges arising from IP multicast support in legacy mobility management protocols, paving the way for its consideration in novel mobility management solutions. The solution adopted by 3GPP for group-based service delivery, MBMS, is also described.

- **Section 2.5**: The subsequent section goes through necessary video notions are described, spanning common delivery methods, the most widely used codecs and transport technologies.

- **Section 2.6**: This section presents challenges associated to mobile video delivery, from Internet design issues, to those specific to the video delivery ecosystem.

- **Section 2.7**: Finally, the last part of this chapter is dedicated to cross layer design and state of the art cross-layer solutions for supporting video in wireless and mobile scenarios.

2.2 Mobility management

In current scenarios, it is typical to have the simultaneous availability of multiple endpoints to which it is possible to establish a connection - access points -, either from a single or multiple radio technologies. Throughout a device's connectivity lifetime, it is necessary to switch between different points of attachment - HO -, due to factors such as user mobility or simply link degradation. As such, an HO will be a result of functions such as the performance evaluation of available networks, MN's position tracking or location management, and ultimately HO target selection taking into account additional context such as user profile. During an HO, the communication channel between the MN and a correspondent node (CN) may be interrupted, which, depending on the interruption duration and the session's rate, may result in packet loss. The goal of mobility mechanisms is to reduce this interruption time - the HO latency.

Mobility management consists of two services to support mobile communications and deliver the data packets during the MN's movements in the wireless networks: location management and HO management [14]. The former refers to database design and signaling required for tracking MN's position during change of access link, while the latter concerns mobility context (mobility trigger, target and execution details) required for preserving the MN's communication.

2.2.1 Characterizing handovers

Handovers can be characterized according to several strategies or contexts. Some of the most commonly used categories are the following:

- With regards to performance, mobility management solutions may produce a **fast HO** when minimizing HO latency or **smooth** HO when minimizing packet loss. Besides, the HO is referred to as **seamless** if the necessary mechanisms used by the mobility management solution make it imperceptible to the user in terms of service quality or capabilities degradation.

- HOs may also be characterized according to the involved technologies: **horizontal** or **intra-technology** HO refers to mobility within a single access technology, while **vertical** or **inter-technology** HO refers to mobility between different access technologies.

- Concerning the control and involved signaling, HOs may be **mobile / network-controlled** or **mobile / network-initiated**, respectively.

- HO may also be operated with distinct connectivity concerns. In **make-before break (soft)** HOs, the new link is established before the previous link is torn down, while in **break-before-make (hard)** HOs the target link is only established after loss of the previous link.

- Moreover, **macro** or **global** HOs involve a large area such as an administrative mobile domain, and **micro** or **local** HOs refer to mobility within a small region or single domain.

- **Layer 2** (L2) mobility refers to the change in Radio Access Network (RAN) link / access point, and does not lead to IP address change. **Layer 3** (L3) mobility on the other hand is associated to the change of IP address, and will be further detailed in section 2.2.2.

- Finally, if the previous network initiates the HO process, the mobility process is **predictive,** while mobility is **reactive** when the mobility signaling was initiated by the target network.

2.2.2 IP Mobility Management

The modification of IP address as a result of mobility in initial legacy services (web browsing, email and other text-based services) didn't have consequences besides content retransmission and consequent slightly longer transmission. With the emergence of voice- and video-based real time and interactive services, the renewal of IP address directly impacts user experience, either through complete connection loss or unbearable jitter, delay or packet loss. This paradigm change was the initial motivation for requiring means for IP address continuity.

IP mobility management is an important research issue for future IP wireless network, and regards the techniques that enable MNs to roam between network domains without any observable service change or disruption. Mobility management protocols can operate at different Open Systems Interconnection (OSI) layers such as network or application layer. However, solutions not applied at network layer are out of the scope of this Thesis.

IP mobility protocols manage mobility at the network layer and provide network level transparency, thus, the upper layers are unaware of MN mobility or the consequences of the underlying IP address change. Handovers can also be scoped differently according to the network domain organization. Global and local IP mobility protocols have been defined to provide mobility within the same domain or across network domains, respectively. Furthermore, the IP mobility protocol can be classified into two main categories; host-based and network-based. In the host-based category, the MN must participate in mobility related signaling. Whereas in the network-based, the network entities are responsible for the mobility related signaling without terminal intervention.

2.2.2.1 Client-based mobility management

In order to solve IP address mobility, the Internet Engineering Task Force (IETF) defined the Mobile IP (MIP) protocol [15], which realizes the mapping between a terminal's Home Address (HoA) and its current location-associated address, or Care-of-Address (CoA), and enables the forwarding of associated traffic using a tunneling mechanism between anchor points. The Correspondent Node (CN) is not required to know about MN's mobility and all transmitted packets go through MN's Home Agent (HA), the entity anchoring MN's HoA. The operation of MIP is as follows. As MN moves to a visited network, it initiates the Agent Discovery phase, in which it exchanges Agent Solicitation / Agent Advertisement messages with the region's Foreign Agent (FA). The Registration phase then takes places, which refers to the attainment of a new CoA using Stateful Address configuration or Dynamic Host Configuration Protocol (DHCP). Finally, the MN updates its mobility binding, stored at the HA, by means of a Binding Update (BU), to which HA replies with a Binding Acknowledge. From this point, all packets sent from and to MN will cross both the FA and the HA. Packets transmitted to MN are intercepted by HA, which encapsulates and tunnels them to the MN's CoA. FA is then responsible for decapsulating the packets and forward them to the MN. MIP has several drawbacks such as triangular routing and consequent communication latency, long distance mobility signaling and others.

Mobile IPv6 (MIPv6)[16] protocol introduced IP mobility over IPv6, and is similar to its IPv4 counterpart. Moreover, it supports a Route Optimization mode, enabling packets to travel directly from the HA to the MN, and as such solving the triangular routing issue, as well as improving network reliability, security and reducing network load. On the other hand, limitations such as high HO latency, which translate into packet loss, or the costly signaling have contributed to its slow deployment.

Mobile IPv6 Fast Handovers (FMIPv6) [17] tackled some of the limitations in MIPv6, namely the HO latency and its impact in service disruption during HO. The key concept behind FMIPv6 is to leverage on MN's awareness to mobility, through L2 HO triggers initiated at the previous network. Two modes are possible: in the predictive mode, L3 HO is initiated before L2 mobility completion, while in reactive mode L3 mobility is initiated after the MN transmits an Unsolicited Neighbor Advertisement to the new Access Router (AR). While FMIPv6 promises seamless mobility, it depends on the availability of a L2 trigger, and affordable triggering latency, which will affect the mobility signaling initiation pro-activeness.

Finally, Hierarchical Mobile IPv6 (HMIPv6) [18] improves MIPv6 by both reducing HO latency and signaling overheads resulting from frequent HOs. For such, it defines a mobility management hierarchy, differentiating local mobility from global mobility. This is achieved by the introduction of a Mobility Anchor Point (MAP), which acts as a "local" home agent. Basically, mobility within a MAP domain (local mobility) will trigger a BU to the MAP, hiding this process from HA and CNs, while mobility to a different MAP will be handled using regular MIPv6 operation.

Host-based solutions require modifications to MN's protocol, and IP address reconfiguration, which translate into waste of resources and complexity. Such schemes were designed at a time where processing and energy resources were restrict, with such signaling translating into additional overhead over the radio access network, further avoided the wide deployment of these proposals.

2.2.2.2 Network-based mobility management

Proxy Mobile IPv6 (PMIPv6) [19] is the network-based counterpart of MIPv6, and its acceptance lead to endorsement within several standardization bodies [20][21]. PMIPv6 provides network-based mobility support within a localized domain, enabling MNs to stay unaware to any L3 modification, and as such not being involved in the mobility signaling. The key advantage of this design option is the efficient use of wireless resources, because tunneling overhead over the wireless link is avoided. PMIPv6 partially reuses the logic from MIPv6 [22] by anchoring the MN's global address over a centralized entity - the Local Mobility Anchor (LMA). LMA is located at the network core and is responsible for advertising the MN's Home Network Prefix (HNP) - providing reachability to MN's address - and to forward any data from and to the MN. MNs attach to Mobility Access Gateways (MAGs), which are responsible for detecting MN's movement and initiate mobility signaling with MN's LMA on behalf of the MN. Such signaling is used by MAG for establishing a tunnel with the LMA, allowing the MN to use an address from its HNP. MAGs emulate MN's Home Network, and intercept and forward any data sent by MNs towards its LMA, by relying on policy based routing. When intercepting traffic destined to the MN, LMA forwards those packets towards the MN's Proxy Care-of-address (PCoA), which is the transport endpoint of the tunnel with the current MAG. When associating to a MAG, the MN configures one or more HoAs from its HNP.

PMIPv6 is a localized mobility management protocol with improved signaling update time against MIPv6. However, the service interruption in PMIPv6 still is a function of the link layer HO, which depends on the underlying technology used. In order to enhance the seamlessness of the PMIPv6 HO to support the Quality of Service (QoS) of real-time sensitive services and multimedia applications, a fast HO version of PMIPv6 was also developed [23], sharing similar L2 trigger limitations to those of FMIPv6.

2.2.3 Multiple interfaces management

To simplify the management of the increasing number of interfaces available on mobile devices, applications for simplifying the configuration and choice among the available interfaces have emerged, allowing users to focus on more demanding tasks. These software components typically store user preferences, security profiles, and manage terminal HO. Answering newer and more complex scenarios and corresponding applications requirements, these "connection managers" are extended to support new features such as smart interface balancing, per-application preferences, operator driven policies, monitoring-based selection, or activation of advanced mobility solutions. Recent proposals offer solutions for Wi-Fi authentication, mobility, offload management and simple traffic balancing mechanisms [24].

The simultaneous utilization of multiple interfaces, i.e. multilink, is a very interesting feature, as it allows users to efficiently "combine" multiple interfaces in order to increase the overall throughput, seamlessly making the aggregate bandwidth available to the user. This is an excellent added value to mobile UGC, since the stream's quality no longer gets restricted by the available network technology. Combining multiple interfaces - of the same or different technologies - on the fly and according to the application bandwidth requirements allows to greatly improve the overall users' experience. Multilink is addressed by multiple standardization bodies for under-L3 technologies, such as Multilink Point-to-Point Protocol (ML-PPP), Inverse Multiplexing over Asynchronous Transfer Mode (ATM) (IMA) or Multilink Frame Relay (MFR). Many of those items have been addressed within IETF, namely in Multiple Interfaces (MIF) WG [25][26]. The utilization of multilink for sending content is particularly important in Personal Broadcasting Service (PBS) scenarios, where the quality of the uplink transmission affects all of the video subscribers.

2.2.4 Facilitating handovers with IEEE 802.21

As referred in previous subsection, mobile devices are now multi-interfaced, supporting technologies such as WLAN and 3G/4G. Thus, operators providing multiple networks desire to facilitate user's access across their multiple technologies when using a single device. The support of seamless roaming and inter-technology HO is a key element to help operators manage and thrive from this heterogeneity. Operators who have the ability to switch a user's session from one access technology to another can more easily manage their networks and accommodate the service requirements of their users. Examples include mobility of applications to access networks providing improved performance, or load balancing operations for improved system performance and capacity [27].

IEEE 802.21 [28] defines a media-independent HO (MIH) framework aimed at improved HO between heterogeneous network technologies, i.e. aims for optimized and facilitated HOs in heterogeneous environments, providing a technology-agnostic way to control and retrieve information from access links. The standard defines the tools required to exchange information, events, and commands to facilitate HO initiation and HO preparation. IEEE 802.21 does not attempt to standardize the actual HO execution mechanism. Therefore, the MIH framework is advantageous to systems applying different mobility management solutions, and also for L2 mobility.

IEEE 802.21 goal is to support and enhance the intelligence behind HO procedures, which is achieved through three different services: Media Independent Event Service (MIES), Media Independent Command Service (MICS) and Media Independent Information Service (MIIS). The first service detects changes in link layer properties and reports appropriate events from both local and remote interfaces; MICS provides the ability to control and manage link layers (e.g. threshold configuration, order specific action, etc.); MIIS enhances the HO decision process with information about network configuration. These services are provided through a cross-layer Media Independent Handover Function, enabling high-level entities (dubbed MIH-Users) to control and access link layer information in a media-independent way.

2.2.5 Discussion

Mobility within heterogeneous scenarios faces a multitude of problems which the explosion of traffic just magnifies. The design of mobility protocols suited to the new traffic patterns and to the overall video ecosystem, while facilitating both horizontal and vertical HOs is necessary. Such mobility management protocols should also be energy-efficient, aware to the application, provide optimized routing, and keep the advantages of network-based mobility management, e.g. the involvement of the MN in the signaling process. Moreover, applications-layer entities should be provided with relevant information efficiently, and

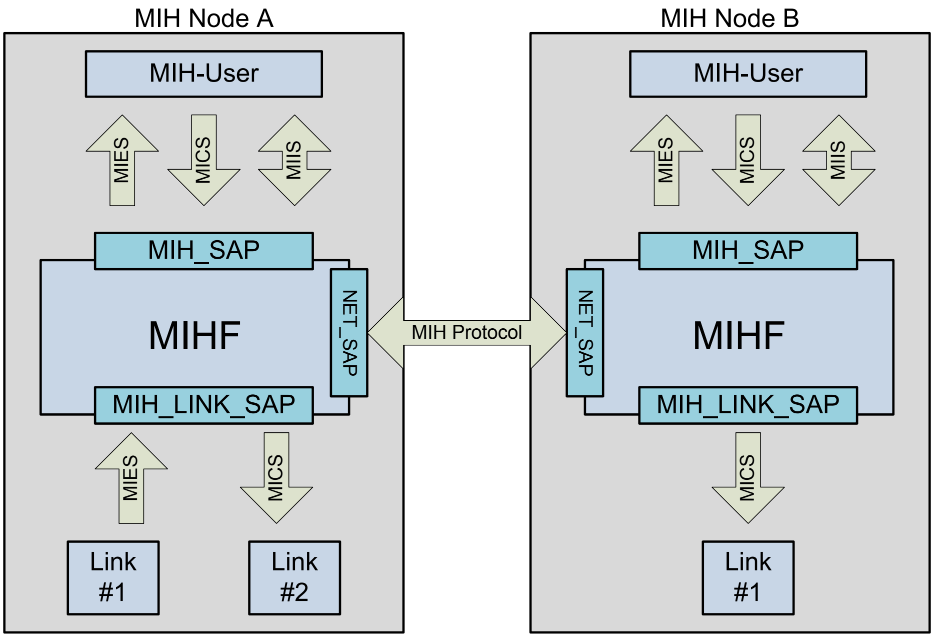


Figure 2.1: IEEE 802.21 framework

should have the necessary intelligence to make enhanced decisions while leveraging novel mobility management solutions.

2.3 Towards mobile access convergence

As data consumption habits shift to mobile, and with heavier and more recurrent traffic loads, radio accesses are put under more intensive stress. This is a concern for mobile operators, which seek means to cope with the increased data traffic peaks while simultaneously assure service quality to its costumers. The widespread availability of Wi-Fi is an opportunity for Mobile operators to naturally increase their networks' capacity and freeing up 3G and LTE accesses. The mechanisms and strategies adopted by mobile operators to deliver traffic originally destined for cellular networks are referred to as data offloading. This trend brings operator services closer to the notion of ubiquitous broadband access, ultimately improving user's brand loyalty. Although to accommodate this vision the whole 3GPP architecture must be adapted in order to achieve the same security and reliability levels of cellular communications when using WLAN access. One first facilitator of this fixed mobile convergence is the cellular industry focus on a single mobile broadband standard - LTE -, contrarily to previous generations. Although many other required extensions were defined during last years, others - such as seamless mobility support independently of technology- are still missing.

This section presents a background on the evolution of 3GPP towards WLAN - 3GPP symbiosis, or fixed mobile convergence, as well as relevant 3GPP data offloading mechanisms. While this section is not directly related to mobile multicast video delivery per se, it is still significantly relevant for comprehending previous and current design challenges associated to mobility within heterogeneous environments.

2.3.1 WLAN / Cellular interworking as a data offloading approach

There are two models of WLAN / Cellular interworking depending on the degree of integration of the WLAN network with the cellular infrastructure. In the first model - **Loosely Coupled** -, the WLAN performance is not under the cellular operator control or has not been integrated into a common converged wireless solution. As for the second model - **Tightly Coupled** -, the cellular operator has control over the WLAN operation, usually consisting of the integration between the two networks in a common core infrastructure, facilitating IP session continuity and overall user experience.

The initial steps for the inclusion of WLAN access in 3GPP environments were given in 3GPP Release 6 with one tightly coupled approach and one loosely coupled approach. In the first approach, named 3GPP-based enhanced generic access network (GAN) architecture [29], 3GPP and WLAN networks share the core infrastructure, which introduced the possibility to reroute cellular network signaling through WLAN, i.e. both Circuit Switched and Packet Switched services are run over WLAN access. In the latter approach, WLAN is not integrated into a common converged wireless solution. This solution, named interworking WLAN (IWLAN) architecture [30], enables the transfer of IP traffic between a User Equipment (UE) and operator's mobile packet core when using WLAN access. Neither of the approaches were widely deployed, and in most cases involved duplication of resources e.g. networks, policies and traffic management systems.

Initial work on offloading mechanisms started to be defined in the next 3GPP releases. Core network offload refers to the break-out of traffic from the mobile network, and to its offload to the Internet. This allows operators to avoid the need to process such traffic using the mobile packet core, this way reducing investment costs due to traffic increase. It is thus a crucial solution to the efficient delivery of video services. The first core network offload proposal was defined in Release 7 with Direct Tunnel [31], which was introduced for enabling Serving General Packet Radio Service (GPRS) Support Node (SGSN) bypass for user-plane traffic. Release 9 first proposed Femtocell work [32], or the possibility to deploy residential small cells operating in licensed spectrum.

Multi-Access Packet Data Network (PDN) Connectivity (MAPCON) is an EPC function developed to allow the UE to gain simultaneous establishment of PDN connections to different Access Point Names (APNs), via both 3GPP and non-3GPP access networks. This function is subject to UE capability and enables the selective or complete transfer of all

PDN connections from one APN to another, suitable for offloading traffic from the core network. Although, applications sensitive to mobility (e.g. VoIP) should not be offloaded, as IP connection may fail during HO.

Two main data offload solutions have been proposed in 3GPP's System Architecture 2 (SA2) working group, Local IP access (LIPA) and Selected IP Traffic Offload (SIPTO) [33]. LIPA enables direct communication between IP-capable devices located within a local network, i.e. without detour via the mobile operator's core network, resulting in higher quality and security. Such solution is aimed for residential or corporate deployment enabling local network access. LIPA functionality is provided by a Local GW (LGW) co-located with the Home eNodeB (HeNB). The other solution, SIPTO, may serve to offload indoor data - which accounts for the majority of the data usage - from home, offices or public places, reducing the cost of delivered data. SIPTO may be applied both at home and enterprise environments, in a similar fashion to LIPA, or at specific eNBs in macro-cellular access networks.

Regarding the breakout point, in LIPA it always takes place at the Local GW (LGW) in the local/home or enterprise femtocell network, while SIPTO-based offload for femtocell can take place at LGW similar to LIPA or above HeNB, such as at HeNB gateway. Considering macro-cellular networks, macro SIPTO offload takes place at or above the RAN. By breaking-out selected traffic closer to the edge of the network, operators may avoid overloading their scarce resources, i.e. PDN gateways (PGWs) and Serving gateways (SGW), as well as avoid inefficient routing in the mobile backhaul network.

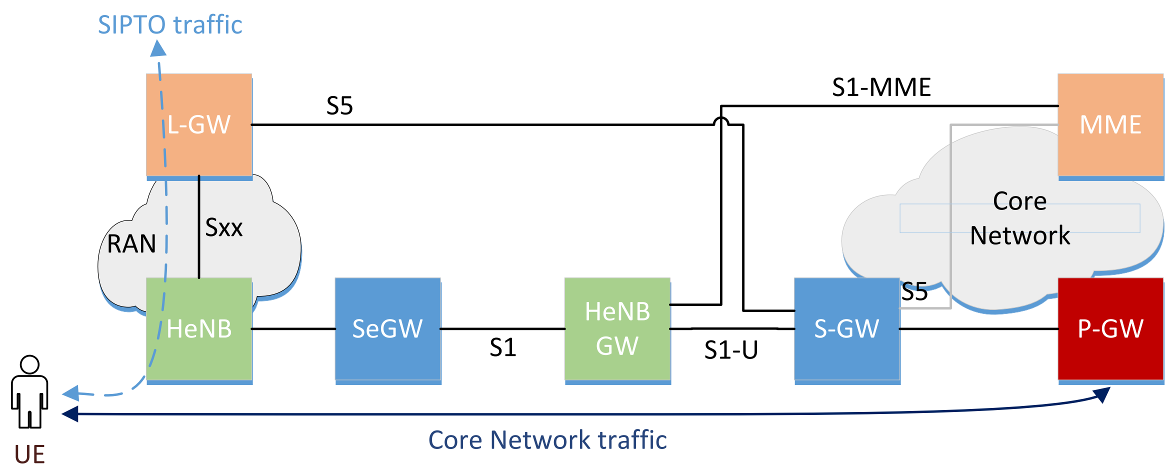


Figure 2.2: LIPA/SIPTO breakout at Local Network with stand-alone LGW

2.3.2 IP Mobility support in non-3GPP accesses

In Release 11, work on trusted non-3GPP access in EPC was developed with S2a-based Mobility Over GPRS Tunneling Protocol (GTP) (SaMOG)[34]. SaMOG allows mobility between 3GPP and non-3GPP networks that have a trusted relationship, allowing UEs to seamlessly HO between cellular and WI-Fi networks (Figure 2.3). As WLAN networks lack security in comparison to cellular ones, a Trusted Wireless Access Gateway (TWAG) is introduced, which acts as the perimeter security entity of the EPC network and connects to the PGW over a secure GTP tunnel - justifying the S2a reference. This solution is applies Extensible Authentication Protocol - Authentication and Key Agreement (EAP-AKA)/ EAP-Subscriber Identity Module (EAP-SIM)-based authentication, which takes advantage of 802.11x protocol to provide security.

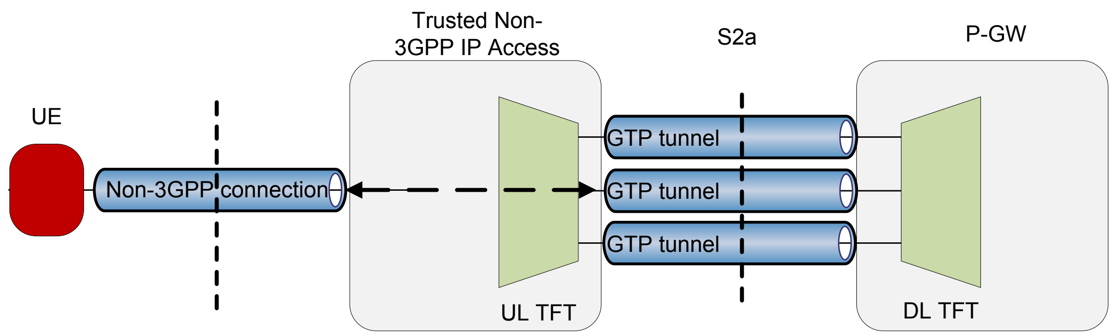


Figure 2.3: Mapping to Non-3GPP Access over GTP

In order to enable IP address continuity for certain IP flows of a PDN connection, IP Flow Mobility and seamless WLAN offload (IFOM) was defined in Release 10. IFOM allows traffic to be routed through either 3GPP or non-3GPP access network, with individual flows to the same PDN connection. IFOM is based on network policies, where different types of traffic are being forwarded to and from the UE through different Access Networks (AN) via individual flows. IFOM requires UE to be compatible with MIP family stack.

LIPA mobility and SIPTO at the Local network (LIMONET)[35] updates the requirements for deploying LIPA and SIPTO at the local network by adding support for LIPA mobility between HeNBs located in the local IP network (i.e. within a single residential or Enterprise network), and functionality to support SIPTO requirements at the local network, including mobility - either between HeNBs of the same local network, from the macro network to HeNBs or vice-versa.

2.3.3 Enhancing access selection and traffic steering

As referred, with the simultaneous connectivity using both non-3GPP and 3GPP accesses, it is crucial to connect to optimal Access Networks (ANs), and to direct traffic using the appropriate link. 3GPP initiated standardization of Access Network Discovery and Selection Function (ANDSF) [36] in Release 8, a framework which allows the customization of network steering policies and their dissemination to the UEs. By providing devices with relevant information and operator-defined policies to guide network selection decisions, smart network selection between 3GPP and non-3GPP access networks is possible. ANSDF is a server that uses the UE-to-ANDSF S14 interface to distribute the network selection information and policies, either using push or pull modes. Therefore, either the UE can query ANDSF server for information, or the ANDSF server can proactively distribute its information to desired UEs. The information provided by ANDSF aid UEs determine and prioritize which network to connect to, in a per-flow granularity - i.e., a particular Access Network (AN) may be preferred for a certain video service.

2.3.4 Discussion

In parallel to the integration of multiple access technologies, future cellular standards must also adapt the networks towards the enablement of distributed anchoring. Two main drivers are pointed out in [37] for this: first, operators are expected to cover a larger geographical area, which requires a significant number of PGWs with a certain degree of locality; second, the efficiency in delivering new services demands a network of service-anchoring gateways, for instance for coping with distributed caching systems.

Current proposals place LGWs as anchors closer to UEs, which poses services with improved round trip delays, etc. Although, seamless mobility is still not provided between LGWs. Taking into account service disruption for sessions requiring IP address continuity in SIPTO scenarios, the normative work in coordinated SIPTO: Change of PGW (CSIPTO) is in charge of defining service requirements enabling traffic to be selectively offloaded towards a defined IP network close to the UE's point of attachment to the access network. Among those requirements are the ability to support multiple connections associated to the same IP network, the (3GPP) system awareness to the session's address preservation, as well as to detect suboptimal connections and to establish a new / reuse an existing connection.

Moreover, the next steps should aim towards traffic tunneling between LGWs, which relies on coordinated work on several of 3GPP architecture's characteristics: enabling features such as multiple addresses over a APN, multi-PDN over a single APN, as well as the selection of LGWs over WLAN accesses are still missing.

2.4 Group-based services

2.4.1 IP Multicast

IP multicast was originally designed for achieving maximum efficiency in delivering packets, i.e., by having minimum packet replication during data transport. Due to its characteristics, it is used in group-delivery services, from file distribution such as software updates, to multimedia delivery and financial data dissemination.

IP multicast services rely on two main network functions: IP multicast subscription management and multicast routing. The former function, using Internet Group Management Protocol (IGMP) [38] in IPv4, or MLD [39] in IPv6, provides multicast routers with awareness to interested receivers and respective subscriptions; it is complemented by the latter function, which allows routers to build the transport paths or multicast trees for delivering multicast traffic. Throughout this document we focus on IPv6, and thus on MLDv2. MLD is an asymmetric protocol where receivers send MLD Reports stating their interests towards multicast routers, which process these and operate according to the multicast routing protocol.

Protocol Independent Multicast (PIM) is a popular family of multicast routing protocols which do not implement its own topology discovery mechanism, but instead use information supplied by unicast routing protocols. Two relevant derivations from this protocol include PIM Sparse Mode (PIM-SM), originally designed for Any-Source Multicast (ASM) model only, and its source-specific multicast (SSM) mode (PIM-SSM)[40], where a subscription is identified both by the IP addresses of the multicast group and the source node.

In a shared media Local Area Network (LAN), there may be more than one multicast router. A Designated Router (DR) is the single router which is elected to act on behalf of directly connected hosts with respect to the PIM-SM operations. In order to assure loop-free data transmission, PIM uses Reverse Path Forwarding (RPF), where the forwarding of a packet is determined based upon its reverse path. As a consequence, PIM routers will only forward packets which entered the interface associated to the Multicast Routing Information Base (MRIB) routing entry of the packet's source, otherwise discard them. PIM Join messages, which are used between routers to subscribe or unsubscribe IP multicast groups or channels, are also forwarded according to RPF which determines the upstream neighbor (or RPF neighbor) for a given subscription. PIM-SM operation comprises three stages:

1. In the first stage, a source sends traffic to its DR, which encapsulates the data into a PIM Register, and forwards it to the Rendezvous Point (RP). The RP, which is the root to multiple multicast trees, is then responsible for decapsulating the PIM Register packets and send the multicast traffic natively down the corresponding (*,G) RP Tree

(RPT), or shared tree.

2. After receiving the first PIM Register from the source's DR, the RP will send PIM Stop-Register packets for establishing a (S,G) Shortest Path Tree (SPT) with the source's DR, and then receive traffic natively (Stage 2).

3. A receiver's DR may then establish an optimized SPT by sending a (S,G) PIM Join directly to the source's DR (Stage 3). This tree is usually established after the data rate crosses a threshold, and will trigger the receiver's DR to unsubscribe from the RPT, through the transmission of a Stop-Register message to the RP.

An IP multicast datagram is transmitted by a source S to a destination address G. While in PIM-SM the MLD Reports only contain the group G of interest, in PIM-SSM receivers subscribe to (S,G) channels by sending a source-specific MLD Report. Thus, PIM-SSM corresponds to PIM-SM's third stage, or the direct establishment of source-specific trees. In PIM-SSM, a router may advertise itself as a RP, but it must not accept packets encapsulated in PIM Register messages, i.e. it only accepts native multicast data.

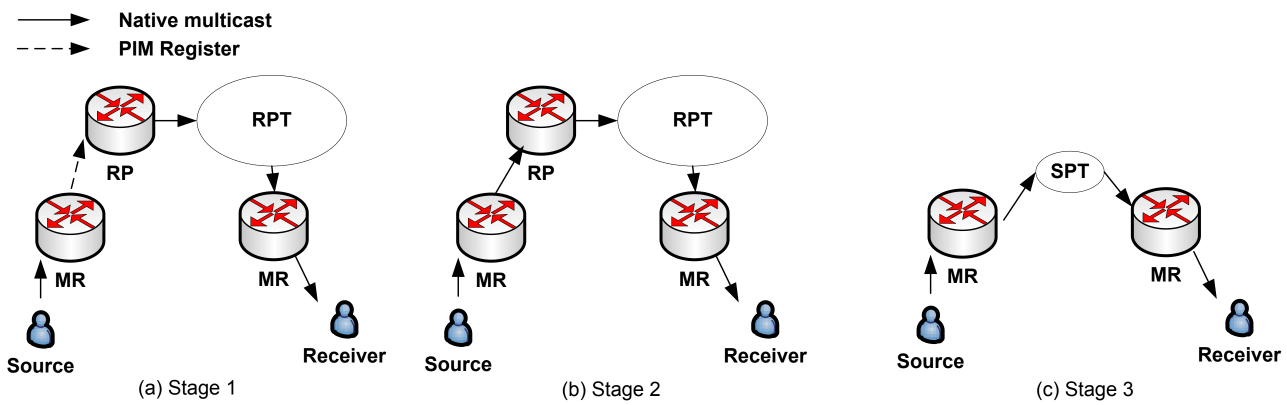


Figure 2.4: PIM-SM stages

2.4.2 IP Multicast in Cellular Networks: (e)MBMS

2.4.2.1 MBMS background

Aiming to optimize the distribution of video traffic over cellular networks, 3GPP introduced multimedia broadcast/multicast service in Release 6. Using new point-to-multipoint (p-t-m) radio bearers and multicast support in the core network, broadcast more efficiently covers all services where multiple users located in a common geographical area consume the same content at the same time. Broadcast enables the efficient and quick push of the same content - for example, caching of software updates, popular content (podcasts, news and music clips) or advertisements - to many devices at the same time without interaction with the user, allowing users to later access the content without any connection to the network. At the time MBMS over UMTS terrestrial RAN (UTRAN) was standardized, the industry was focused on the mobile TV use case. Digital video broadcasting for hand-held (DVB-H) networks was being deployed in several countries, and MBMS was seen as a way for mobile broadband operators to offer the same services over the UTRAN access.

MBMS did not reach commercial success due to several reasons, such as: 1) lack of user interest in programmed-start content, prefering instead immediate consumption such as VoD, 2) inferior service quality compared to other video services, e.g. in terms of bit rates (with up to 20 channels of 64 kb/s or 5 channels of 256 kb/s [41][42]); or 3) lack of interest from other necessary players such as hardware companies: these need real incentives for including MBMS-compatible chips, which translates into the construction of more costly devices. For addressing 1), broadcast could also be used for an efficient on-demand experience. In order to address issue 2), Rel-7 introduced MBMS over Single Frequency Network (SFN) transmission (MBSFN)). In MBSFN, a set of synchronized eNodeBs (eNB) transmit in the same resource block, i.e. at the same time. This enhancement increases the capacity by a factor up to 3 or 4 in certain deployment conditions [42], but also leads to new challenges such as the transmission scheme selection [43]. As for 3), it may be minimized by making multicast / broadcast technology on backend systems depend on software and not on hardware extensions.

2.4.2.2 LTE and eMBMS

The initial design of MBMS in LTE was motivated by a further optimization in radio efficiency and by the compliance with the flat LTE architecture without any Radio Network Controllers (RNCs). For this purpose, 3GPP redesigned the transmission schemes and considered both a multi-cell and a single-cell transmission [44]. In 2009, however, 3GPP suspended MBMS in favor of other, more general, short-term features not related to MBMS for the first LTE release (Release 8), and today some basic MBMS functionalities are specified for Release 9 and 10 [45]. MBMS standard has evolved into enhanced MBMS (eMBMS) that builds on top of the 3GPP LTE standard. eMBMS evolution brings improved performance thanks to higher and more flexible LTE bit rates, SFN operations, and carrier configuration flexibility. 3GPP Rel-11 brings improvements in the areas of service layer with, for example, video codec for higher resolutions and frame rate, and forward error correction [46]. eMBMS enables operators to offload the LTE network and backhaul, by introducing the possibility to deliver premium content to many users with secured quality of service in defined areas, pushed content via user equipment caching and machine-to-machine services.

eMBMS architecture is depicted in Figure 2.5. The management of eMBMS content and resources is done using a multi-cell / multicast coordination entity (MCE), a control entity responsible for admission control and resource allocation. During MBSFN operation, it jointly manages the radio resources of all eNBs in the MBSFN area. Another logical entity, the MBMS Gateway (MBMS-GW), is responsible for forwarding the eMBMS packets to the eNBs using IP multicast. In order to use broadcast, an entity named broadcast/multicast service center (BM-SC) is used to control broadcast sessions of the MBMS-GW and map incoming server traffic to broadcast bearers. BM-SC also owns the schedule of services as configured by the operator. Such schedule information is delivered to end-user registered devices so that applications or users themselves can decide to tune in to a particular service.

One of the persisting limitations with eMBMS is the lack of mobility support between different domains, which is a concern given the nature of most traffic aimed to be delivered, namely live video.

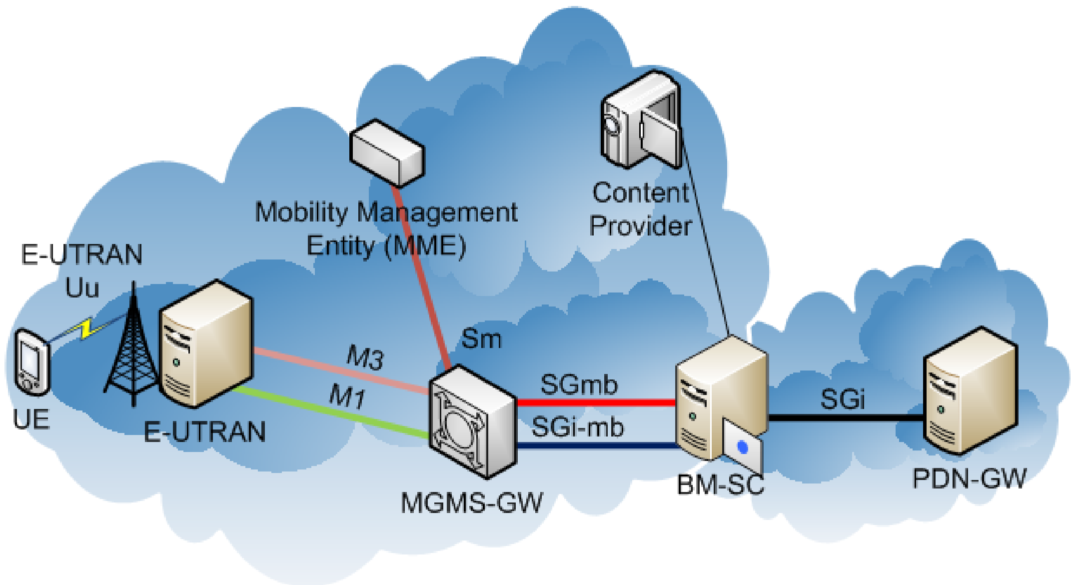


Figure 2.5: 3GPP Reference Architecture for eMBMS (without UTRAN)

2.4.2.3 Video streaming over eMBMS

The industry of video streaming has in the last few years moved from Real Time Transport Protocol (RTP)-based streaming to HTTP based streaming protocols like Apple's HLS [47] HTTP streaming has the benefit of using plain HTTP servers, and overcoming Network Address Translator (NAT) and firewall traversal issues. For HTTP streaming of files, the client requests media segments that are then pushed by the HTTP server. 3GPP standardized the Adaptive HTTP Streaming (AHS) protocol in Rel-9[48] - this and other 3GPP streaming standards are overviewed in [49]. AHS was then introduced in Rel-9 to allow the streaming of AHS content via eMBMS, and adopted by the Moving Picture Experts Group

(MPEG) as the baseline of what subsequently became Dynamic Adaptive Streaming over HTTP (DASH) [50]. 3GPP then defined its Rel-10 version of DASH, called 3GP- DASH [51], which is a compatible profile of MPEG-DASH.

2.4.2.4 Personal Broadcasting Services

From the perspective of 3GPP, PBS refers to a content distribution service using 3GPP services for enabling users to generate and broadcast content on air. For video services, this regards services where Internet users, private companies or mobile users share their videos using available wireless networks. It is supposed that a variety of broadcast services and a new device market may emerge once users are able to access the content distribution service. This vision has been considered within the Technical Specification Group for Service and Architecture (TSG-SA) of 3GPP in SA1, more involved in services, use cases and requirements [52].

User generated video can be distributed either using a time unconstrained VoD method, or a time constrained, real-time method. In the former case, the video contents are uploaded to a server and downloaded at the request of users, when desired. Most VoD applications may be delivered using unicast bearers; alternatively, VoD users may share the content using the MBMS in multicast mode. In the case of live broadcast, content is delivered to multiple users simultaneously, so a broadcast or multicast bearer is necessary for resource saving. A broadcast bearer service implies a pre-allocation of resources, so it may be efficiently deployed only in specific cases, e.g. when a large number of receivers are expected and the receivers are spread in many cellular areas. Major TV or Radio services are examples that may utilize such a broadcast bearer. UGC stream may also be transmitted using broadcast bearer if the population of receivers justifies the cost for delivering the content. The show of a very popular celebrity or daily episodes of a small production company are potential examples. Another use case is localized broadcast in areas such as a campus, theater or theme park.

3GPP defines a set of PBS use-cases; in the following we extract a subset which concerns video.

- Receive only personal streaming content services: this service is similar to real-time Internet TV or radio services. The idea is that an Internet or 3GPP user may register the service, and distribute contents at their convenience. The benefit is that 3GPP users may enjoy an abundance of real-time streaming content in addition to MBMS TV service. This service is a typical broadcast service.

- Interactive Personal Broadcast Services: A bi-directional content distribution service where the receivers may use uplink channel to interact with the content distributor.

Essentially, the interactive service offers users the opportunity to participate in broadcasting.

- Shared VoD Service: this service offers VoD-like broadcast service by sharing content streams with multiple users. Similar to VoD service, users browse a list of contents and send their request for content distribution to a service provider. When a specific content has been requested by a large number of users, the service provider checks if a multicast stream for the selected content can be allocated in the cells where the users are located. If there are sufficient resources available, the service provider broadcasts a notification to users, so that they may join the multicast session and start receiving the content stream, maximizing the opportunity for resource sharing. The key aspect of this service is the use of a multicast service bearer for providing VoD-like service in order to save bandwidth. However, the level of efficiency may also be achieved using a broadcast service bearer if the population of users justifies global allocation of resources.

2.4.3 IP multicast mobility

Multicast and mobility protocols were designed independently from each other, leading to several limitations in their interaction. As the resulting issues are intrinsically related to the multicast role performed by the MN, we organize the related work and related issues based on that viewpoint.

2.4.3.1 Mobility of multicast receivers

The main problem arising from multicast receivers mobility is the noticeable interruption in the associated service. The originating latency is comprised from two processes: the multicast subscription discovery at the new AR, and the consequent multicast join process. Most solutions are aimed at solving this, and as a consequence of handling IP multicast as a afterthought to mobility management, result in transport efficiency shortcomings.

Host-based mobility

Two main solution classes were proposed for IP multicast mobility support in MIPv6: Home Subscription or Remote Subscription [53]. The first takes advantage of the bidirectional tunnel to forward subscription requests and multicast traffic between the Home Agent and the MN. It does not solve subscription latency and introduces the tunnel convergence problem, which refers to a Home Agent having to encapsulate the same subscription to multiple users, even if located in the same network. The latter enables optimal multicast routing, but does not reduce latency driven from multicast subscription after mobility. Hybrid solutions attempt to combine advantages of each solution class [54] - such class of solution usually relies on overlays, thus is out of the scope of this work, which tackles native IP multicast transport only. Other problems typically affecting multicast mobility include out-of-order reception or packet duplication, which depend on the aid from upper layers or smart buffering schemes.

Network-based mobility

PMIPv6 does not explicitly define support for IP multicast. Given MNs unawareness to mobility, the total interruption time will derive not only from the subscription learning and join processes, but also from the delay pertaining to the attachment detection in the new link. The moment the MN's subscriptions are done will strongly depend from the MLDv2 timers, whose default and recommended values are 10 (RFC2710) [55] and 5 to 10 seconds (RFC6636)[56], respectively. Multicast Mobility (MULTIMOB) WG [1] proposes base support for mobile multicast receivers by deployment of MLDv2 Proxies in MAGs [57]. This proposal leads to multicast replication driven from tunnel convergence - known as Tunnel Convergence problem -, and does not avoid subscription latency, not being a satisfying solution for all applications. The former issue refers to the case where a MAG receives the same multicast subscription from multiple LMAs, and is tackled by using a dedicated multicast anchor or using direct routing [58]. The latter issue led to several proposals towards fast multicast subscription during HO, such as [59], [60] or [61], none of which tackles latency and transport efficiency issues simultaneously.

2.4.3.2 Mobility of multicast sources

The mobility of a multicast source has the potential to impact the whole multicast communication, and comprises address transparency and temporal HO constrains [54]. For the former, native forwarding of multicast will be bound to the source's topological network address, due to RPF checks. For the latter, the reconstruction of a source-specific tree due to the mobility of a source imposes interruptions in the service for all subscribing users.

Host-based mobility

When forwarding traffic down the shared tree, mobility may lead to temporary interruption of the transmission, and when forwarding traffic natively towards a source-specific tree, the consequence may be the complete reconstruction of the associated (S,G) SPTs at the new location, due to the bound to the source's topological address and the associated RPF check. As such, address transparency must be assured at two levels: first, the source's address must be in topological agreement with the multicast forwarding tree, due to the RPF check;

[1] http://datatracker.ietf.org/wg/multimob/

secondly, the logical node identifier, typically the HoA, must be presented as the stream's packet source to the transport layer of the receiver side [53]. Other limitations include packet loss, which affects all receivers, and multicast scoping restrictions at the new network. The SSM model was designed as a lightweight approach to group communication, and the addition of mobility management should preserve this feature. However, the forwarding of native multicast leads to several of the aforementioned routing problems. [62] presents a concrete way to support mobile sources in MIPv6, by elongating the root of the previous tree to the current DR.

Network-based mobility

A solution based on MLD Proxy provides minimal support for sending IP multicast in PMIPv6, by keeping involved multicast routing trees stable. However, it does not assure seamless data transmission, resulting in the loss of several packets during the mobility signaling process as a consequence of the application unawareness to mobility [63]. The deployment of multicast routing functions in MAGs enables always optimal routing, as opposed to solutions based on MLD Proxy functions. Although, it introduces an additional issue: the SPT creation and its update after HO. When coupling PIM with PMIPv6, LMA's Multicast Router (MR) cannot determine the upstream router for reaching the source MN: while it knows that MN is reachable using the bi-directional tunnel interface, it does not see any neighbor MR from the tunnel interface, but only from the PCoA. In other words, it does not associate the MAG's PCoA address as the upstream router for reaching the source. Moreover, as PIM is agnostic to source mobility management, these trees will be lost after HO, and tunnel re-establishment will take place at the new MAG. Even assuming that PIM is able to recover the states properly, this process is considerably slow [63]. The work in [64] addresses HO latency in PMIPv6 source mobility scenarios, and proposes both solutions based on MLD Proxy (MLD Base solution, hereby represented as MBS) and on PIM routing (represented as Direct Multicast routing scheme - DMRS). The solutions are analyzed and evaluated experimentally. MBS scheme presents smaller HO latency at the cost of non-optimal routing and potential for tunnel convergence problem, while DMRS assures local content distribution with locally optimized traffic flows. In this work, a solution retrieving the best of the two worlds is proposed.

2.4.3.3 Multicast mobility using multiple interfaces

Packet loss is further exacerbated when considering inter-tech multicast HOs. General management issues of multiple interfaces is specifically targeted in Multiple Interfaces (MIF) WG[2]. Focusing in PMIPv6 scenarios, inter-technology mobility and other features such as

[2]http://datatracker.ietf.org/wg/mif/charter/

multihoming can only be supported through specific software configurations at the MN, which motivated the research of Logical Interface solutions [65]. Still, the dynamic transfer of a IP multicast session to a new interface completely lacks considerations in any of these efforts. [66] presents an efficient bicasting scheme duplicating the downlink packets over the old and target MAGs, in order to minimize packet loss when receiving multicast data. A similar issue exists when a MN is the traffic source, as it may be lost at the previous MAG's buffer once the tunnel between the LMA and the new MAG is established. Although, to our knowledge there are no efforts addressing either uplink connectivity or vertical HO issues when moving between MAGs with different radio technologies - which also differently impacts source and receiver mobility. Specifically, the vertical HO of a source adds the need for synchronizing the L2 and L3 HO processes - i.e. the switch between interfaces and the tunneling update, respectively - in order to effectively transmit using the interface associated to the active mobility tunnel. When acting as a receiver, applications relying on IP multicast must use specific service interface calls whose listening state is both socket- and interface-specific, meaning that vertical HOs require the use of a different service interface after HO. Without mobility-awareness, the application cannot invoke the subscription on the target interface and receive the multicast channel(s), even if the target multicast network already has the subscription(s) of interest. Summarizing, the unicast nature of the network stack fabric implies modifications to enable inter-technology multicast mobility of both sources and receivers in order to effectively support time-constrained services such as video. We highlight there is no single solution - and consequently implementation - supporting both rapid multicast receiver mobility and seamless source mobility in heterogeneous networks, and claim that IEEE 802.21 can be key for this goal if wisely adapted. As such, we will now briefly present relevant state of the art work on the topic.

[67] proposes a solution for fast multicast subscription taking advantage of IEEE 802.21. It introduces new MIH Information Elements and messages for applications' QoS requirements, but the focus is on the HO decision algorithm for L2 mobility of multicast receivers between PoAs - no considerations on L3 mobility are done. Besides, the evaluation of the solution is limited to mathematical analysis. [68] presents a dynamic playback control for multicast streaming based on IEEE 802.21, aiming to reduce the influence of HO between heterogeneous networks. It should be highlighted that neither of the aforementioned solutions considers multicast source mobility support or presents experimental results.

2.4.4 Discussion

IP multicast is one of the enablers for efficient distribution of multimedia content. Although, it has been added in previous mobility management solutions a posteriori, i.e. being "stacked" over independently defined unicast IP mobility solutions. Alternatively, IP multicast support, and how to preserve its transport efficiency, should be taken into account in the design of novel mobility management solutions. Moreover, the efficient delivery of video in cellular networks using eMBMS lacks considerations beyond intra-domain mobility.

2.5 Video Playback and Transport technologies

2.5.1 Background on Internet video

Several systems are currently used for video delivery over the Internet. IP Television (IPTV), for instance, is used to deliver video through ISP's networks, and which replaced legacy broadcast TV formats. It enables different viewing modes, such as live TV, where IP multicast is typically used, using streaming or on-demand modes. Other widely used system is Content Delivery Networking (CDN), consisting of the placement of multimedia content closer to users in order to minimize routing distances and bottlenecking possibilities. It is particularly important for high-rate multimedia delivery, unburdening core networks and links from significant loads though the optimal positioning of caches at the edges of the networks.

From a transport perspective, initial mobile video players had two basic options for video transfer: download using HTTP or streaming via Real Time Streaming Protocol (RTSP) [69]. HTTP was only used for download and play, with the whole file being downloaded before playback. The possibility for rendering while simultaneously downloading, known as progressive download, was introduced later. Although, downloading prior to playback incurred significant latency in case of bandwidth limitations, which led downloaded videos to become shorter and of low quality, in order to reduce file sizes and shorten download times. This fueled the adoption of RTSP by many mobile devices, as a protocol requiring little data to be buffered before playback, meaning lower playback latencies. Given the importance of the download and streaming modes in the video delivery background, they are now further described.

2.5.1.1 Video download

The most practically used protocol for video file delivery, in particular in mobile platforms, is HTTP. The simplistic and as-fast-as-possible operation of HTTP was designed for delivery of small amounts of data with minimal latency. Video files, however, tend to be much larger than HyperText Markup Language (HTML) pages. HTTP is built upon TCP to ensure data integrity. For video, data integrity ensures that the intended picture quality is achieved, given timely delivery of data. In broadband-connected desktop environments, packet loss is relatively rare and bandwidth is relatively plentiful. However, in constrained wireless networks such as 2G or even 3G networks, where lossy, congested, high latency, low bandwidth conditions proliferate, the relatively high number of high latency retransmissions required to support data integrity can disrupt playback continuity. Besides, the greedy traffic delivery pattern leads to premature delivery of data, which may more easily provoke congestion, and result in a waste of resources for cases where the user only watches part of the video.

HTTP is the de facto option for most data delivery, including video. The main advantage of HTTP is its ubiquity, given factors like the availability of Web browsers in all mobile devices, the wide acceptance of HTTP for firewall traversal, the interchangeability of stateless HTTP servers, and the existing Content Delivery Network (CDN) data hosting and delivery infrastructures.

2.5.1.2 Streaming

Streaming relies on just-in-time data delivery with just-in-time rendering. A plethora of proprietary streaming protocols exist, with the most common standardized protocol being RTSP, a control protocol supported by RTP to deliver individual video frames over unreliable User Datagram Protocol (UDP) transport. Just-in-time delivery uses less instantaneous bandwidth than as-fast-as-possible delivery and typically requires less client buffer space to be reserved. Reducing bandwidth usage over time reduces congestion probability, as long as the delivery rate is inferior to the available bandwidth. This paced delivery can also prevent unnecessary bandwidth usage when user access patterns include random seeks or incomplete viewing.

The real-time nature of streaming also makes it suitable for delivering live video. Though streaming is more bandwidth-efficient because frames arrive at the last possible moment, there is no time for retransmissions. As such, there is no advantage in using a reliable transport like TCP. UDP provides “graceful” degradation of picture quality with minimal playback stoppages. As individual frames are lost, pixelization or rendering distortions will be noticeable to the user. Frame-based delivery allows for intelligent dropping of frames (e.g., non-key frames, or frames from low motion scenes), but this requires an intelligent network that knows which frames to drop. This level of intelligence is not generally found in the Internet today. Frame-based delivery combined with feedback from the Real Time Transport Control Protocol RTCP can also be used to implement dynamic video bit rate adaptation.

2.5.1.3 Hybrid solutions

Hybrid schemes rely on HTTP for data delivery, which make them ideal for use with CDNs, whose infrastructure is already optimized for distributing mass quantities of data via HTTP.

Though CDNs do support RTSP and other streaming protocols, the overhead of maintaining separate, more specialized servers makes supporting those protocols expensive and less desirable. Beyond the significant processing overhead, RTSP, in particular, also requires the use of four UDP channels (two RTP connections, one for audio, one for video, along with their corresponding RTCP connections) which further limits server scalability and complicates network design. For many networks, dynamic provisioning for a large range of UDP ports is undesirable as it typically requires real-time firewall "fixes" which tax the firewall and in many cases violates security policies. With RTSP/RTP, there is also the issue of gracefully degraded quality, due to random packet loss (e.g., network error or discard-based traffic shaping). Graceful degradation is non-deterministic and undesirable to content providers. Many interesting schemes have been shown to improve quality and predictability by limiting key frame loss [70], recovering from key frame loss [71] or proactively dropping non-key frames [72], but the non-deterministic nature of loss remains unchanged. With hybrid schemes, TCP-based transport guarantees frame delivery, though not necessarily on-time delivery. Late delivery may result in playback stoppage, but stoppage is deterministic in terms of rendering (unlike pixelization due to frame loss).

We now present a selection of the most relevant video compression and decompression protocols, both those currently in use and emerging ones.

2.5.2 Video transport

2.5.2.1 RTP and RTCP

RTP is a standard for audio and video delivery over IP, and is used for transporting media streams in a timely manner - contrarily to TCP, which prioritizes reliability. It is a data transfer protocol used together with the RTCP, its control protocol counterpart responsible for monitoring transmission statistics and QoS, and for synchronization of multiple streams. RTP is originated and received on even port numbers and the associated RTCP communication uses the next higher odd port number. Sessions are characterized by an IP address and those ports, with video and audio using different RTP sessions. Thus, typically 4 ports are required for delivering a video stream over the network.

Moreover, RTP can be used for transporting either unicast or IP multicast streams.

2.5.2.2 Adaptive bitrate streaming

Adaptive bitrate (ABR) streaming solutions work by continuous inspection of user's bandwidth and CPU, adapting the video transmission rate accordingly (as depicted in Figure 2.6). It depends on two entities, the encoder and the client. The encoder generates multi-bit-rated versions of a single source video, typically distributed using a CDN, and the client selects among the different encoding as necessary. Several proprietary versions were defined such as Adobe's HTTP Dynamic Streaming or Microsoft's HTTP Smooth Streaming, but the most popular is HTTP Live Streaming (HLS), currently under standardization. HLS relies on a hierarchy of m3u8 playlist files, which are extensions to the m3u format used for mp3 audio playlists. The top level playlist contains static pointers to separate playlists for the individual bitrates, and each of the bitrate playlists contains a rolling list of pointers to segments. A segmenter is responsible for recording from the live stream and transcoding segments into the different target bitrates. Once new segments are available, the bitrate playlists are updated, adding the new segment and removing the oldest segment. The new segments and the updated playlist are pushed to the CDN for delivery at regular intervals. A client is passed a link to the master playlist, from which it obtains a list of available bitrates. Once a bitrate is selected, the client begins polling the playlist file corresponding to the selected bitrate. After the initial read of the playlist, subsequent reads ideally occur at a regular interval equal to the duration of the segments. For uninterrupted playback, there must be alignment between the asynchronous upload of segments and playlists, the polling for playlist updates, and download and rendering of segments. The HLS specification provides guidelines for polling and polling retry delays, where the delay should be equal to the segment duration if the playlist has changed, or half the segment duration for the first retry if the playlist has not changed.

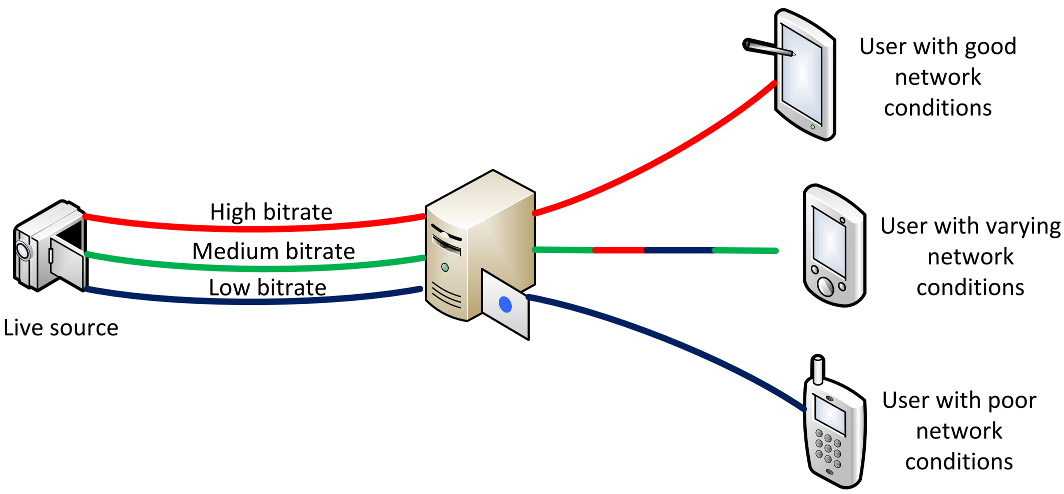


Figure 2.6: General scenario applying adaptive bitrate streaming

2.5.2.3 MPEG transport standards

MPEG-2 Transport Stream (TS) is the most notable standard devised by MPEG for multimedia delivery. MPEG-2 TS provides efficient mechanisms to multiplex multiple audio-visual data streams into one delivery stream according to consumption order. Audio-visual data streams are packetized into small fixed-size packets and interleaved to form a single stream in the order of synchronized playback. In addition, information about the multiplexing structure is contained in the data packets allowing the receiving entity to efficiently identify each stream. Moreover, sequence numbers identify missing packets at the receiving ends, and timing information is assigned after multiplexing with the assumption that the multiplexed stream will be delivered and played in sequential order. These design principles made MPEG-2 TS a perfect solution for multimedia streaming for a large number of users.

However, the increase of personalized, on-demand viewing of multimedia content over the Internet has introduced the requirement for more flexible access to the content. MPEG-2 TS cannot efficiently support personalized advertisement or audio-stream selection with a language suitable for a specific user because these require streams demultiplexing and remultiplexing.

Dynamic Adaptive Streaming over HTTP (DASH) is an open source adaptive bitrate streaming solution designed by MPEG group. It leverages on HTTP and follows a segment-based solution like HLS, breaking content into segments of short duration. As such, its goals are to be universally implemented, unlike vendor-specific solutions such as HLS. DASH specification considers two types of containers: MPEG-4 and MPEG-2 TS. DASH was MPEG's standard for efficient and simple streaming of multimedia applying currently available HTTP infrastructure (CDNs, proxies, etc), and allows the deployment of streaming services using existing low cost and ubiquitous internet infrastructure without any special provisions adaptation[3].

On the other hand, the MPEG Media Transport (MMT) is being developed as part 1 of International Organization for Standardization / International Electrotechnical Commission (ISO/IEC) 23008, High Efficiency Coding and Media Delivery in Heterogeneous Environments (MPEG-H), and represents the next stage in the MPEG protocol evolution. MMT is inline with Content-Centric Networking (CCN) network paradigm[4], where the existence of intelligent caches is assumed. These caches are positioned close to the receiving devices, and are responsible for actively caching content and for adaptively packetizing and pushing it to the receiving devices. In MMT, the content may be accessed at a finer grain through uniquely identifiable names instead of just their location. Besides, small chunks of content are cached close to the receiving entity regardless of the specific service provider and their location [73]. MMT is thus MPEG's recognition that future networks will require a multimedia transport solution more aware to the characteristics and requirements of the underlying delivery networks.

[3]http://ride.chiariglione.org/new_ways_of_transporting_bits.php

[4]While CCN paradigm is not given particular attention in this Thesis, the author is aware of its importance and the role it may take in future video delivery.

2.5.2.4 Multicast video transport

In order to meet the heterogeneity of converged networks and the diversity of user terminals, video delivery must adapt to network changes. There are two main approaches to this issue in the research community. In the **single-rate multicast** mode, the source transmits at one fixed rate, or the rate is adaptive and defined by either the receiver with the lowest bandwidth capacity [74], or by an inter-receiver fairness objective [75]. In **multi-rate multicast** schemes a video file is transmitted using multiple layers (streams). Although single-rate multicasting is easier to implement, it has scalability limitations. Multi-rate multicasting exhibits better scalability, is more flexible and can make more efficient use of network resources.

Moreover, multi-rate multicasting has two basic modes. In the **layered mode**, each video file is encoded to one base layer and several enhancement layers. The layers may be interrelated for cumulative layered multicast, or may be operated independently. **Simulcast** is referenced to the transmission of a number of independent streams with the same content that differ in quality and hence in bandwidth requirements. The advantage of having different versions of the same content is that it does not require more sophisticated encoders [76]. Simulcast technology is fairly simple when compared with layered encoding. The drawback, however, is that the multiple versions of the same multimedia information are transmitted over the network in parallel, thus representing additional spent bandwidth. While this is done so that users can choose the appropriate version at any given time, it is crucial to minimize the number of transmitted streams to free up network resources. This is feasible only if the transmitted streams are adaptive, so that they can serve a large number of users with similar receiving capabilities.

2.5.3 Video codecs

2.5.3.1 H264

H.264-Advanced Video Coding (AVC) was defined by an effort from the Video Coding Experts Group (VCEG) and the MPEG, which formed a Joint Video Team (JVT). The standard uses two layers: the Video Coding Layer (VCL) and the Network Abstraction Layer (NAL). Briefly, the former is used to create a coded representation of the source content, providing flexibility and adaptability to video transmission, while the latter is used to format the VCL data and to provide header information on how to use the data for network video delivery [77]. In VCL, each image is partitioned into smaller coding units (macroblocks) which are themselves comprised of independently parsable slices. These slices are partitioned into three groups which allow flexible partitioning of a picture: Partition A, which defines macroblock types, quantization parameters, and motion vectors; Partition B, for intra partition; and Partition C, for inter partition. NAL units are the video data encoded by VCL plus a one-byte header that shows the type of data contained in the NAL unit; one or more NAL units can be encapsulated in each transport packet. NAL units are classified as VCL NAL units, which are coded slices or coded slice partitions, or non-VCL NAL units, which contain associated information, namely sets of parameters and supplemental enhancement information (SEI). Each coded video sequence is an independently decodable part of a NAL unit bit stream, and starts with an instantaneous decoding refresh (IDR) access unit. Furthermore, the IDR access unit and subsequent access units are decodable without decoding any previous pictures of the bit stream. The NAL payload is transmitted with different priority according to the Nal_Ref_Idc, of the NAL unit header.

2.5.3.2 H264 SVC

H264-Scalable Video Coding (SVC) has improved scalability over AVC by enabling the transmission of different layers of a video sequence from the same file [78]. The Base Layer contains the most basic representation of the video sequence, and consists of the lowest quality representation in each of the spatial, temporal and quality dimensions (Figure 2.7). Other layers may then be encoded, referred as Enhancement Layers. Each of these layers represent a point in the 3-dimensional space, and is seen as an improvement in terms of one or more of the 3 dimensions. In order to decode an Enhancement Layer, it is required that all of the lower layers have been received and decoded successfully. The scalability of SVC is achieved by tailoring the visual quality of each sequence to the target device(s), or the network status (e.g. bandwidth restrictions or effective congestion). It is possible to achieve scalability at each of the three dimensions: Spatial scalability refers to the resolution of decoded video, Temporal scalability refers to the video in terms of displayed frames per second, and Quality scalability refers to the level of compression degree used during encoding of the source video.

A H.264 SVC stream is generated by selecting one or various of these scalable dimensions, with this selection being decided prior to the encoding phase - as a consequence, during playback it is required to scale up or down along the same path chosen during the encoding process. For instance, when encoding using spacial and then temporal scalability (two layers), the video is then upscaled along the same order. In case the goal is to switch to a lower layer during playback, the reverse path must be followed - in this case, the temporal dimension would be the first to be decoded, and so on.

2.5.3.3 HEVC / H265

High Efficiency Video Coding (HEVC) standard is aimed at decreasing the bandwidth requirements of video services by providing a 50% increase in compression efficiency over

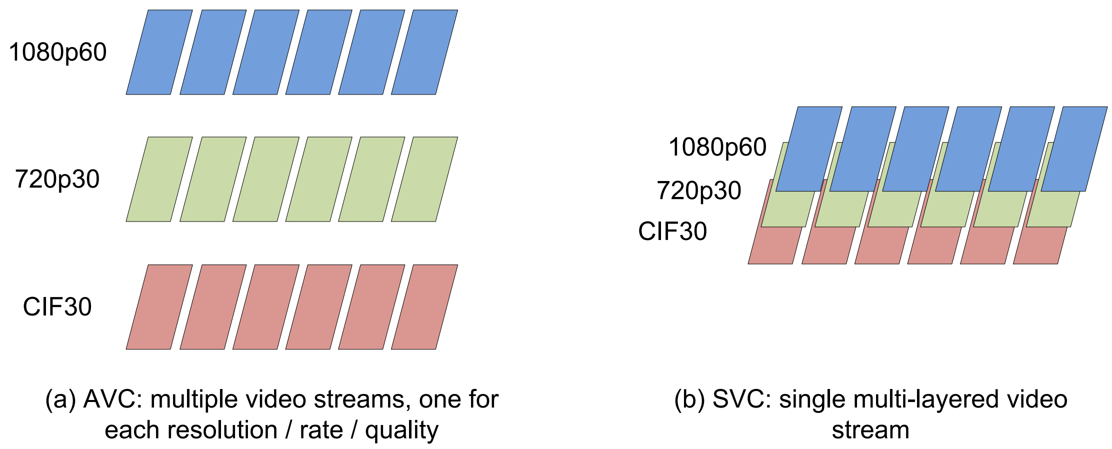


Figure 2.7: Example reflecting the differences between AVC and SVC

previous H.264-AVC standard, while keeping the same level of perceptual visual quality. Similarly to H.264-AVC, HEVC consists of a VCL and a NAL layers. The coding structure used in a VCL layer for each picture is significantly changed [79]. While in H.264/AVC each picture is divided into macroblocks (with 16x16 luma samples), which can be further divided into smaller blocks (16x8, 8x16, 8x8, 8x4, 4x8 and 4x4), in HEVC pictures are split into Coding Unit (CU) treeblocks of up to 64x64 luma samples, with the highest level of the treeblock structure referred to as the Largest Coding Unit (LCU). Tree block structures can be recursively split into smaller CUs through a quad-tree segmentation structure - CUs can vary from squared 8x8 to 64x64 luma samples. The higher compression gains can be achieved using larger CUs on homogeneity regions within a picture with little or no motion between two adjacent pictures, when using intra-prediction and transforms.

2.6 Mobile video delivery and associated challenges

The challenges associated to video delivery go beyond the requirement for a massive network capacity, and involves several factors. The first one - and the root for most of the challenges associated with video delivery - is Internet's base architecture itself. Internet was designed with a limited set of requirements, not taking the specificities of video traffic into account. The evolutionary nature of Internet, motivated by the emergence of other requirements such as mobility and security, led to an ever-morphing pile of patches,which ultimately resulted in the increase of complexity and the risk of failure [80]. Besides, with the multiple technological advances, the user's expectations for improved services increased and further elevated the stress on existing networks. For instance, Internet's Best-Effort nature collides with the fact that different data types, such as video, rely on varying requirements in terms of delay and jitter, etc, and processing overhead resulting from decoding and other costly tasks. The switch of Cellular networks to an all-IP model highlighted some of Internet's limitations, such as TCP/IP performnce under wireless scenarios.

Some of the challenges pertaining to the video delivery ecosystem include significant fragmentation, as a result from the diversity of mobile equipments and their characterisitics (e.g. Operating Systems), which imply the availability of each specific "content" in compatible video compression formats. Similarly, different video transport technologies and protocols such as RTP, HLS or Smooth Streaming, among others, are adopted by different Service Providers.

Further increasing the costs and the difficulty for efficiently delivering video to subscribers, there is a variety of mandatory mechanisms and optimizations, such as caching, buffering or adaptive bit rate technologies, which add complexity to the networks. Even with today's advances in computation power, storage and network connectivity, mobile devices still present some limitations for handling intensive processing tasks, such as 3D graphics applications, which significantly drain the limited battery. To minimize this, streaming-based techniques are used to remotely visualize multimedia content on mobile devices, shifting the processing task from the client to the cloud (i.e. powerful remote computers) [81].

Another recurring problem is that of scalability. The diversity in user equipments introduces not only the need for multiple transport technologies, but also for the simultaneous delivery over multiple rates when using a single protocol, for coping with both vanguard and less powerful devices. Besides, this introduces the need for deciding between the placement of the encoder / transcoder either in the same or in different network sections: while the first leads to higher bandwidth usage, in the latter any lost packet at the core will affect all devices, independently of the video characteristics (e.g. quality, frame rate, resolution).

Concerning IP mobility, some video streaming protocols will endure IP address change better than others. For instance, while in progressive download the whole video file needs to be requested after the IP change, in adaptive streaming or other chunk-based protocols like Real Time Messaging Protocol (RTMP) only the latest chunk needs to be requested again. Stateless protocols like HTTP are compliant with IP address change in the sense that no re-negotiation at TCP layer is required after mobility. As such, the applications will send requests as soon as the new IP address is configured, and, considering CDN operation, the Service / Request Router will redirect MN's requests to a new cache, based on its new IP address. On the other hand, applications relying on stateful transport protocols (e.g. RTMP or RTSP) are required to go through the negotiation process, thus IP address continuity is required in order to assure a smooth HO. Whatever is the case, activation of IP mobility mamagement protocols will bind the MN's requests to the anchor's location, obfuscating most CDN-optimization mechanisms and leading to non-optimal routing.

The diversity in service requirements motivates the need for mobility management solutions taking into account the video properties. Not having IP mobility service at all represents the failure of several services at IP address change, but it's also not efficient to activate IP mobility by default, due to the amount of traffic unnecessarily anchored - and consequent overhead - or non-optimal routing. Awareness to the need of a session's IP address continuity is thus a requirement if efficiency is to be achieved, either with the application explicitly requiring it, by simply providing awareness to the network about such need, or both.

2.7 Cross-layer design for video delivery in mobile networks

2.7.1 Background on cross-layer design

Previous protocol frameworks such as OSI reference model follows a black box paradigm.which rely on stratification, a composition mechanisms which defines each protocol layer impervious to the functionality embedded within other protocol layers [82]. Using this model, information within a protocol stack may only be exchanged between adjacent protocol layers, following the concept of service access point (SAP). In recent years, research efforts have explored ways to exchange information between non-contiguous layers, thus violating OSI reference model; these methods are referred as cross-layer design. Looking at the particular case of wireless communications, cross-layer approaches emerged as a way to overcome design assumptions leading to wireless network performance degradation. One typical motivation is TCP performance over wireless.

This section presents state of the art work on cross-layer approaches aimed towards improved video support over mobile scenarios.

2.7.2 Cross-layer optimizations for video delivery

There are several cross-layer proposals targeting video support, some of them summarized in[82]. A cross layer optimizer (CLO) is used in [83] for optimizing the operational parameters of multiple layers via abstracted layer parameters. The rate distortion factor - the difference between average peak signal-to-noise ratio (PSNR) of the encoded and displayed video stream - is disseminated from the video server to the CLO, which then distributes the values of the abstracted parameters to the corresponding protocol layers. The two processes represent additional communication and processing overhead, respectively. In [84], a cross-layer approach addresses QoS provisioning over IP-based Code Division Multiple Access (CDMA) networks. The idea is to have a centralized cross-layer scheduler placed at the base station which interacts with UE to exchange information about traffic, power level and others, and where video frames are compressed to batches of link layer packets according to their priority. This way, base stations are aware of maximum tolerable delays over the wireless link. [85] describes an adaptive streaming algorithm using 3GPP standard that improves significantly the quality of service in varying network conditions and monitors its performance using queuing methodologies, while [86] proposes a low complexity system for determining the optimal cross-layer strategy for wireless multimedia transmission based on classification. The authors show that significant improvements can be achieved using the proposed cross-layer techniques relying on classification, against optimized ad-hoc solutions, in particular in scenarios with high packet loss rates. [87] tackles both rate adaptation and resource allocation in order to maximize the sum of achievable rates while minimizing the distortion difference among multiple videos. The optimal algorithm relies on information exchange between the application and the Medium Access Control (MAC) layers, which independently process parameters from a single layer. Additionally, sub-optimal algorithms are proposed for reducing the solution complexity.

Several works specifically target CDN-based video delivery, whose service performance greatly depends on the server selection. [88] proposes a cooperative server selection scheme designed to maximize robustness to wireless-related changes thanks to the cooperation between the Content delivery system and its users. Similarly, [89] presents a video control plane which uses a global view of client and networks conditions for dynamically optimizing video delivery, aiming to provide high quality viewing experience in current unreliable delivery infrastructures. Based on measurement-driven extrapolation, it is shown that optimal CDN selection may improve buffering ratio by up to 2x in normal scenarios, and more than 10x under extreme scenarios.

In wireless broadcast services, the number of receivers and the average video quality of the received video may be maximized by adjusting physical and application layer's parameters, taking into account the characteristics of the video. In [90], a system supporting a multitude of transmission data rates using H.264 is proposed, aiming to achieve optimal compromise between maximum average received PSNR and minimum video broadcast service outage probability.

Other solutions address the energy efficiency issue. A cross-layer optimization framework targeting improvement of QoE and energy efficiency of mobile multimedia broadcast receivers is proposed in [91]. This joint optimization is achieved by grouping the users based on their device capabilities and estimated channel conditions, and broadcasting adaptive content to these groups; such content is obtained through optimal SVC source encoding parameters achieved by applying a novel game theory model. Energy savings result from using a (SVC) layer-aware time slicing approach during the transmission stage.

2.8 Concluding Remarks

We have dwelled in the distinct areas and concepts delimiting the boundaries of this Thesis' work. Optimal video support and continuous delivery demands the insufficiently explored interation of orthogonal notions such as multimedia characterístics, underlaying transport and access technologies. Their harmonized cooperation is of paramount importance for future mobile video services, with previous solutions merely tackling protocols or layers, instead of providing a generic approach to justify Operator's interest.

Chapter 3

Multicast mobility support with network-based and centralized mobility management

Given the previously identified issues resulting from mobility of multicast users, and the lack of unified solutions addressing them, this chapter focuses on achieving an integrated solution for the support of both multicast sources and receivers, leveraging IEEE 802.21 MIH and PMIPv6 protocols, as well as standard IP multicast routing.

3.1 Introduction

The mobile data boom is simultaneously a threat and an opportunity for mobile operators, allowing them to improve their role beyond a bit pipe of Internet access providers, while challenging current mobile networks capacity. With the scale of today's networks, transport efficiency is of paramount importance. Designed for efficient data transport, IP multicast might prove an essential mechanism to overcome such challenges. While its inclusion within operator networks found initial inertia due to issues like difficult service management, it has been incrementally incorporated for delivering services such as IPTV. With the definition of clearer use cases in mobile scenarios, its support within cellular networks is closer to real deployment with 3GPP's eMBMS [46]. The massive increase in user content production is expected to originate novel scenarios and services where not only the content subscribers, but also the multicast source [64] are on the move, also referred as Personal Broadcasting Services [7]. With the proliferation of devices such as Go-Pro cameras and wearables like Google glasses in the near future, a variety of scenarios such as real time journalism in warfare or natural catastrophes, and the live showcase of a locality or city during seasonal festivities to a group of subscribed users are to be expected[1]. As with unicast communications, IP multicast was not designed taking into account mobility scenarios, with the resulting issues depending on the role performed by the Mobile Node (MN) or the adopted communication model [53]. For instance, receiver mobility has local impact only, while source mobility affects all subscribed users. Such complexity has led to partial solutions, first in Mobile IPv6 (MIPv6) [54], and more recently in its network-based counter-part, Proxy Mobile IPv6 (PMIPv6) [19], which is being addressed in Multicast Mobility (MULTIMOB) WG. For instance, some solutions preserve IP multicast efficiency in mobility environments, but do not provide fast multicast mobility. Additionally, there's no reference on how to jointly support multicast mobile receivers and sources when deploying Protocol Independent Multicast (PIM) in Mobility Access Gateways (MAGs), although such is described for source mobility [63].

Adding to this, limited work was focused at the specificities of multicast mobility within heterogeneous environments. A switch between interfaces means buffering, address and general connectivity management challenges, and there is no holistic approach harmonically providing seamless IP multicast mobility, which limits its wide adoption in mobile networks. IEEE 802.21 Media Independent Handover (MIH) [28] is a popular technology proposed for mobility management in heterogeneous scenarios. It enhances and facilitates mobility procedures /e.g. network selection and HO) in heterogeneous access technologies by providing a framework able to 1) abstract the specificities of each link technology, and 2) exploit that abstraction to control and obtain information from such links within a geographical area. In this work, we follow two design goals: avoid multicast transport inefficiency driven from tunnel replication and minimize service disruption due to a host HO. In order to be able to meet these goals, in this chapter the usage of multicast routing is proposed as an alternative to typically considered Multicast Listener Discovery (MLD) Proxy [92], which is known to originate transport efficiency problems such as tunnel replication. Besides, the use and extension of IEEE 802.21 as the common multicast mobility enabler and facilitator is proposed, as opposed to solutions relying on a stack of redundant and costly mechanisms and protocols.

Chapter Contents

- **Section 3.2** presents a full-fledged architecture enabling mobility for multicast MNs in PMIPv6, which leverages on the interaction between IEEE 802.21 and multicast routing information for supporting transparent mobility for multicast MNs, either when acting as receivers - achieved via MIH-triggered multicast context transfer - or as sources -

[1]Products such as those from LiveU, for instance, enable reports to take advantage of the multiple available radios for broadcasting.

through proactive source path tree reconstruction.

- **Section 3.3** presents the aforementioned architecture operation for two main use cases: multicast source mobility and multicast receiver mobility, and describes involved signaling.
- **Section 3.4** shows a qualitative analysis of the architecture against a set of Challenges.
- **Section 3.5** presents the practical evaluation of the architecture over an experimental testbed.

3.2 Reference Architecture

The hereby proposed architecture, dubbed Media Independent Multicast Mobility Management (MI3M), aims for the seamless support of multicast mobility in PMIPv6. It leverages on IEEE 802.21, which was adapted in order to integrate the operation of both mobility and multicast management planes. For such, PMIPv6 entities were implemented as IEEE 802.21 MIH-Users. By crossing the already available link-related information with awareness to IP multicast operations, they are able to preemptively activate the necessary network- or host-side mechanisms for each of the two referred scenarios.

The architecture revolves around three core entities responsible for managing all mobility and multicast related decisions: the Connection Manager (CM), the Multicast Flow Manager (MFM) and the Multicast and Mobility Decision Entity (MMDE). The CM is introduced for providing the MN with awareness to mobility and enable it to preserve the session during horizontal or vertical HO - preventing packet loss due to L2 and L3 lack of synchronization. It will exchange information with each access network by means of MFM, which is responsible for timely executing the multicast context transfer during the MN HO. Through an integrated threshold or event-based framework, the target network will join the necessary multicast subscriptions before the HO process is complete. Finally, the MMDE resolves the problem of source mobility in PMIPv6 by synchronizing the multicast state of the LMA with the HO.

3.2.1 Entities

3.2.1.1 Connection Manager (CM)

The CM is located in the MN, and incorporates a Logical Interface (LI) function, which "denotes a mechanism that logically groups/ bonds several physical interfaces so they appear to the IP layer as a single interface" [65], acting as the interface between CM and the operating system. CM is responsible for managing available interfaces, activating them and requesting the activation of the required resources from the network during a mobility process. For such, it contains a list of subscribed groups associated with the corresponding sources. CM employs a Mobility Awareness Layer which is responsible for updating multicast service interfaces during mobility, abstracting the application from Layer 3 updates.

When receiving multicast traffic, CM will trigger network-side multicast context transfer procedures, informing the target MFM about its current subscriptions. As such, the MN is able to inform the target network about the multicast groups to which it is subscribed, as well as to propagate that information through several network entities, when required. For realizing the context transfer process, we extend the IEEE 802.21 MIH MN HO Candidate Query and MIH N2N HO Query Resources request messages with an additional Type-length-value (TLV) named Subscribed Multicast Group, as depicted in Figure 3.1.

When acting as multicast source, the CM operates the LI so that the current and target interface are managed during the HO with minimal loss. In case of an intra-technology and single interface HO, a potential strategy would be to take advantage of MAGs' buffers. In inter-technology HO though, the availability of two interfaces allows to take advantage of redundancy. Thus, the Logical Interface is used to apply a make-before-break approach and to deliberately broadcast the IP multicast traffic over the previous and new interfaces.

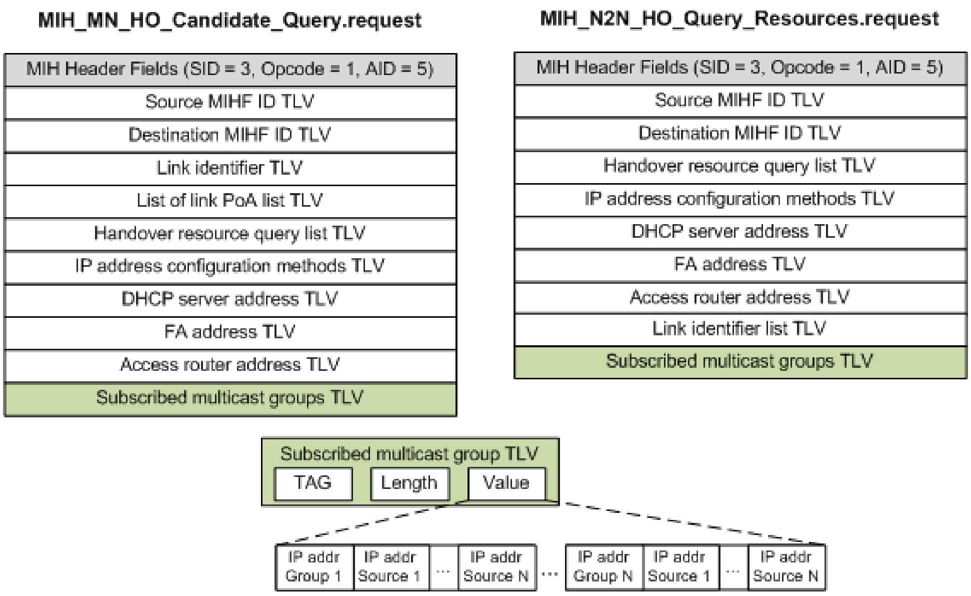


Figure 3.1: Proposed “Subscribed Multicast Group” TLV for IEEE 802.21

3.2.1.2 Multicast Flow Manager (MFM)

The MFM is located at the MAGs, and is a logical entity interfacing with the MR and MAG components. As a MIH-User, it triggers IP mobility operations, i.e. bi-directional tunnel updates, based on subscribed events such as the link quality degradation or availability of preferred PoAs. Enforcing IEEE 802.21 as a control plane, it executes multicast context transfer by transporting multicast subscriptions via the MIH Protocol signaling. It is informed by the CM about such subscriptions, executing the multicast context transfer process and the multicast subscription with added pro activeness. This way, the advantages of using a protocol like [93], such as multicast receivers' transparency to mobility signaling or the decoupling of multicast from unicast mobility protocols are done without introducing overhead due to Tunnel Convergence problem. Moreover, the followed approach discards the usage of a dedicated protocol for the desired mechanism, naturally embedding it over MIH signaling.

3.2.1.3 Multicast and Mobility Decision Entity (MMDE)

Located in the LMA, MMDE provides the full mobility system view by acting as a Point of Service. It has control over required operations for transparent source mobility. As identified in [94] in order to quickly recover the PIM (S,G) tree after HO, the multicast state maintenance should be synchronized with the unicast HO. To do so, MMDE anchors the DR function for multicast sources at the LMA. When the mobility of a multicast source is triggered, MMDE transfers the source path tree state from the previous tunnel to the new one, triggering the PIM (S,G) Join message to be sent towards the new tunnel. This mechanism applies to both ASM and SSM communications modes, whenever source specific paths are active.

For its operation, MMDE is required to store the list of multicast source addresses, in order to know when to trigger the referred process.

The resulting architecture is depicted in Figure 3.2. MRD6[2], ODTONE [3]and OPMIP[4] refer to the open-source software that were extended, adapted and deployed for instantiating the Multicast Routing, IEEE 802.21 and PMIPv6 mechanisms, respectively, and the underlying tools for the correct operation of CM, MFM and MMDE. The introduced entities are described in more detail in the following sub-section.

3.3 Architecture operation

The reference scenario, depicted in Figure 3.3, is one in which a mobile source is transmitting multicast traffic, tunneled by current MAG to its LMA. We focus on the case where both receiver(s) and source are registered at the same LMA, and on SSM communications, namely PIM-SSM, which faces bigger challenges in mobile scenarios. As such, the SPT is established between the source's MAG and the receiver's MAG, and processes related to PIM-SM stages

[2]http://fivebits.net/proj/mrd6/
[3]http://helios.av.it.pt/projects/opmip
[4]http://atnog.av.it.pt/odtone/

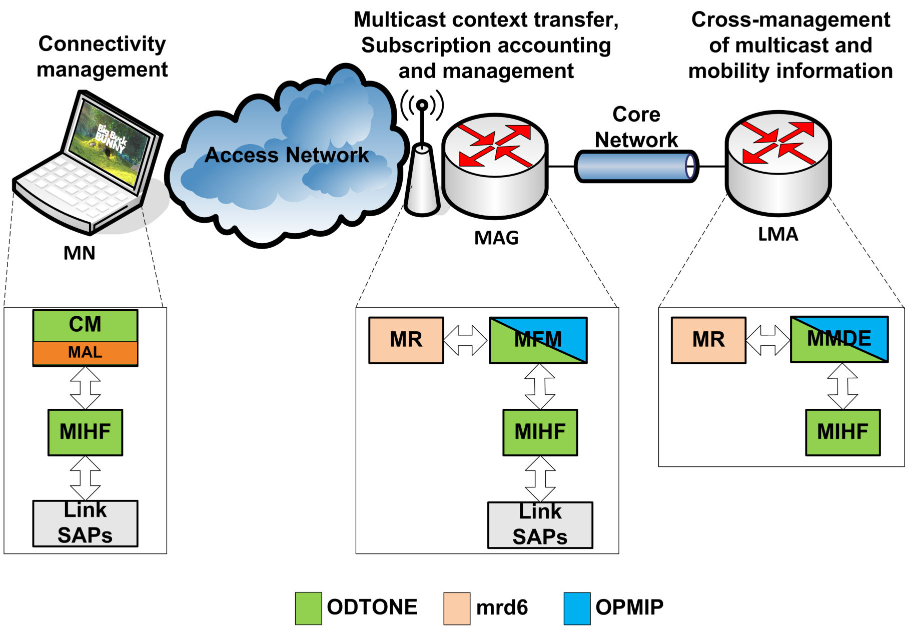


Figure 3.2: Proposed Architecture

1 and 2, namely PIM Register transmission and RPT establishment towards RP, are disregarded. Moreover, the single receiver is intended to represent a group of subscribers, except for the mobility event.

We consider two independent stages, which aim to fully demonstrate the provided features. In the first, we consider that the mobile source moves to a new MAG. In the second, it is the mobile receiver which moves from its initial MAG to another. Both scenarios occur independently of each other.

For both multicast source and receiver mobility, IEEE 802.21 provides information about nearby HO candidates, facilitates and optimizes the HO by providing technology independent commands and events able to trigger the L2 attachment on the MN side and necessary PMIPv6 and multicast procedures on the network side. Moreover, its coupling with multicast context transfer enables substantially reduces HO impact in delay-sensitive multicast applications.

3.3.1 Service initiation

The signaling involved in the mobile multicast service initiation using SSM model is depicted in Figure 3.4. The source MN is considered to be initially connected to a access network (e.g. LTE), but also within range of a more suitable or preferred one such as Wireless Local Area Network (WLAN) (typically considered for offloading scenarios), belonging to the same operator. The MN sends a IP multicast video stream through its serving MAG (1). After a receiver reports interest for group G from source S (2), a PIM (S,G) Join is sent towards the

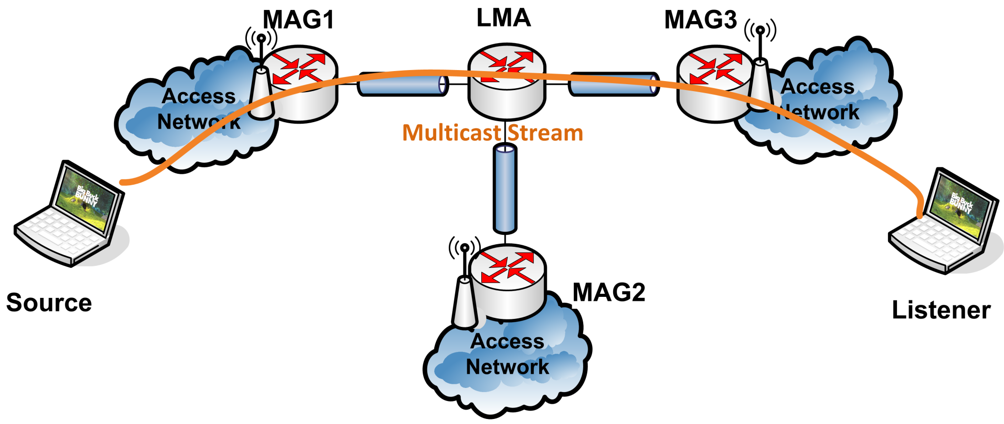


Figure 3.3: Reference Scenario

LMA (3), and the MMDE is triggered to update the existing MRIB, so that S is reachable via the tunnel interface. This way, the SPT is established between LMA and MAG1 (4), enabling multicast traffic to flow through the respective tunnel, and from LMA to other interested DRs (5). As will be seen, this enables the DR function for multicast sources to be anchored at the LMA.

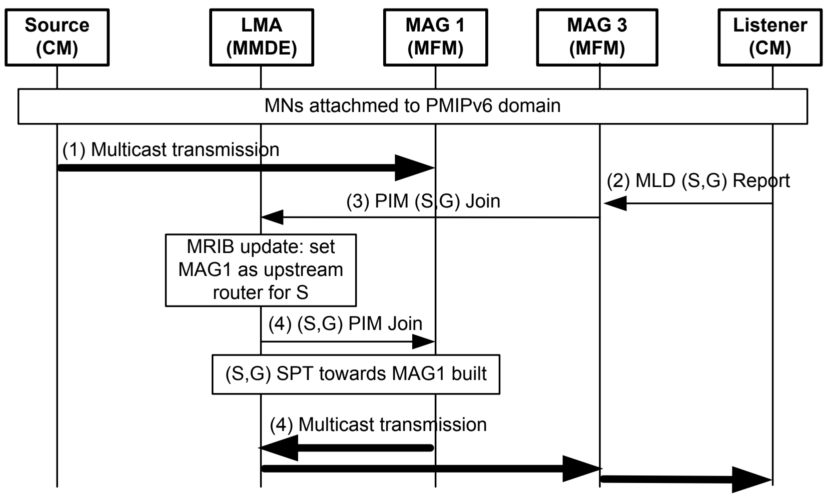


Figure 3.4: Multicast service initiation

3.3.2 Multicast source mobility

This section describes the involved signaling in the mobility event of a multicast source. The CM, in presence of the WLAN network, issues a MN-initiated HO towards its Point of Service (in this case MMDE), as shown in (2) from Figure 3.5. The HO trigger can have different origins, either by the MN (e.g. low received signal strength indication) or by the network side (e.g. detection of a non-optimal technology access for the current flow type). By receiving the CM's request, the MMDE will query the candidate networks about available resources (3) and based on the responses will reply to the CM with the best candidate networks. Upon the reception of this message, the CM requests MMDE the resource preparation in the selected target network (6), which is then responsible to communicate to MAG2's MFM the eventual arrival of a MN (7). At this stage, the MMDE is aware that one of its multicast sources is preparing a mobility process. When the MN receives the confirmation that the resources were successfully prepared by the target network, and verifies it corresponds to a distinct access technology (9), it activates the necessary procedures to transfer the upstream multicast flow to the new interface. Namely, the CM uses the LI function to merge the previous and new interfaces' buffers into a single logical buffer, through which it will send the multicast data. This way the packets will be broadcasted over the two interfaces for a very short period, minimizing upstream packet losses. Simultaneously, the MAG's PoA detects the attachment of the MN (10) and, since the chosen network belongs to the same PMIPv6 domain as the old connection, initiates the mobility tunnel update (11) with the LMA. Following PMIPv6 procedure, LMA updates the routing and Binding Cache Entry (BCE) information, and replies to MAG2 (14). MMDE updates the MRIB, triggering the prune of the connection with MAG1's MR (12) as well as the subscription of the multicast tree through the new MAG's MR (13). Upon reception of the Proxy Binding Acknowledgment (PBA), MAG2 sends a Router Advertisement (15) to the MN with the required information to configure its IP address on the new interface. At this point, the MN is still sending the video to the network via both interfaces, and notifies MMDE about the completion of its HO process (16). MMDE then forwards this notification towards the old serving network (17) in order to terminate its current binding and release the allocated resources. Lastly, when the MN is notified about the tear down of the resources, it detaches and turns off the network interface connected to the old serving network.

3.3.3 Multicast receiver mobility

The signaling pertaining to the receiver mobility scenario is represented in Figure 3.6. It is partially the same as the signaling for multicast source mobility, thus we will focus on the main differences between the two cases, which regard the multicast procedures. The same initial assumption applies, with MN initiating the service while connected to a network (e.g. 3G), and simultaneously within range of a different access technology network (such as WLAN), belonging to the same operator. The traffic is considered to flow through a similar path to the previous scenario. When the HO is triggered, the MMDE queries the candidate networks about the available resources (3), indicating the multicast records that the MN intends to subscribe. Each queried MFM stores the multicast information during a limited time interval. Thus, after the selection is made by the MMDE, MAG2's MFM is requested to

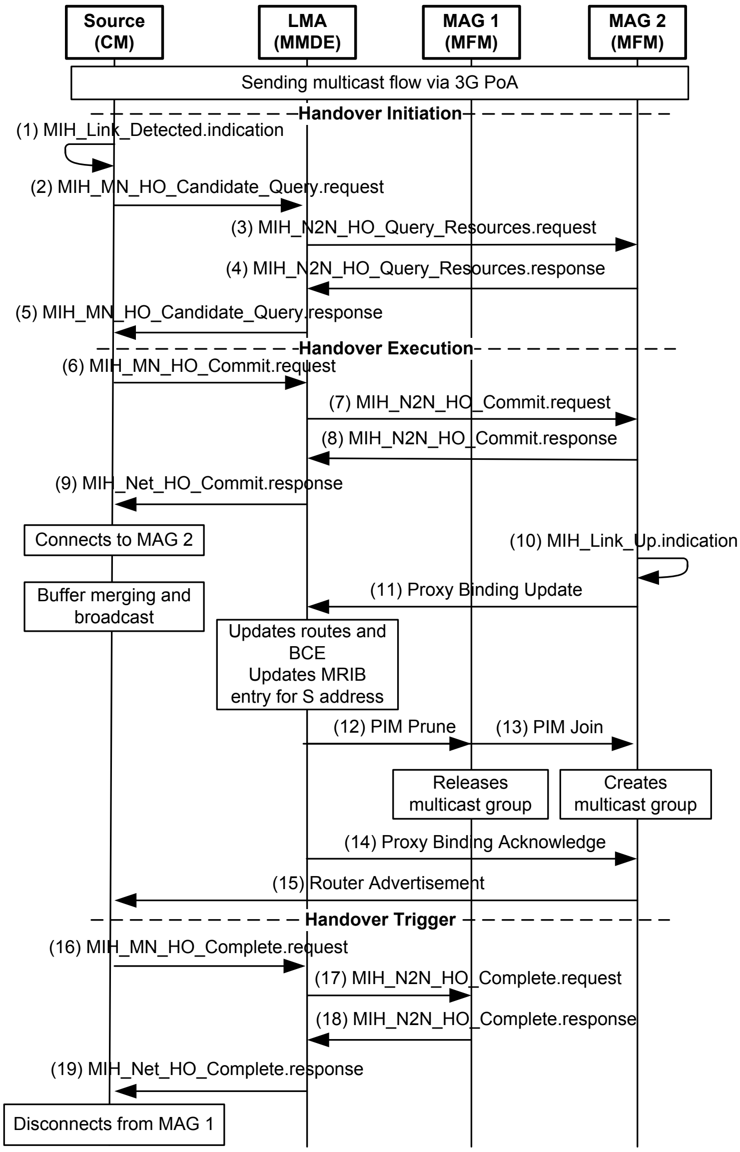


Figure 3.5: Source mobility signaling

prepare the resources for the attachment of the new MN, including the list of subscriptions (7). MAG2 then initiates the multicast join procedure for each multicast record that the MN was receiving, following the standard PIM operation (8). The remainder of the procedure is the same as in the previous scenario, with the exception that the MMDE does not need to update the multicast tree at the mobility event, until the notification of the old serving network to terminate its current binding and release the resources allocated the MN. Upon this notification, if the old serving network does not have more subscribers regarding one or more of the MN's subscriptions, the MFM will trigger their prune (17). Finally, when the MN is notified about the tear down of the resources, it detaches and turns off the network interface connected to the old serving network.

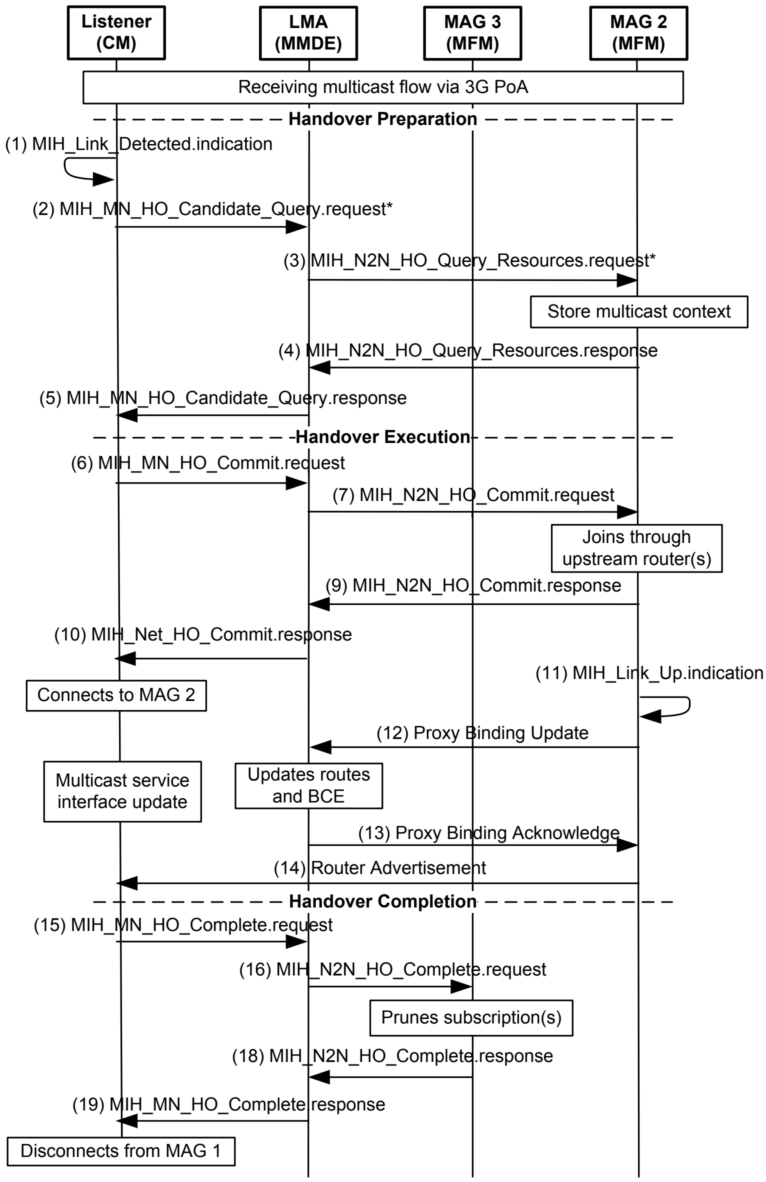


Figure 3.6: Receiver mobility signaling

3.4 Qualitative analysis

3.4.1 Identification of Challenges

We hereby summarize the list of challenges to be addressed by the proposed architecture, some of which were on previous work, and others are addressed and solved in this work for the first time, namely those which involve mobility between different interfaces. The challenges refer both to general issues associated to multimedia delivery using IP multicast under mobility, as well as to those specific to source or receiver mobility.

- Challenge 1 - Packet loss due to lack of synchronization between L2 and L3 mobility

One of the main issues with multicast source mobility is that multicast packets may be lost during HO in one of two cases: 1) they are transmitted to the previous MAG after the mobility tunnel is associated with the new MAG, or 2) they are transmitted to the new MAG before the tunnel and the mobility entry is updated.

- Challenge 2 - Service disruption due to receiver mobility

This is the typical problem resulting from multicast receivers mobility. When the user moves, it typically is required to resubscribe the content at the new network by means of IGMPv3/MLDv2 signaling, during which the service will be interrupted.

- Challenge 3 - Inability for receivers to cope with vertical HO

This problem is well detailed in section 3.3; basically, the service interface associated to a multicast application must be updated when moving between different access technologies. In time-strict applications, this must be done while simultaneously tackling Challenge 2 seamlessly.

- Challenge 4 - Assure Quality of Experience taking into account link dynamics

Some of the problems in heterogeneous wireless environments include the instability and varying characteristics of distinct access networks, which lead to varying - and unsatisfying - user experience. Even considering the availability of multiple interfaces, QoE assurance is not straight-forward.

- Challenge 5 - Overhead due to mobility tunneling

Solutions leveraging on mobility management protocols may lead the tunnel convergence problem, in which multiple copies of the same multicast stream reach the same MAG [57]. Overhead may also be originated by the first stage of PIM-SM, where all data is encapsulated in PIM Register messages.

3.4.2 Resolution of Challenges

This section provides an analysis of MI3M. First, the challenges it aims to overcome are enlisted, both general ones and those specific to source or receiver mobility. In the second part, the solution is evaluated as a whole in light of the properties presented in Section 4, for the sake of comparison with the other state of the art multicast context transfer proposals - thus, tackling pros and cons for its adoption as a fast multicast receiver mobility solution.

- Challenge 1: CM broadcasting operation during the multicast source HO assures the synchronization between the two switching processes, i.e. the one between upstream interfaces and the one between mobility tunnel. The transmission of redundant packets using the target interface assures that when the mobility tunnel is updated, the target MAG will forward the multicast traffic towards LMA. Besides, MMDE behavior removes the need to restart the PIM from Stage 1 at the target MAG, allowing the SPTs from the LMA to the receiver's DRs to be stable by simply updating the SPT between the LMA and the current MAG synchronously to the unicast mobility process.

- Challenge 2: MI3M leads to the proactive multicast tree update during the user HO, by applying a new flavour of multicast context transfer - by containing the multicast context at the MIH signaling itself, the need for a dedicated protocol is discarded.
- Challenge 3: The introduction of CM resolves the problem deriving from the application's mobility unawareness. Instead of giving mobility awareness to the application, it provides a fabric to the host's operating system, which will be responsible for updating the multicast service interface and thus assure always-on connectivity.
- Challenge 4: The integration of IEEE 802.21, with facilitates automatized network selection and attachment properties, enable users to roam between heterogeneous accesses while attached at the best possible network at each moment.
- Challenge 5: First, the double encapsulation due to the coupling of PIM Register and mobility tunnel when in Stage 1 is avoided, by prioritizing the rapid convergence to Stage 2 / 3 of PIM-SM; secondly, the tunnel convergence problem does not occur thanks to the application of multicast routing, more specifically the RPF mechanism, which provides MAG with the full view on all received subscriptions, avoiding the subscription of the same multicast stream through distinct interfaces.

As a summary, this approach enables the coupling of the HO latency speed of using a MLD Proxy (i.e. MBS [64], refer to section 3.1.2) with the transport efficiency achieved with PIM-SM (e.g. DMRS), while simultaneously avoiding multicast traffic replication, supported by RPF mechanism.

3.4.3 Comparison of fast mobility and multicast context transfer solutions

Several proposals have been presented for decreasing the time for transferring MN's multicast subscription information (i.e. multicast context) during a multicast receiver HO. However, the advantages and issues of each proposal have not been clearly evaluated, for instance preventing operators from doing an informed decision on which solution is suitable for their networks and needs. In this section, the major alternatives providing multicast context transfer-alike support in PMIPv6 are compared, according to a relevant set of parameters which we devise as requirements to be filled. The solution space comprises: Tuning the Behavior of the IGMP and MLD protocols in mobile environments [56] - denoted as TBIM; extension of the Context Transfer Protocol for multicast [61] - denoted as MCXTP; multicast extensions to Proxy Fast Mobile IPv6 - denoted as MPFMIP [60] and multicast HO optimization by Subscription Information through the LMA - denoted as SIAL [59].

3.4.3.1 Overview of multicast context transfer Techniques

TBIM - Tuning the Behavior of the IGMP and MLD protocols

This solution considers the configuration of several timers for mobile scenarios. Among those, the General Query Interval (QI) and the Query Response Interval (QRI) are of special interest during the HO: the former defines the time between General Queries sent by a Querier, while the latter tunes the interval in which MLD hosts must reply to a Query, translating the burstiness of MLD messages on a link. The reduction of QI and QRI enables the disruption time after HO to be reduced, but leads to the increase of signaling over the radio medium.

Additionally, the tuning of the Startup Query Interval (SQI) is considered, which reflects the rate at which a Querier sends MLD Queries after startup or after a new link is configured. The fact that PMIPv6 follows a point-to-point link model reduces signaling impact over the wireless media, and enables the optimal configuration of SQI to a value close to 0.

MCXTP - Multicast Context Transfer Protocol

This approach enhances the Context Transfer protocol [93] with a multicast subscription mobility option. As soon as a MN attaches to the new MAG, the context transfer process can take place between the previous and new MAGs.

MPFMIP - Multicast extension for Proxy Fast Mobile IPv6

This solution introduces extensions to both MIPv6 and PMIPv6's Fast Handover protocols. For the latter, the context exchange is done between the previous MAG and the new one, like MCXTP. Two possible HO modes are considered: predictive and reactive mode. The difference between the two modes is how the new MAG gains knowledge of the receiver's active multicast subscriptions. In the former, the previous MAG will detect the receiver's movement, and, after learning about the ongoing multicast subscriptions either by using the explicit tracking function or a general MLD Query, it will send the information to target MAG via an Handover Indication (HI). In the reactive HO, the new MAG gets the receiver's multicast subscriptions using the regular MLD process.

SIAL - Subscription Information through the LMA

This work proposes the multicast membership of the active receiver to be stored in the LMA, which acts as a multicast subscription manager, and thus relies on extensions to the standard PMIPv6 protocols. Similarly to SIAL, two HO modes exist: in the proactive mode, the previous MAG embeds the subscription information within the de-registering Proxy Binding Update (PBU) sent to the LMA; in the reactive mode, after receiving the PBU from the new MAG, LMA initiates a Subscription Query process with the previous MAG. In either case, the new MAG will be informed by LMA about the existing multicast subscription of the MN.

3.4.3.2 Solutions comparison

The different alternatives are now compared regarding a set of properties not exclusively related to the multicast HO latency - the major evaluation metric of a HO solution - but also general characteristics relevant from the network operator perspective. The considered properties affect different stakeholders: the application, the user, or, for the larger part, the network. The considered properties are the following, and are summarized in Table 3.1.

- **Proactiveness**

Proactiveness expresses the degree of proactiveness introduced by the context transfer mechanism, and is related to the trigger originating the process. A more proactive solution will typically translate into less HO latency. MCXTP can allow proactive subscription, depending on a proper mobility trigger. although, the trigger is always the attachment at the new MAG. For both MPFMIP and SIAL the subscription proactiveness is assured with predictive and proactive modes, respectively, where detachment provokes the subscription process. IEEE 802.21 trigger-oriented flexibility allow the context transfer to take place before the mobility protocol signaling is complete, enabling MI3M to act proactively.

- **Complexity**

Complexity expresses solutions' ease of deployment (e.g. simple protocol extensions or dependency on additional protocol). MCXTP implies running an additional protocol, Context Transfer Protocol. Both MPFMIP and SIAL require extensions to the base mobility protocols, while TBIM is the less complex approach, achieved by simple tuning of intervals and timers. MI3M implies running an additional protocol, IEEE 802.21; though, in most deployment scenarios it is expected to act as a common framework for overall HO (unicast and multicast) preparation, justifying its deployment.

- **Signaling overhead**

Signaling overhead represents the additional signaling cost induced by the solution. TBIM does not introduce any additional signaling, while MPFMIP and SIAL add limited signaling to the base PMIPv6 protocol. MCXTP adds the signaling overhead associated to the CXTP protocol signaling. MI3M adds the signaling overhead associated to the included new protocol suite. The exact values when applying Mi3M for source and receiver mobility are shown in sections 6.2.1 and 6.3.1, respectively.

- **Out-of-band**

This parameter expresses whether the context is transferred by using the mobility protocol signaling (in-band) or using another protocol (out-of-band). TBIM is the only approach exclusively relying on multicast protocols. Both MXCTP reuses a specific protocol, while MPFMIP and SIAL methods are in-band, introducing new messages aimed for the multicast context transfer within the mobility protocols: PFMIPv6 and FMIPv6, respectively. MI3M is out-of-band and incorporates multicast context within IEEE 802.21, which is assumed to be used as the control plane in heterogeneous environments, and thus will be involved in other HO-related processes.

- **Scalability**

This parameters expresses whether the solution scales well with large amounts of users. TBIM, MCXTP and MPFMIP have similar scalability, not implicitly depending on specific central entities for storing multicast context. On the other hand, SIAL centralizes part of the processing and storage overhead at the LMA, thus require careful planning before being deployed. MI3M does not centralize its context transfer mechanism on any specific entity, but adds responsibility to LMAs. Thus the solution is considered to scale to the increase in the number of users and network size in a similar way to that of the mobility management protocol.

- **L2 dependency**

This factor expresses the solution's dependency on host-side trigger and L2 specific capabilities (e.g. MN-ID transmission, radio particularities like framing etc.). The proposed solution leverages on the use of technology-agnostic signaling for HO initiation, and supports both network or host-side mobility trigger. SIAL operates by network detection of HO, whilst MPFMIP's predictive multicast HO relies on a report from the MN side. TBIM is independent of any L2-specific properties, with the SQI timer bootstrapped with the router's downstream link activation. Given that IEEE 802.21 is mainly designed towards mobility in heterogeneous environments, one of its main features is providing independence from access-specific mechanisms.

- **Unicast Synch**

This parameter expresses whether the solution allows multicast HO to be fully synchronized with unicast HO. SIAL is fully synchronized with unicast, being triggered by mobility protocol's registration and de-registration messages. MPFMIPv6 is integrated with FMIPv6 [95] and PFMIPv6 [23], which are the fast HO variations of MIPv6 and PMIPv6, respectively. MCXTP is not synchronized with mobility protocol, thus latency is possible (refer to [93], Appendix A), the same applying to TBIM. Unicast and multicast mobility processes will be triggered by the same MIH signaling message, i.e. MIH N2N HO Commit.request. Although, IP multicast mobility and unicast mobility are not tightly coupled.

- **Explicit tracking dependency**

Describes the need for Explicit Tracking mechanism in routers for the multicast HO process. In order to have per-user subscriptions knowledge, explicit tracking function is used. It allows routers to keep track of downstream multicast membership state created by downstream hosts, in order to save network resources and achieve fast leaves, and is required for preemptive multicast context transfer from the previous to the new network. With PMIPv6's adoption of point-to-point model, MAGs may extract membership status from forwarding states from the MN's-exclusive link, and organize this information for achieving explicit tracking. This applies to all solutions except TBIM, where routers do not exchange multicast context. MI3M does not depend from an Explicit Tracking function, as the context is always transferred from the MN, similarly to MLD but preemptively to the mobility process.

- **Independence of Multicast function**

These factors represents whether multicast context transfer can be achieved deploying MLD Proxy or MR in MAGs. All the solutions may be applied while using a MLD Proxy or a MR over the MAG, i.e. the context transfer process is independent of the multicast function. MAGs can apply a MLD Proxy function instead of full multicast routing stack capabilities. Concerning source mobility, the process is the same: the route towards MN from the LMA perspective is updated along with the tunnel creation. Additionally, the corresponding downstream interface of the MLD Proxy which has its upstream interface configured towards MN's LMA is setup.

3.5 Quantitative analysis

3.5.1 Experiment description

In order to evaluate the feasibility and performance of the proposed mechanisms, the two scenarios presented in the previous section were run in a physical testbed [13], located on the

Table 3.1: Context Transfer mechanisms' comparison

	TBIM	**MCXTP**	**MPFMIP**	**SIAL**	**MI3M**
Proactiveness	None	Trigger-dependent	On predictive mode	On proactive mode	High
Complexity	Low	High	Average	Average	High
Signaling overhead	None	Average	None	None	Average
Out-of-band signaling	No	Yes	Yes	Yes	Yes
Scalability	High	High	High	Average	High
Layer 2 dependency	No	Yes	On predictive mode	No	No
Unicast synch	No	No	Yes	Yes	Yes
Explicit Tracking dependency	No	Yes			No
Multicast role	Supports MR or MLD Proxy				

rooftop of the Instituto de Telecomunicações building in Aveiro, Portugal[5]. We integrated MRD6, OPMIP and ODTONE open-source softwares, implementing the required changes according to the proposals presented in Figure 3.2 As referred, ODTONE and OPMIP are open-source implementations of IEEE 802.21 and PMIPv6 protocols, respectively, while MRD6 is an open-source multicast routing daemon for IPv6, including support for multiple protocols, such as PIM-SM and MLDv2.

Besides the custom configurations of those softwares, several extensions were added as follows:

- Inclusion of a Subscribed Multicast Group MIH TLV in ODTONE;
- An interface between the MFM MIH-User and MRD6 for joining missing subscriptions and pruning outdated subscriptions;
- An interface between MMDE MIH-User and MRD6 for updating source's prefix route in MRIB.

The LI was implemented through a bonding mechanism. For the specific case of the multicast source, a broadcast strategy was followed, allowing the outgoing packets through all slave interfaces when active.

The two scenarios were deployed using the set of machines depicted in Figure 3.7, which includes two MNs, three MAGs and one LMA. The three MAGs (providing WLAN access to the MNs) and the LMA have IPv6 multicast routing capabilities provided by MRD6, besides providing their PMIPv6 functions. A sample video[6] is subscribed by the receiver using the

[5] http://amazing.atnog.av.it.pt

[6] http://www.bigbuckbunny.org/.

Table 3.2: Big Buck Bunny statistics

Video codec	H.264 - MPEG-4 AVC
Audio codec	MPEG ACC audio
Packet rate (packet/s)	96
Average throughput (Mbps)	1.2
Maximum throughput (Mbps)	4.1
Video size (MB)	86.6

SSM model and transmitted via a Real-time Transport Protocol (RTP) stream by using VLC clients. Information regarding the video is presented in Table 3.2. Finally, all machines have ODTONE installed, which provides MIH protocol communications. The nodes were setup so that control plane packets have higher priority, placing them at the top of the queue and overtaking the multicast RTP packets. This way, the potential delay in the control plane is controlled in case of high amounts of queued RTP packets. MAGs and MNs are setups in Amazing nodes.

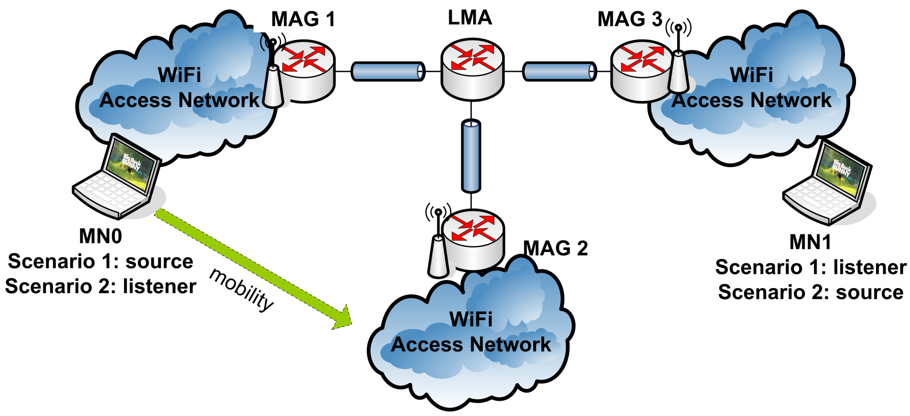


Figure 3.7: Testbed setup deployed in the AMazING testbed

3.5.2 QoS performance results

Each experiment was run 10 times, presenting average results with a 95% T-Student confidence interval. Although only WLAN access was used, we highlight that the media-independent mechanisms provided by IEEE 802.21 allow the network to be realized over different wireless and wired technologies, maintaining all architectural and signaling aspects. In both scenarios (i.e. source and receiver mobility), MN0 moves from MAG1 to MAG2 after 20s, and a video is streamed during 40s. In the first scenario, MN0 acts as source and MN1 acts as receiver, while in the second scenario the multicast roles are exchanged, with MN0 acting as receiver and MN1 as source. The available bandwidth over MAG1 is 1 Mbps, emulating a overloaded PoA, while MAG2 and MAG3 provide 11 Mbps. Several QoS metrics were collected, and the video aspect was described for both scenarios. Although the monitoring of both network transport (e.g. packet loss, round-trip-time) and video service QoS metrics (e.g. Signal to Noise Ratio) is crucial for assuring high performance in a video transport system, these are known to neglect the user perception, which might not be translated by excellent QoS values. For this reason, we extend the performance analysis with a QoE evaluation for the receiver mobility case, with a focus on the QoE properties before and after the HO.

3.5.2.1 Source mobility

In this experiment, we measured the performance and the impact on the video streaming resulting from the HO of the source (MN0) to a network with more available bandwidth, according to Figure 3.5. We collected QoS metrics from both network perspective (e.g. overhead) and from the receiver perspective (instantaneous, maximum and average throughput, packets per second and maximum burst).

Signaling Footprint

In order to evaluate the signaling footprint of the proposed architecture, we measure the amount of data exchanged between each entity. Results from Table 3.3 show that almost 68% of exchanged signaling data involves the MIH protocol, in particular between the MN and LMA (about 50% of all signaling). This was an expected result since the MN and the LMA exchange several messages to optimize the HO, negotiating the best candidate network and informing the network about the HO status. However, MIH is leveraged to enable the network to optimize the use of the available resources and the HO process, while lightly loading the network [96][97]. The remaining 30% and 20% of the MIH protocol signaling correspond to the resource querying and committing in the new network (MAG2), and to the resources' release of the old network (MAG1), respectively. Note that 200 bytes (about 24%) are related to MIH acknowledgments - given that it's a significant fraction, we argue that the research on methods aimed at its reduction, such as compression, are worth. We also highlight that no extra overhead is added to MIH signaling for the purpose of IP multicast source mobility, with MMDE being able to operate based on the mapping between the MN ID and its role as a multicast source. The other protocols require less information exchange, with about 15% of the signaling for PMIPv6, 12% for PIM and 5% for ICMPv6, corresponding to the base operation of each protocol. The exchanged information pertains to route and binding updates for PMIPv6, multicast subscription updates for PIM and address configuration for ICMPv6. Lastly, concerning the entities involvement, almost 93% of the exchanged signaling involves the LMA, with a MN involvement in about 40%, due to its participation in the candidate query and HO completion processes. Such involvement by the MN means that the network-based mobility management nature of PMIPv6 is partially kept.

	Entity				
	MN<->LMA	LMA<->MAG1	LMA<->MAG2	MN<->MAG2	Total
Protocol	**Signaling footprint**				
MIH	432	153	245	0	830
PMIPv6	0	96	96	0	192
PIM	0	70	70	0	140
ICMPv6	0	0	0	64	64
Total	432	319	411	64	1226

Table 3.3: Total signaling footprint in source mobility scenario (bytes)

Packet Loss and Throughput

Table 3.4 presents the video performance before and after the HO, comparing the packet loss and other metrics related with bandwidth usage. It was observed that before the HO approximately one in each four transmitted packets were lost, as a consequence of the limited available bandwidth in MAG1. The measured throughput values at the source's Logical Interface, and the previous and target MAG access interfaces are shown in Figure 3.7: the behavior prior to the HO can be observed in the time interval between 0 and 20s. When the required video bitrate goes beyond the available bandwidth, the throughput in MAG1 is not able to keep up. In turn, after the HO to a MAG with better bandwidth, the packet loss is significantly reduced.

	Before HO (MAG1)	**After HO (MAG2)**
Packet loss (%)	25.95 +- 0.05	0.20 +- 0.09
Packet rate (packet/s)	76.1 +- 0.3	104.2 +- 0.4
Avg throughput (Mbps)	0.9 +- 0	1.2 +- 0
Max throughput (Mbps)	1.1 +- 0	3.9 +- 0.2
Max burst (packet/100ms)	10 +- 0	35.3 +- 1.7

Table 3.4: Performnce comparison in source mobility

Handover Latency

As discussed, the HO latency for a multicast source has added importance, as it affects all interested receivers. Besides, in the considered scenarios, as soon as the HO is performed, the faster there is a service experience improvement, given the relocation to a better resource-wise network. Table 3.5 presents the time related to HO control plane, i.e. all stages of the HO signaling. A total HO duration of about 245 ms was observed. This value is influenced by the

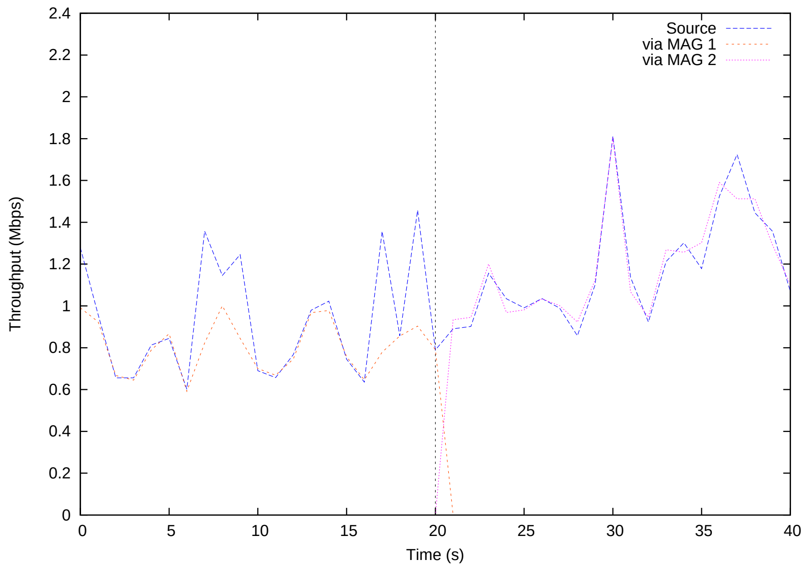


Figure 3.8: Throughput per interface (source mobility)

HO execution phase which, in turn, is highly affected by the L2 attachment procedures (an issue outside the scope of this work, and that could be improved using techniques as [98]. The presented architecture provides the means to flexibly pre-configure the thresholds for such decisions, mitigating the involved HO initiation problem. Concretely, IEEE 802.21 allows the CM to configure Signal to Noise Ratio thresholds to trigger the search for better connectivity solutions. Regarding the remaining phases, the query and reservation of resources in the network lasted approximately 14 ms, while the release of the resources from the old network took about 18 ms. The impact on the data plane in the context of source mobility is reflected in the HO latency, regarded as the link switching time, i.e. the time between the transmission of the last and first packets at the previous and new interface, respectively. Given the broadcasting strategy, this delay is associated to the PMIPv6 tunnel update operation, and independent of the wireless media; referring to the values of the used implementation [99] the PMIPv6 procedure (PBU, PBA, route setup) only takes about 1 ms, which translates into the continuous reception of the multicast stream at the LMA. Its additional task is, from a multicast router role, the update of its route towards the multicast source. For the sake of comparison, the two solutions in [64], MBS and DMRS, achieve an HO latency of 293 ms and 323 ms, respectively. It is not possible to do a direct comparison against MI3M though, as the referred solutions perform horizontal HOs.

Packet Duplication during Handover

As described previously, the adoption of a make-before-break / L2 broadcasting strategy during the HO leads the L2 connection to the target MAG to be established before the previous connection is lost, enabling the source to transmit through both network interfaces

Latency parameter	Total signaling time (ms)
HO preparation	13.78 +- 1.37
HO execution	245.09 +- 91.18
HO completion	18.37 +- 3.00
Total HO signaling	245.12 +- 91.19

Table 3.5: Handover signaling latencies in source mobility scenario

Duration (ms)	18.37 +- 3.00
Duplicate packets	1.7 +- 0.64
Duplicate packets (bytes)	2363 +- 890

Table 3.6: Properties of source's handover completion phase

during the HO. The broadcast strategy minimizes packet loss and brings the solution closer to a seamless HO, but also represents data overhead, which we analyze in this section. We claim such an approach is needed for premium services where multiple users could be affected due to a noticeable source HO, and that simply relying on network layer mechanisms (e.g. [64]) is not enough. The resulting duplication traffic is now inspected. The time interval that the source is sending through both interfaces is intrinsically related to the HO Complete phase. The source took about 18.37 ms to disconnect from the MAG1 after connecting to MAG2, resulting on the transmission of approximately 2360 duplicated bytes over the two interfaces (Table 3.6).

Video Aspect

Figure 3.9 illustrates the video quality before and after the mobility of the MN. Before the HO, the image suffered severe problems including blockiness and blurriness, as well a significant amount of artifacts. After the HO, the video maintains a good quality, with minimal artifacts. This can also be verified by analyzing the throughput values shown in Figure 3.10, where it is possible to observe that before the HO the receiver throughput cannot match the video bitrate sent by the source, but is then improved after the HO to the target MAG.

3.5.2.2 Receiver Mobility

Signaling Footprint

Similarly to the previous scenario, we measured the performance and the impact on the reception of the video stream resulting from the HO of the receiver to a network with more available bandwidth (MAG2). In order to evaluate the signaling footprint, we measure the amount of data exchanged between each entity. Results in Table 3.7 show that the

Figure 3.9: Video quality comparison (source mobility)

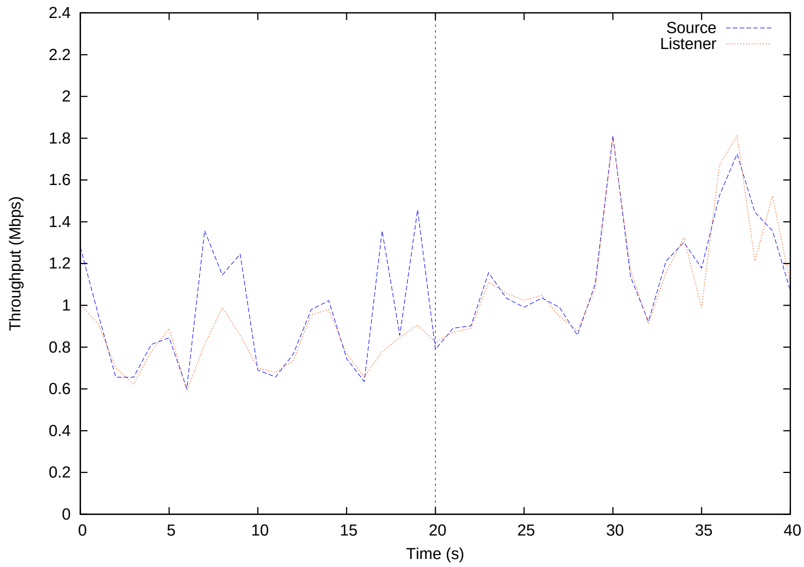


Figure 3.10: Source and receiver throughput comparison (source mobility)

amount of exchanged data is identical to the source mobility scenario, differing only on 20 extra bytes exchanged by the MIH protocol, which results from embedding the subscribed multicast channel in the two extended MIH messages. Notice that this extra overhead will vary according to the number of subscriptions transported, and the associated encoding; still, it is foreseeable that the portion of bytes necessary for IP multicast subscriptions is limited.

Packet Loss and Throughput

Table 3.8 presents the performance of video reception before and after the HO, and compares the packet loss and bandwidth-related metrics. In this case, before the HO there was a packet loss of approximately 9% as a consequence of the limited bandwidth available in MAG1. The throughput values at the source's Logical Interface, and the receiver's previous

	Involved entities				
	MN<->LMA	LMA<->MAG1	LMA<->MAG2	MN<->MAG2	Total
Protocol	**Signaling footprint**				
MIH	442	153	255	0	850
PMIPv6	0	96	96	0	192
PIM	0	70	70	0	140
ICMPv6	0	0	0	64	64
Total	442	319	421	64	1246

Table 3.7: Total signaling footprint in receiver mobility scenario (bytes)

and target MAG access interfaces are compared, as shown in Figure 3.11. Before the HO, there's a significant fluctuation between the source throughput and the throughput received at MAG1. Immediately after the HO, it's noticeable that MAG2 starts receiving the traffic at a very close throughput to the one transmitted by the source, as a consequence of having enough bandwidth for the multicast data.

	Before HO (MAG1)	After HO (MAG2)
Packet loss	8.83 +- 0.17	0.97 +- 0.88
Packet rate (packet/s)	81.9 +- 0.5	103.3 +- 0.9
Avg throughput (Mbps)	0.9 +- 0	1.2 +- 0.1
Max throughput (Mbps)	1.1 +- 0.1	4.0 +- 0.1
Max burst (packet/100ms)	10.3 +- 0.6	36.7 +- 0.8

Table 3.8: Performance comparison in receiver mobility

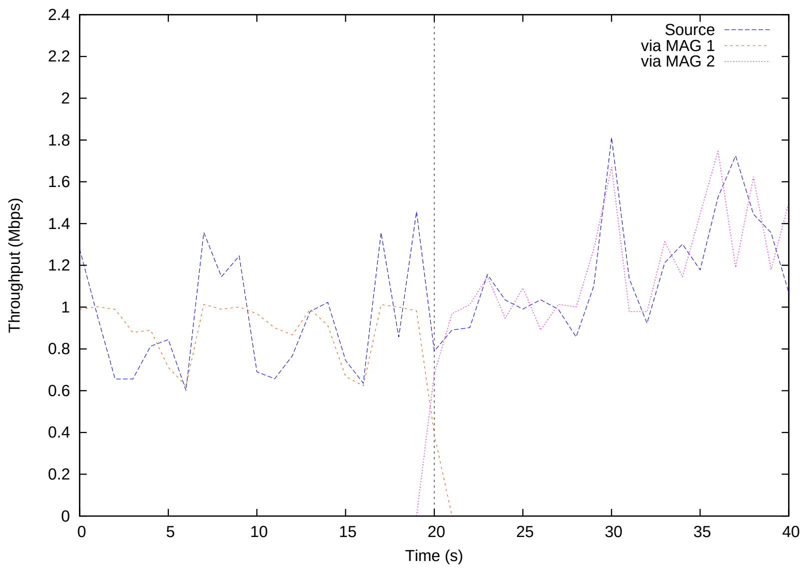


Figure 3.11: Throughput per interface (receiver mobility)

Handover Latency

The total time required for the HO operation is shown in Table 3.9. It is observed that the total HO signaling time is about 217 ms, a value which is again highly affected by the HO execution phase. In turn, the latter is dependent on the L2 procedures to attach to the new network. Regarding the remaining phases, the query and reservation of resources in the network took about 27 ms, while the release of the resources from the old network took about 18 ms. We verified that the HO preparation took more time than the source mobility case, explained by the storing process of the multicast context at each MAG. From the receiver perspective, HO latency is regarded as the time between the reception of the last and the first packet at the previous and new interface, respectively. It was observed that the MN is able to seamlessly continue the video stream reception, supported by two key facts: 1) the connection to the new PoA is done before detaching from the previous one; and 2) the multicast subscription by the new MAG is done before the detachment from MAG1. Unlike the source mobility scenario, in this case the mobility management / PMIPv6 process does not have direct impact in the HO latency; as soon as the new MAG interprets the request for new information, it will subscribe it towards the selected upstream multicast router. The only related solution which presents results is [61], achieving a service disruption time of 366 ms.

Latency parameter	**Total signaling time (ms)**
HO preparation	27.40 +- 6.68
HO execution	217.76 +- 88.96
HO completion	18.25 +- 1.66
Total HO signaling	217.17 +- 88.95

Table 3.9: Handover signaling latencies in receiver mobility scenario

Duration (ms)	18.25 +- 1.65
Duplicate packets	2.3 +- 0.64
Duplicate packets (bytes)	3197 +- 894

Table 3.10: Properties of receiver's handover completion phase

Packet Duplication during handover

The time that the MN is receiving through both links is correlated to the duration of the Handover Completion phase. We verified that the receiver took about 18.25 ms to disconnect from MAG1 after connecting to MAG2, which translates into the reception of approximately 3200 duplicated bytes (3.10). It is important to notice that the inclusion of multicast context transfer mechanism along with multi-linked reception is aimed at balanced duplication while assuring seamless HO; i.e., the sole application of a make-before-break strategy would either result in more significant duplication overhead or in noticeable packet loss due to HO.

Video Aspect

Similarly to the previous case, a comparison over the video quality before and after the mobility event was done. Two sample images, one before and the other after the HO, are shown in Figure 3.12. Once again there is a noticeable improvement in the video quality due to the HO to a PoA with higher available bandwidth, verified by analyzing the throughput values of both source and receiver (Figure 3.13). Before the HO the receiver throughput cannot match the source's, which is inverted after the HO. Since we are considering transmission of multimedia content, the quality perceived by the multicast receiver is subject to be impacted by slight changes in the end-to-end path between the source and the receiver. Namely, given that the two mobility scenarios were tested by switching the roles employed by MN0 and MN1, coupled with the fact that the uplink and downlink properties of WLAN (e.g. transmission rate) are asymmetric, there is an added factor for the performance variations between the source mobility and the receiver mobility scenarios.

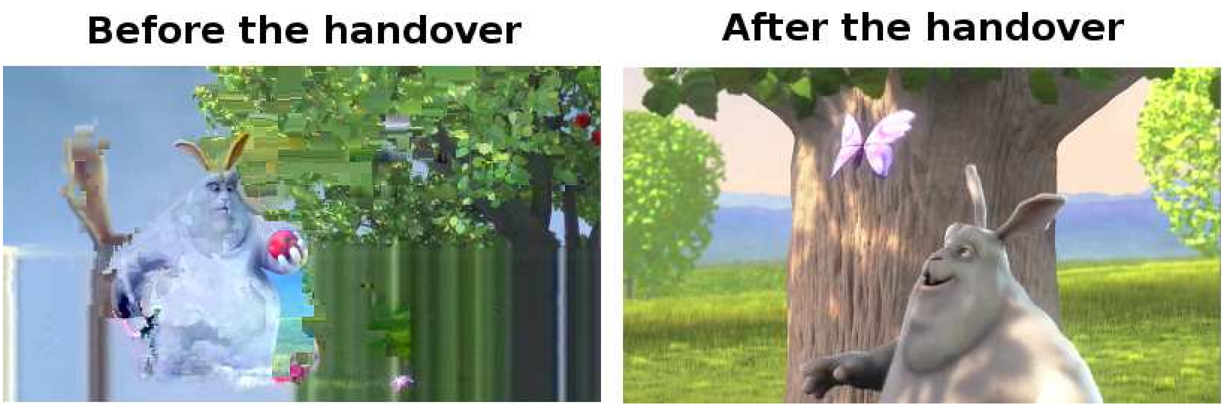


Figure 3.12: Video quality comparison (receiver mobility)

3.5.3 QoE Performance results

Video transmission has strict requirements, which can be evaluated not only based on network-side conditions but also considering user perception, i.e. QoE. As such, human-sensitive subjective video quality assessment techniques are also required for validation of network-side improvements. In this section, we do a full-reference QoE evaluation, comparing the received video quality against the reference video. We focus on perceptual-based objective video quality measurements, which use human vision system models to determine

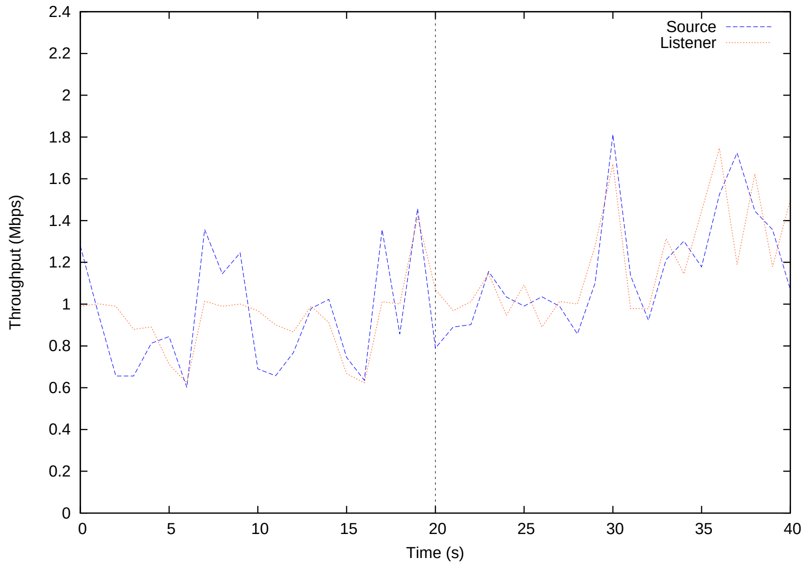


Figure 3.13: Source and receiver throughput comparison (receiver mobility)

the perceptual contrast between processed and reference videos; specifically, we obtain the Difference MOS (DMOS) by using Video Quality Analyzer from AccepTV[7]. We aimed to determine how much the received video quality was improved after realizing the receiver HO from the constrained network to the stable network, and to verify how quick the improvement occurred relatively to the HO instant. It is important to notice that these goals are distinct from the video aspect analysis done in sections 5.2.4 and 5.3.4. For this reason, and given the similar results obtained for the source mobility scenario, this evaluation was only done for the receiver mobility scenario. The results regarding DMOS are represented in figures 3.14 and 3.15. In the former, DMOS values are shown as a function of time, where it can be seen that the degradation of the received video was significantly reduced after the HO. The latter figure visually displays DMOS mapping for the varying bitrate throughout the video. It helps to show DMOS values before HO are closer to 100. After the HO, while the values do not reach the best quality (0), they are substantially improved.

3.6 Concluding Remarks

In this section, we have shown that the support of IP multicast in heterogeneous mobility scenarios demands coordination between hosts and network in order to be able to resume the session transmission/reception at the new access network, and to do so in a seamless way for the multicast receivers. A framework integrating multicast support over PMIPv6 with IEEE 802.21 optimizations, MI3M, was presented to address the latter issues, taking

[7]http://www.acceptv.com/

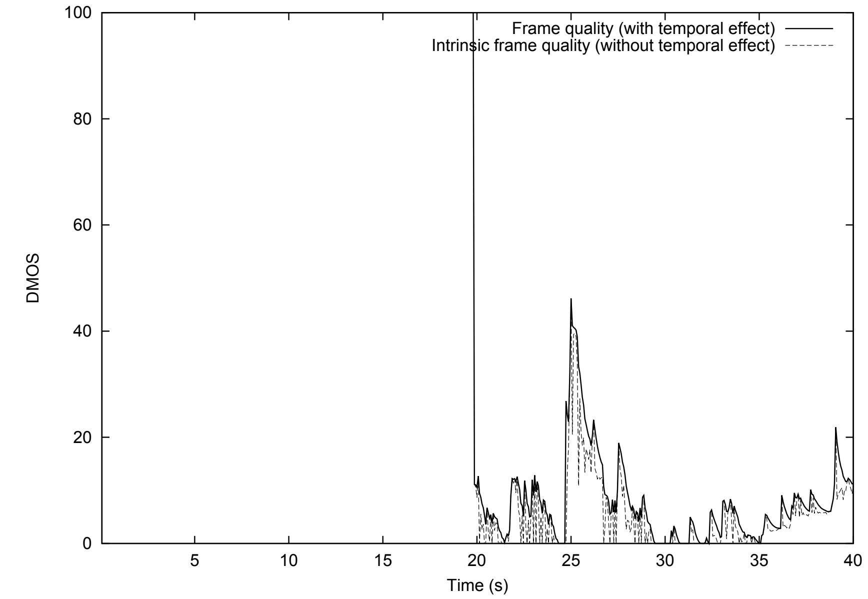


Figure 3.14: DMOS variation with time

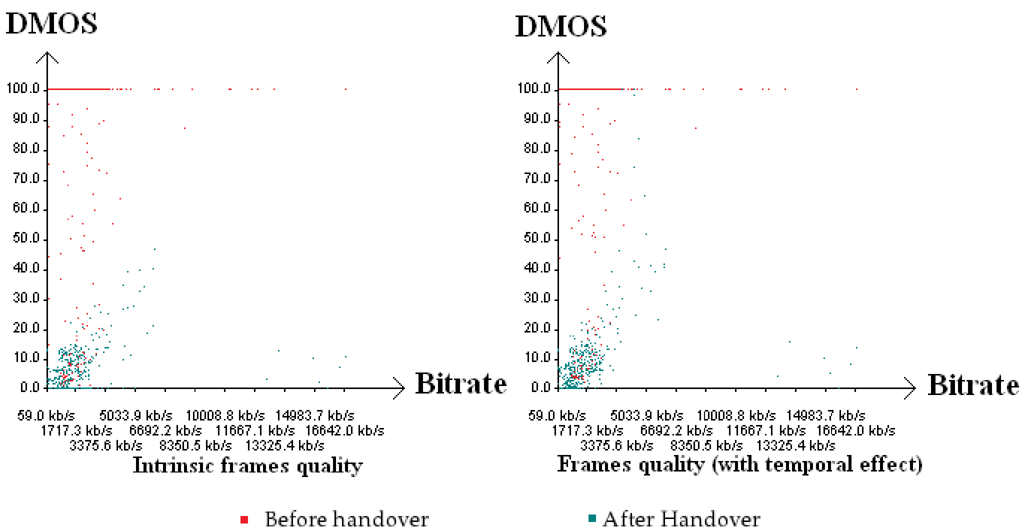


Figure 3.15: DMOS distribution per bitrate

into account the preservation of multicast nature of the data plane. Moreover, while each of the comprising technologies has its advantages and limitations, the framework was evaluated as a whole against a set of relevant characteristics, either reflecting how it addresses the proposed challenges and its suitability as a multicast context transfer solution. Besides, the comparison between distinct context transfer proposals enables operators and multicast service providers an informed design and deployment decision.

The results achieved through experimental evaluation show that the inclusion of IEEE 802.21 as a multicast mobility enabler is advantageous and enables quick HO at the cost of minimal additional signaling overhead. We consider the Network Localized Mobility Management nature of PMIP, and the unicast nature of the network stack fabric of operating systems require adaptations to enable vertical mobility in multicast environments, both from a source or receiver mobility perspective.

Chapter 4

Further evolving Mobile Architectures for improved video delivery

We have seen in previous chapter how legacy mobility management protocols, such as PMIPv6, were designed disregarding IP multicast support, leading to several limitations both for base support and in mobility scenarios. Moreover, a solution for IP multicast support in such environments was presented. Although, given the properties of centralized mobility management, such solution is not optimal, for instance due to dependence on centralized mobility anchoring, which required in multicast source mobility.

Thus, it was one of the goals of this Thesis to contribute to the development of future mobility management approaches, avoiding the repetition of the same mistakes in novel protocol generation. This chapter addresses the author's perspective and contribution to one major design tendency in mobile environments: Distributed Mobility Management.

4.1 Introduction

Mobile data traffic continues its tremendous growth path, leading the increase in investment cost by mobile operators for coping with the consequent and overwhelming capacity requirements. Mobile operators have been looking for intelligent ways of significantly reducing the risk of having CAPEX and OPEX costs outstrip data revenues, e.g. stretching their network capacity with data offloading technologies. Particularly focused on data offloading for mobile core networks, Local IP Access (LIPA) and Selective IP Traffic Offload (SIPTO) [100] mechanisms have been defined, guiding users to access locally available peering points via small or macro cells, thus freeing up mobile network capacity. Such data offloading solutions may alleviate the traffic burden over current hierarchically centralized mobile architecture, where all the traffic is directed to a centrally deployed mobility anchor, i.e. IP mobility anchor in an IP-based network and PGW in 3GPP's EPC. However, they do not eliminate scalability problems such as single points of failure, sub-optimal routing, and unnecessary use of mobility resources; therefore, such optimizations do not reach the disruptive degree to potentially cope with the eve-increasing traffic volume traversing mobile operators' core. Distributed mobility management (DMM) is an alternative to the previously presented centralized mobility management, characterized by the flattening of mobile networks and facilitated anchoring of traffic closer to the user's point of attachment. Following this trend, DMM allows a MN to employ multiple anchors, resulting in more optimal packet routing as MNs change point of attachment.

To realize these concepts, DMM faces various design issues such as mobility anchor selection at startup and on runtime, distribution degree of the mobility scheme or source address selection. Each design issue can be handled based on architectural aspects which are differently emphasized according to the involved player. So, it is critical to know the resulting performance impact of the different design options for each architectural aspect. There are prestigious articles dealing with DMM topic, but these have so far been focused at introducing conceptual scenarios, DMM protocol design proposals, or summarizing progresses from IETF (DMM WG) standardization perspective as initial efforts for flat-based mobile networking [37][101]][102][103]. Many design ideas have been proposed and evaluated from both the academic and industry communities, but each was focused at individual or a subset of the design issues, lacking a wider and more comprehensive perspective.

In this chapter, the main intended characteristics of DMM are presented, paving the way to the identification of the main design issues and the comprehension of their effects in a final DMM solution. Through this study, we aim to provide clearer and concise perspectives, ultimately leading to the understanding of the resulting effects of the multiple DMM design approaches, benefiting both vendors, operators as well as research engineers. Moreover, this chapter is complemented by presenting the main observations resulting from the evaluation of a generic network-based DMM approach over a specific network topology Such analysis resulted from work within MEDIEVAL Project led by Doctor Seil Jeon, while the qualitative analysis was carried out after the completion of the same project.

Chapter Contents

- **Section 4.2** describes the notions of Dynamic and Distributed mobility management, comparing them against previous centralized management approach.

- **Section 4.3** details the generic operation of DMM.

- **Section 4.4** presents a thorough analysis of several DMM proposals. The identified solutions are organized and classified, and then evaluated against a set of network-relevant factors, such as its efficiency, scalability, etc.

- **Section 4.5** presents the performance analysis of DMM against PMIPv6 in a concrete network topology. This work was not led by the author of this Thesis, but its observations and conclusions were a necessary step for the upcoming work from next chapters. For this reason, it is also included in the Thesis.

4.2 Dynamic and Distributed Mobility Management

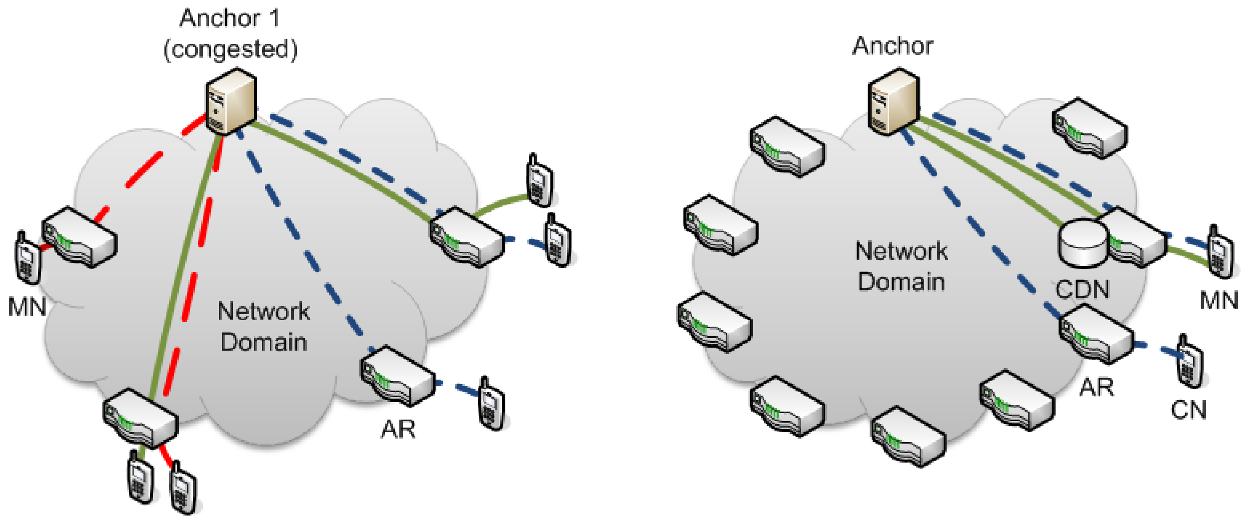
4.2.1 Motivation

IP mobility management solutions leverage on the concept of IP mobility anchor, an entity responsible for a set of crucial functions such as forwarding all traffic from and to associated MNs, advertising their public IP address and maintaining its mobility context (i.e. tracking their current location or AR). First solutions adopted a Centralized Mobility management (CMM) design, which leads to several limitations. Given the role employed by anchor entities, they are placed at the mobile core, leading to the convergence of all data at the network backhaul (Figure 4.1(a)). As data is forced to converge at these core anchors independently of the communication endpoints location, data is mostly routed through triangulation or "boomerang effect" if both endpoints are located in the same region (Figure 4.11(b)). The introduction of extensions for optimizing CMM operation, such as localized routing or run-time anchor assignment, have slightly alleviated non-optimality, but solutions still do not scale with the exponential increase in traffic and / or users due to the backhauling of all data traffic. From the control plane perspective, in order to maintain MN's mobility context, the mobility anchor must be involved in the mobility signaling every time an associated MN moves, originating signaling storms. Moreover, such signaling occurs independently of the applications the mobile user is running, which may for instance be text-based services like email or browsing. The aforementioned issues compromise the network core performance in envisioned scenarios; furthermore, while each of the planes (control and data) represents a limitation by itself, the combination of the two problems significantly increases the potential for failure. This motivates for a more significant reformulation of mobility management.

4.2.2 Desired Characteristics

DMM mainly leverages on the logical and topological distribution of mobility functions, and the dynamic mobility activation. The main advantages brought by these features are the following:

- Shorter, optimized routing paths



(a) Single point of failure and unbalanced network usage

(b) Long routing (CN and local content access)

Figure 4.1: Examples of limitations with centralized mobility management

The distribution of mobility functions closer to the network's edges and nearer to mobile users overcomes the need for all associated traffic to traverse the same restrict core-positioned network nodes, which tend to be off the direct communication routing path. With dynamic mobility management, mobility is activated according to the users' or application needs. Thus, if the MN does not move, end-to-end communication will traverse the optimal, non-anchored path. If the MN moves, the possibility for anchor selection enables improved routing. Overall, both data from static and mobile users will be able to traverse shorter and closer to optimal routing distances.

- Data load distribution

In CMM all traffic from a MN would traverse a single mobility anchor entity - refer to Figure 4.2(a) for PMIPv6 example – independently of its location, whereas in DMM traffic can be anchored at distinct and "optimal" anchor points - DMRs[1] - on a HNP or HoA granularity, as depicted in Figure 4.2(b). This way, the routing effort is naturally distributed among the mobile network, dissipating the backhaul congestion issue and the potential for single point of failures.

- Reduced and distributed mobility signaling

The distribution of mobility functions, such as mobility tunnel establishment and location management, leads to the implicit distribution of the signaling load involved in updating the mobility bindings among the multiple deployed mobility anchors. The mobility signaling is expected to be reduced in DMM, as it only refers to sessions requiring mobility.

[1]In this Thesis, both MAR and DMR terms refer to a generic mobility router in DMM environments

- Reliability and robustness

The processing cost distribution decreases potential for backhaul congestion and single point of failures, as mentioned. Additionally, in DMM control and data mobility management functions are no longer necessarily bound to the same entity. DMM proposals consider the possibility to isolate the control plane (i.e. mobility signaling) functions in dedicated dynamic databases, with such design feature adding robustness to mobility management solutions: for instance, the failure of a specific mobility anchor or AR will not compromise the integrity of location information of any MN, or the communications of other MNs besides those attached to the AR.

- Efficient mobility management operation

With the concept of "dynamic" mobility management, previously obligatory signaling load will not take place for prefixes associated with static users or mobile users whose applications are compatible with IP address modification, such as UDP-based ones, signifying an overall increase of efficiency in IP mobility management operations.

- Alignment with CDN-based contents access

The placement of anchors at the network's edge avoids triangular routing, facilitating efficient access to locally available resources such as CDNs, both at session initiation and after mobility. Thus, DMM paradigm is aligned with the "content everywhere" trend.

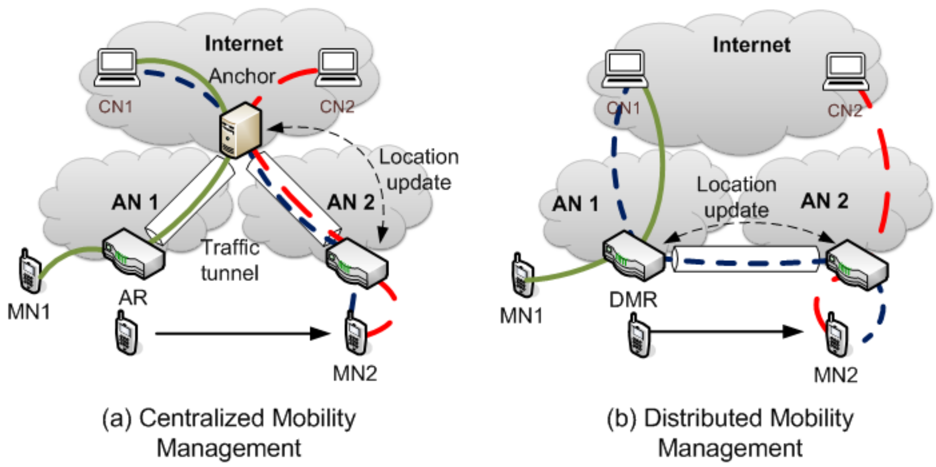


Figure 4.2: (a) Centralized Mobility Management vs (b) Distributed Mobility Management

4.3 Reference Architecture and Operation

Figure 4.3 shows packet routing operations for different IP flows initiated at different serving routers as a MN moves to DMM router 3 (DMR3) from DMR1, regardless of the design of the control plane. The description of the different stages in Figure 4.3 is as follows: (i) the MN performs IP communication with CN1 by regular IP routing, with no mobility activation; (ii) the MN is currently attached to DMR2; when DMR1 receives the packets destined to MN by CN1, the packets are forwarded to the MN through the established tunnel between DMR1 and DMR2; (iii) a new session is initiated by CN2, and IP packets belonging to the new session are exchanged between the MN and CN2 by regular IP routing, while CN1's session is anchored at DMR1 and the packets are forwarded through the established tunnel between DMR1 and DMR3; (iv) when the MN is attached to DMR3, each of existing sessions initiated by CN1 and CN2 is anchored at DMR1 and DMR2, and the packets destined to MN, sent by CN1 and CN2, are intercepted at each anchor and forwarded to DMR3, respectively.

Such a packet routing scheme has already been applied in several DMM protocol proposals with different designs of the control plane. In [104], it proposed a network-based mobility approach taking benefits of MN-unawareness into DMM, with the classification of fully/partially-distributed DMM. The concept of home/visited mobility anchors was applied in DMM protocol design [105]. P. Bertin *et al.* proposed a flat-oriented mobile architecture named dynamic mobility anchoring (DMA), which was evaluated against MIP in terms of HO latency [106], TCP segment delay and end-to-end packet delay [107]. Recently, it was evaluated against PMIPv6 in terms of packet delivery cost, signaling cost, and processing/tunneling costs in [108]. In [109] MIPv6-based DMM, PMIPv6-based DMM, and SIP were evaluated in terms of HO latency and packet loss.

Contemplating the previously proposed DMM studies from performance metric perspective, most efforts have been dedicated to show user-centric performance improvement with the relevant metrics as listed above. These metrics may be necessary to see the overall performances of proposed mobility protocols having the DMM strategy but are not sufficient to definitely address DMM-specific characteristics, which represents distributed workload throughout the network, released traffic intensity, and reduced link stress on the mobile backhaul, compared to the CMM.

4.4 Qualitative Analysis

4.4.1 Identification of Design Issues

When considering network computing background, distributed approaches (e.g. Peer-to-peer (P2P) or grid mechanisms) are typically more effective in terms of load distribution than centralized approaches such as the client/server model or clustering mechanism. On the other hand, the former require complex management which translates into additional cost. The same principle applies to distributed mobility management, where features such

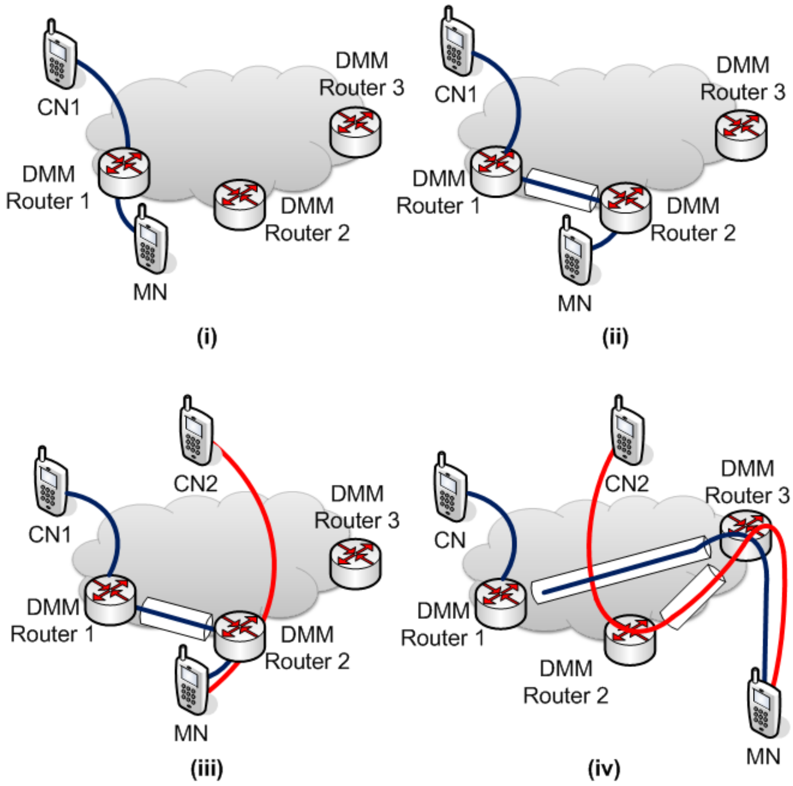


Figure 4.3: Distributed anchoring and dynamic mobility activation in flat-based IP mobile architectures

as the distribution of the location function represent a technical challenge, with very specific trade-offs which must be well researched.

First we'll address design issues related to the Data Plane and then those associated to the Control Plane.

4.4.1.1 Host involvement in mobility management

DMM solutions design can be categorized in function of the host involvement in the mobility management as network-based or host-based. This property has been given significant importance in the design of previous mobility management solutions, as seen in the shift from host-based (i.e. MIPv6) to a network-based approach (i.e. PMIPv6). As such, it needs to be considered in the DMM solution design, identifying which functions and information structures can be inherited and what should be differently applied or changed. In this section, we focus on the main properties derived from the host involvement in mobility signaling management and packet processing for a DMM architecture. Other properties related to the host involvement are mentioned and handled in the associated sections.

Network-based

In the network-based approach, the mobility management operation is provided by the network on behalf of the MN. That is, movement detection is supported by the network and the home emulation is ensured to make the MN unaware of its mobility. All the signaling procedures for retrieving and sending a message are executed between network entities, located in the remote (anchor function) and local networks (access function). Being associated with multiple mobility anchors introduces new design issues such as determining which address(es)/prefix(es) should be emulated, which are the respective anchoring DMRs. This mode is represented in Figure 4.4 (a).

Host-based

Unlike the network-based approach, the host-based approach requires modifications and intelligence from MNs enabling it to handle IP mobility, by managing the binding update lists associated with the established sessions and mobility resources in use for packet tunneling [110]. This solution is characterized by the MN's strong involvement in the mobility signaling, as depicted in Figure 4.4 (b).

In [37][101], a new host-based DMM design is proposed, which can be interpreted as a semi-host-based DMM approach. This approach introduces a mobility access router which interacts with the MN for the registration signaling - resembling host-based operation - while also taking care of binding update process - resembling network-based operation. The solution leverages on tunneling between anchoring and access routers, and employs a binding update process with two types of control signaling: Binding Update (BU)/Binding Acknowledgment (BA), used to deliver information regarding MN's context to the attached access router, while the extended BU and BA – ABU and ABA, respectively – are exchanged between DMRs for the tunnel establishment. It inherits the radio resource efficiency from the network-based approach by avoiding the extra packet encapsulation between access DMR and MN, while allowing the MN to initiate the control of mobility management, as shown in Figure 4.4 (c).

4.4.1.2 Distribution of Control Plane

DMM employs data plane distribution. On the other hand, the control plane design may be implemented in distinct ways, and is subject to the specifics of the deployment and access methods of the mobility database. This database tracks and maintains MN's mapping information between MN's ID and the IP address(es) or prefix(es), as well as information about the associated access and anchoring DMRs. So, control plane distribution is associated with the mobility database design, defining where and how the mapping information is

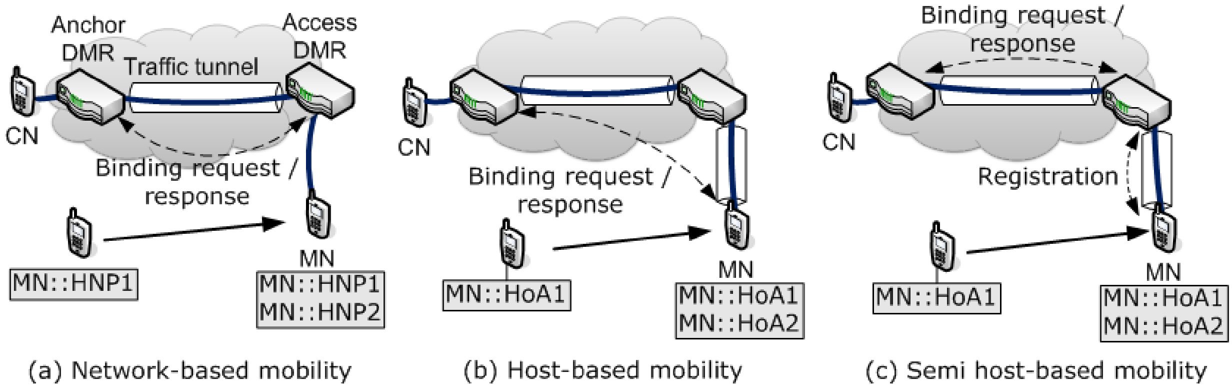


Figure 4.4: Different degrees of MN involvement in the mobility process

distributed. There are two control plane models: partially-distributed and fully-distributed.

Partially distributed

In partially-distributed model, there is a dedicated server to keep the MN's mapping information, and this information is accessed and updated by using the mobility management protocol. This server may take different roles which define its signaling involvement. One option is to use a single server model, where DMRs obtain the IP addresses of anchor DMR for the attached MNs [111]. This model is simple and easy to implement, but it may lead to a single point of failure as a consequence of mobility signaling storms. This issue may be minimized by adding redundancy to the database by deploying backup servers, taking over the primary mobility database in case of failure. Moreover, three roles of mobility database were proposed [112]: anchor locator, signaling relay, and anchor proxy. In relay mode, all the signaling messages pass through the mobility database synchronously, while in anchor locator mode, some signaling messages are directly delivered between previous anchor router and new anchor router. Focusing on fast handover, the use of anchor proxy was proposed for actively working as a mobility broker between two anchor routers.

Alternatively to a single database model, multiple database servers may be deployed into fragmented network domains, thus distributing the burden of querying and processing. The partially distributed approach is represented in Figure 4.5 (a).

Fully distributed

Unlike the partially-distributed mobility database model, a fully-distributed model does not rely on a dedicated server but attributes its forwarding path management and mapping responsibility to the deployed DMRs. Such a DMM deployment model has been sketched in [102], proposing alternative distributed and autonomous mechanisms such as peer-to-peer (P2P) to distribute and retrieve MNs' binding information into the distributed mobility agents, or the leveraging on an external mechanism such as IEEE 802.21 MIH, particularly for the network-based DMM approach [111]. Fully distributed approach is shown in Figure 4.5 (b).

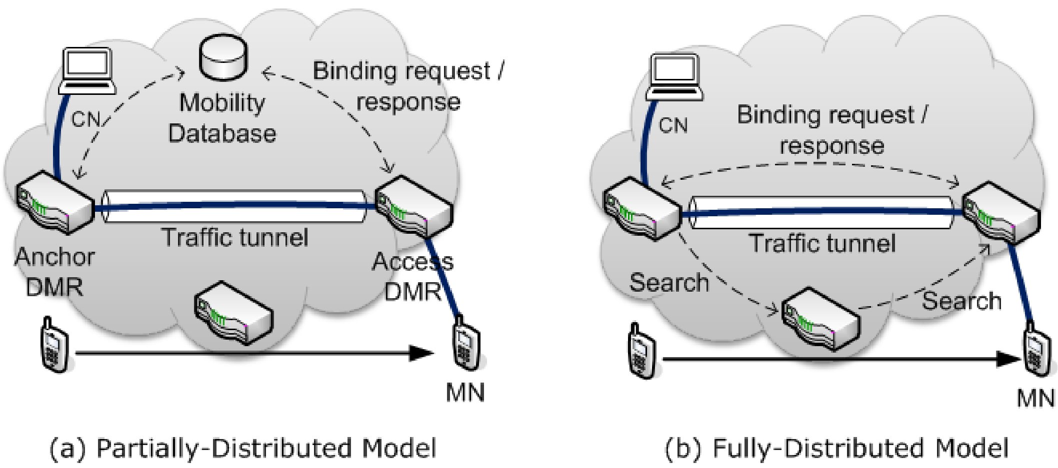


Figure 4.5: Design comparison of Mobility Database

4.4.1.3 Anchor Point selection

In anchor-based DMM solutions, the anchor point selection can be determined based on various decision aspects. We check the possible reference criteria and their general differences.

Distance-based

A simple method used for anchor point selection is distance-based selection, which assigns an anchor point depending on a reference host, such as MN's or CN's nearest available anchor, where the CN can be a server or corresponding mobile host. The main benefit of selecting MN's nearest anchor (depicted in Figure 4.6 (a)) is that it can be effective in accessing local contents from the anchor DMR to which the MN is attached [37][101][111]. Besides, the session connectivity after mobility can get quickly recovered due to relatively shorter signaling path between the anchor DMR and a new access DMR, where the MN is not too far from its anchor DMR. But this approach introduces suboptimal routing issue when considering not CN's location but MN's location only. The selection of CN's nearest anchor can be advantageous for avoiding potential suboptimal routing independently of MNs position after mobility [113] - depicted in Figure 4.6 (b). The opposite "anchorless" approach is represented in Figure 4.6 (c) for the sake of demonstration.

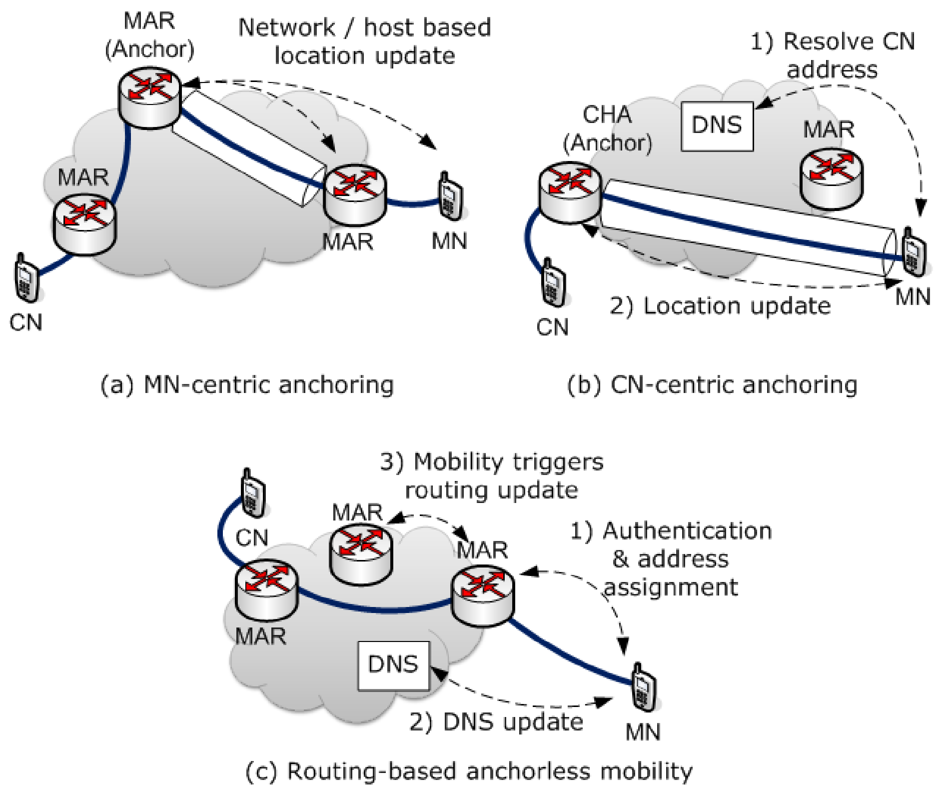


Figure 4.6: Data plane options for mobility management

Load-based

The anchor point can be selected taking into account the load condition at each anchor, since MNs' are not uniformly distributed, neither do they have similar traffic consumption patterns. Such load-balancing allows more reliable mobility management and contribute to better user and network performances.

Context-based

Anchor point selection can consider context such as MN's velocity (user context) and requested applications (application context). For supporting them in a network-based approach, the network needs to obtain necessary context implicitly, through intelligent monitoring mechanisms, while in the host-based approach, some explicit indication mechanisms will be required to deliver the MN's context to the related network entity.

4.4.1.4 Source IP address selection

In DMM scenarios, MNs will have its sessions with no mobility requirements being forwarded natively – using local assigned IP address -, while its sessions requiring IP address continuity will be anchored through one or more mobility anchors – using previously assigned IP addresses. Consequently, new applications will undergo a more complex IP address selection, as it directly influences the way that each session may be routed, resulting transport overhead, among others. For instance, applications leveraging CDNs achieve better performance when accessed with locally assigned IP addresses, while other applications requiring session continuity need anchoring-supported source IP addresses (e.g. Home Address in MIPv6 context).

Overall, applications may be bound to IP addresses either using default source address selection rules, or by explicitly reverting some of these rules (analogously to [12]), for instance by exposing their address type preferences. The two methods are now described.

Application-agnostic selection (default source address selection)

Applications which do not have any IP mobility management requirements may be assigned to a source IP address using default rules [114]. In this case, the IP address selection will be based on the IP address scope, the prefix which best-matches the destination IP address, and others. In DMM, MNs are expected to configure a local-only IP address at each DMR, which changes at each handover. Thus, in this approach a source IP address is configured mostly unaware to application-level mobility preferences, which goes against the observed demand for establishing network connection optimized to mobile applications.

Application-based selection

In order to effectively enable differentiated per-flow anchoring, so that Applications may flexibly use a specific – and not default – IP address, further extensions are necessary both in preference indication and address request mechanisms.

Work [115] classified applications' mobility management requirements according to two criteria: IP session continuity and IP address reachability. For assuring differentiated IP service according to such requirements, distinct IP address types are proposed: Nomadic – no mobility requirements -, Sustained – for IP session continuity or Fixed – for both IP session continuity and address reachability. Finally, in order to enable applications to express the required type of address, the same work proposes the extension of IPv6 socket API as indication mechanism, which is the method gathering more research interest. Based on the assumption of distinct IP address types, the API was further extended with a flag for enabling applications to explicitly request a new / fresher IP address [11]. Such flag is required for effectively enabling the differentiated per-flow anchoring.

To effectively convey the differentiated mobility management, and its request by the host to the network, both stateful and stateless IP address configuration approaches require modifications. Both are described below. In [116], and following stateless address autoconfiguration (SLAAC), the modification of Neighbor Discovery Protocol is proposed, namely by extending Router Advertisements. Access routers are provided with the means to deliver new IPv6 address prefix properties to MN: Remote and Local prefixes. In [117] the Prefix Information Option is extended with a mobility and a security property flag bit, as well as a 'class' describing the properties of the prefix. In order to enable stateful IP address configuration, extensions for DHCP will similarly be required, but presently there are no concrete proposals. Figure 4.7 depicts interactions pertaining to source IP address configuration and selection in DMM.

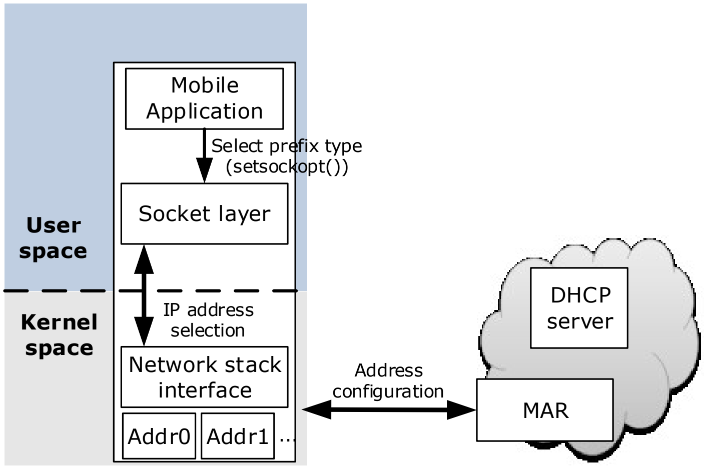


Figure 4.7: Overview of extensions for informed source address selection by hosts in DMM

4.4.2 Qualitative Evaluation of Design Issues

This section discusses qualitatively the impact of several design choices, as identified in the previous section.

4.4.2.1 Impact of control plane distribution

Partially-distributed model gives a reliable and realistic option for deploying and running mobility database by mobile operators. It enables easy installation and control over additional functionality that can enhance mobility performance. To facilitate the fully-distributed model, using a P2P strategy as a representative distributed autonomous mechanism is not a convincing approach for operators, due to its complexity and potentially unreliable mobility management support. In addition, a large volume of control signaling messages can be introduced when managing distributed mobility databases, with binding cache updates and synchronization between DMRs, representing potential broadcast storms in the network.

4.4.2.2 Impact of MN's involvement in mobility management

Removing the MN's involvement in mobility management eliminates the complexity of developing some needed functions and firmware updates as new features and functions are deployed. Also, it will contribute to the better user QoE since mobility signaling delay over the air is avoided and give more available control options to mobile operators. But a proper indication mechanism that represents user and application preferences over a DMM solution needs to be considered, meaning a fair amount of complexity and incremental energy drain.

4.4.2.3 Impact of Anchor Point selection

In the anchor point selection, there are many factors affecting the performance of terminal and network, depending on user and application contexts. If the MN is highly mobile, selecting an anchor close to the MN would cause frequent handover and consequently service disruption, which is critical to the application performance. In such a situation, keeping the routing optimality will be in a trade-off with signaling overhead. One useful idea might be deploying the anchor point at the core as well as at the edge, thus facilitating the optimal decision of the anchor point taking into consideration the terminal speed with additional factors.

4.4.2.4 Impact of source address selection

Source address selection is critical both for assuring user QoE performance, and for enabling penetration of specific services, e.g. for security in accessing enterprise applications. Thus, it is implicitly related to optimal routing – and reduction of its detouring -, as it effectively determines whether an anchored or local prefix is used, and whether the best anchor (e.g. closest) is used. Extensions to the socket API will enable application developers to make optimized and flexible applications, but the associated information that these applications store will be synonymous of additional complexity. Furthermore, the per-flow IP mobility management will result in additional signaling between the MN and the network, necessary for configuring and advertising the multiple prefixes.

Table 4.1 summarizes the impact of each design issue in a set of relevant performance factors.

4.5 Quantitative Analysis

Previous section provided an extensive qualitative analysis of several design options and strategies for delivering DMM solutions. Given the amount of possibilities, a generic DMM

	Control plane distribution	Host involvement	Anchor point selection	Source address selection
Routing distance	N/A	N/A	Depends on distance between anchor and MN – CN path	Increases with application-agnostic
Packet loss	NA	Increased in host-based approach	N/A	N/A
Complexity	Increased in fully-distributed approach	Increased in host-based approach	Depends on nature of input information	Increases in application-based method
Scalability	Increased in fully-distributed approach	N/A	N/A	Decreases with application-based
Transport and signaling efficiency	Decreased in fully-distributed approach	Decreased in host-based approach	Depends on nature of information used for decision	Decreases with application-based

Table 4.1: Impact of each design issue in different performance parameters

protocol, whose behavior is orthogonal to those options, needs to be evaluated in order to more clearly ascertain DMM's performance advantages.

Thus, one of the steps in this Thesis was the collaboration with the author in [2]. The article followed an event-driven simulator running on a given topology in order to measure several performance values based on mathematical analysis. Specifically, the evaluated performance metrics were (a) Packet delivery cost, representing how many routing hops have been traveled to deliver packets from a CN to the MN, (b) the ratio between anchored and non-anchored packet ratio, and (c) Traffic distribution ratio, representing how widely packets have been routed over the network. The main observations are now summarized.

4.5.1 Main observations

- Packet Delivery Cost

DMM achieved high overhead gains due to regular IP routing, as a consequence of the on-demand mobility, showing lower packet delivery cost than PMIPv6. In the initial mobility stages, the majority of the sessions are either generated in the current DMR - and thus are not anchored - or they are routed using relatively shorter routing hop distance, through tunneling between neighboring DMRs. In PMIPv6, the packets for all sessions are routed with a fixed, relatively long routing hop distance, due to the need to traverse the LMA. The packet delivery cost gain for distinct residence times was also analyzed, where for higher residence times, the HO count decreases and the anchored traffic traveling long routing distance also decreases.

- Packet Anchoring / Non-Anchoring Ratios

The analysis of the ratio between anchored and non-anchored packets allows to extrapolate the mobility protocol transport efficiency and the overhead introduced by mobility tunneling. It was verified that the the ratio of anchored packets in DMM is lower than what would be the result of dividing PMIPv6's ratio by the number of available anchors, which is easily explained by dynamic mobility anchoring, i.e. activation of IP mobility only after mobility. This confirms DMM's potential against PMIPv6 to increase transport efficiency in future mobile architectures.

The impact of residence time in the anchored / non-anchored packets ratio was also analyzed, where it was observed that in PMIPv6 the number of anchored and non-anchored packets is not significantly affected over the considered range of average residence time. As for DMM, this fraction of anchored packets decreases, due to the termination of mobile sessions, complemented by the initiation of new and natively-delivery sessions.. On the other hand, DMM's efficiency is significantly decreased in case of low average residence time, i.e. high mobility rate. Thus, the dimensions of the area covered by each DMR must be considered before deployment, taking into account the predicted user profiles and density.

- Traffic Distribution Ratio

Traffic distribution ratio translates how efficiently the mobile network resources is utilized. In the followed experiments, packets in PMIPv6 have traveled a more limited range of routing paths than in DMM. In was observed that DMM contributes to the distribution of overall network traffic as well as the reduction of network stress due to packet anchoring. And it can be foreseen that DMM would indirectly contribute to avoiding link traffic congestion and to improved data transmission speed.

4.6 Concluding Remarks

This chapter focused on Distributed Mobility Management, a promising new paradigm which leverages distributed mobility functions and dynamic mobility activation. Being DMM under intensive research at the time of the writing of this Thesis, a two-fold work was presented, consisting of qualitative and quantitative analysis. In the former, it was intended to organize identified design issues and their relationship to relevant network and service performance metrics. In the latter, and as part of a co-authored work, a network-based variation of DMM was evaluated against PMIPv6 under a concrete network topology, aiming to observe both its main performance benefits and disadvantages, by evaluating metrics such as packet delivery cost, anchored / non-anchored packets ratio, or traffic distribution ratio.

Chapter 5

IP multicast support with Network-based and Distributed Mobility Management

The challenges which emerge from the delivery of IP multicast under mobile environments, such as service disruption, traffic replication or complete service failure, were well detailed in previous chapters. In order to avoid these and other widely known problems from previous mobility management solutions, i.e. "History repetition", research is needed in pushing IP multicast considerations into initial DMM design solutions. This chapter is mainly dedicated to show the author's contribution towards the achievement of this goal.

5.1 Introduction

Services requiring IP session continuity, such as live video, are expected to proliferate in future years [1]. These cannot suffer from significant service disruption during HO, for the sake of user's QoE. As shown in previous chapter, legacy IP mobility protocols such as MIP or PMIP have applicability issues, motivating the work on DMM. With the potential relevance of this topic, work for determining how IP multicast may be supported in such environments must be developed. As an efficient data distribution method [118], advantages of IP multicast should be brought to future DMM solutions.

This gap motivated initial efforts towards the push of IP multicast into DMM design, as well as the identification of use cases taking advantage of MLD Proxy functions, and subsequent optimizations aimed at overcoming some of the identified limitations. As an alternative to MLD Proxy, the employment of IP multicast routing as an enabler was explored,

[1] Cisco's Visual Networking Index 2015

resulting in the definition of different schemes adapted to distinct requirements. Lastly, the support of multicast source mobility in DMM environments was researched.

A subset of these sections, namely sections 5.5 and 5.6, was the result of collaborative efforts under the framework of the MEDIEVAL project and other initiatives.

Chapter Contents

- **Section 5.2** introduces the support of IP multicast over DMM environments, by leveraging on MLD Proxies. Different solutions are defined, based on two features: mobility signaling proactiveness and degree of distribution.
- **Section 5.3** addresses a framework for improved subscription management when applying MLD Proxies, relying on a centrally located multicast channel management database.
- **Section 5.4** explores the utilization of multicast routing as a means to overcome the limitations of MLD Proxy over DMM scenarios, designs and mathematically evaluates such an architecture.
- **Section 5.5** addresses the issue of vertical IP multicast mobility in DMM, and provides an architecture for handling this issue.
- **Section 5.6** focuses on multicast source mobility over DMM, describing potential options.

5.2 MLD Proxy as Mobility Enabler

The utilization of MLD Proxy for delivering IP multicast in mobile scenarios has been considered due to its light properties, as seen in [57]. For this reason, it was considered as the starting point for the study of IP multicast support in DMM.

5.2.1 Reference Architecture

This section presents different schemes for multicast listener support in DMM solution. A generic architecture was considered, and design options were identified as a function of three properties, as follows.

Distribution Scheme

The mobility management solution may be fully distributed, where the MAG and the LMA roles are collapsed in a Mobility Access Router (MAR); or partly distributed, where the control plane is kept centralized at a single entity, referred to as Multicast Mobility Information Server (MMIS), and acting as mobility signaling relay [111]. In the latter, the data plane, which comprises routing function and data forwarding, is distributed throughout the MARs. MMIS is responsible for storing the prefix and PCoA's of each MN, identified by respective MN-IDs, and has an active role in the mobility decision.

Multicast Mobility Solution Proactivity

This is determined by whether the layer 3 HO is triggered by the previous MAR (pMAR) - proactive case - or the new MAR (nMAR) - reactive case. This impacts the speed of the mobility process execution, and as such, the potential disruption of the HO process. Considering that the requirement for IP mobility is associated to an already time-sensitive service such as live video, proactive schemes would benefit interactive services such as video-conferencing.

Subscription method after mobility

After mobility, the target multicast router may be informed of the user's multicast subscription in multiple ways. In this section, we describe only mobility solutions where multicast subscription is realized out-of-band relatively to the mobility protocol, namely using MLDv2 signaling. Other methods include multicast context transfer which has been proposed for DMM architectures in [119].

5.2.2 Scenarios Description

Figure 5.1 presents the proposed schemes, and the involved signaling. The description of the operations involved in each scheme, i.e. in terms of PMIPv6-based signaling and multicast joining procedure was done first for reactive schemes (named RP[2] and RF, for Reactive and Partially distributed scheme and Reactive and Fully distributed scheme, respectively) and then for proactive schemes (named PP and PF, for Proactive and Partially distributed scheme and Proactive and Fully distributed scheme, respectively).

5.2.2.1 Reactive Schemes

These schemes represent PMIPv6 common mobility procedure, in which the mobility tunnel update is triggered by the target access router (nMAR in this case). When using RP scheme, the mobility process is as follows: as the nMAR detects the MN's presence, it will send a regular PBU to the MMIS (Figure 5.1 (a)), which, after checking its database, forwards it

[2]RP will not stand for Rendezvous Point exclusively for the rest of this section

to the pMAR. The pMAR will then reply with a PBA to the MMIS, which forwards the message back to nMAR, completing the tunnel establishment. An alternative would be to send a PBA to both the MMIS and to nMAR, as seen in [120]. Afterward, nMAR will query MN for its multicast interests, and subscribe accordingly using an aggregated MLD Report.

In RF scheme (Figure 5.1 (b)), it is assumed that the nMAR knows the pMAR(s) address(es). The exact process for obtaining this information is out of scope of this section, but one possibility is to use 802.21 MIH protocol for dealing with the HO optimization [97]. The mobility process can be summarized as a PBU/PBA exchange between nMAR and pMAR, followed by the MN's multicast query, respective report, and finally the aggregated MLD Report.

5.2.2.2 Proactive Schemes

Proactive approaches are triggered by the pMAR, requiring prompt identification of the MN's detachment. Following a partially distributed approach (Figure 5.1 (c)), the pMAR signals the MMIS using a deregistration PBU. This message contains two new options: destination MAR option, which should contain the nMAR's address, and multicast subscription option, embedding the multicast context relative to the MN. When the pMAR doesn't know destination MAR, it will send the option empty, leaving the decision to MMIS. The MMIS will then reply with a PBA and send a HI containing the multicast context to the nMAR, which replies with a Handover Acknowledgment (HAck) and then subscribes the missing multicast channels.

As for PF scheme (Figure 5.1 (d)), the mobility process is initialized with an HI containing the multicast subscription context, and will be replied with a HAck in case of successful tunnel establishment by the nMAR, followed by the subscription process. A requirement for all schemes to work properly is that all MARs implement explicit membership tracking function [121].

5.2.3 Performance Evaluation

In this section, we analyze the performance of the different approaches in terms of service disruption latency during HO. An analytical model employed in [122] was followed.

5.2.3.1 Reference Model

Figure 5.2 shows the reference network topology for performance analysis. The delay factors consisting of total delay are defined as follows:

- t_{HW}: the delay between the mobile host and the wireless access network.

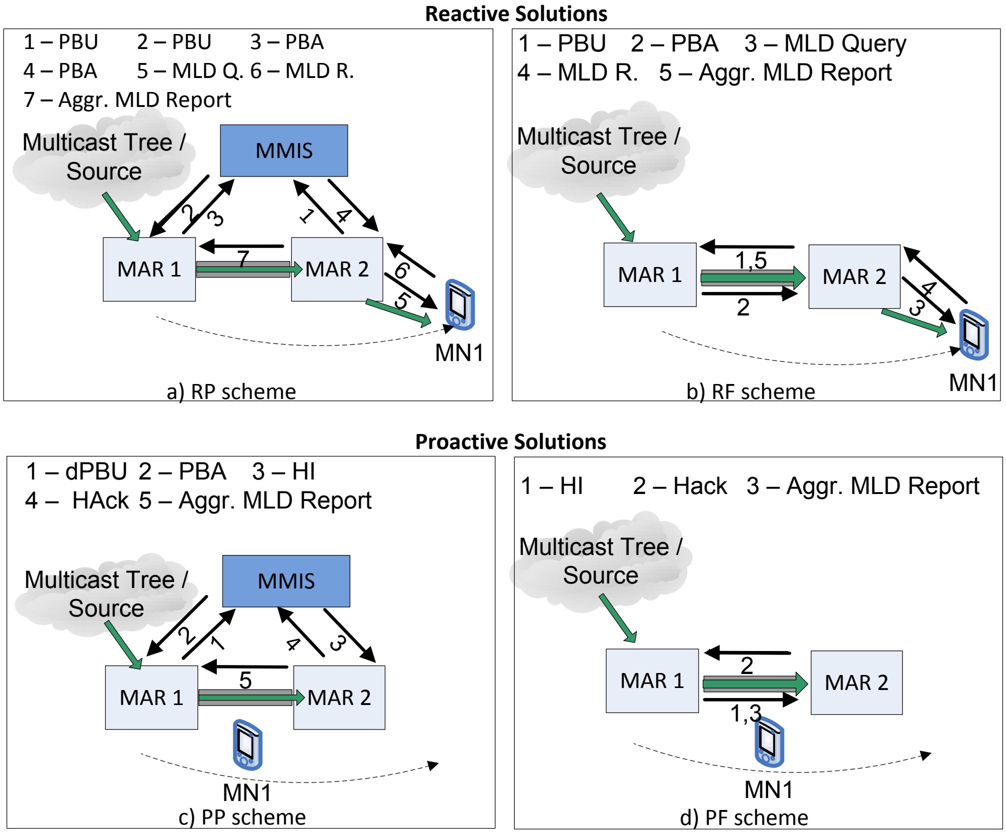


Figure 5.1: Schemes for multicast support in DMM using MLD Proxy

- t_{WM}: the delay between the wireless access network and the MAR.
- t_{MM}: the delay for MAR-MAR tunnel traversal.
- t_{MS}: the delay between MAR and MMIS.
- t_{MI}: the delay between nMAR and the IP multicast tree (PIM Join process). Because all solutions take advantage of the MLD Proxy tunnel towards pMAR, no subscription is required towards the IP multicast infrastructure. As such, t_{MI} is null.

The delay associated with the processing of the messages (e.g. nMAR address retrieval on HI reception) is included in the total value of each variable. Only intra-domain HOs are considered, i.e. between MARs of a same domain, covered by the network-based DMM protocol.

5.2.3.2 Service Disruption Time Analysis

The service disruption time is a relevant user-based metric for HO performance evaluation, as it translates the time slot in which no packet arrives to the MN. The base considerations for

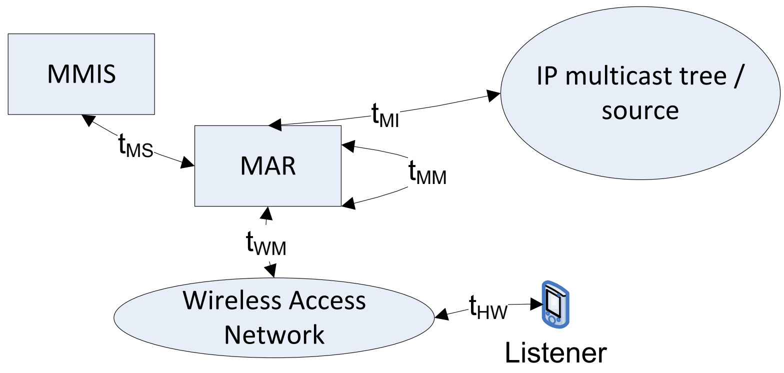


Figure 5.2: Reference network topology

its calculus are as follows. The multicast session duration is t_S , and follows the exponential disruption with factor λ_s. A MN stays in a subnet for t_c, following exponential disruption with value λ_C , so average session duration E(t_S) = 1 / λ_S and average time per subnet is E(t_C) =1/λ_C. The performance of DMM multicast schemes considers the HO frequency as well as the session activity. In this sense, we use a session-to-mobility ratio (SMR) that is widely used for performance evaluation in mobile networks. SMR represents the ratio between the session arrival rate and the HO rate, and is defined as ρ = E(t_C) / E(t_S) =λ_S / λ_C. Intra-domain HO probability is defined as ρ HO = 1 / (1 + ρ) in the literature [122]. The average multicast disruption time for intra-domain HO is computed as T = D × ρ_{HO}, being D the average value (i.e. considering both inter and intra-HOs).

The service disruption time is defined as the total time taken to complete all the signaling procedures for IP HO, multicast subscription, and transmission time of first packet from pMAR to the MN.

$$D_{TOTAL} = D_{HO} + D_{Join} + D_{Delivery} \tag{5.1}$$

It is considered the traffic is always received via the tunnel because the upstream interface of MLD Proxy is set towards the anchor (pMAR). The packet delivery time is given by:

$$D_{DELIVERY} = t_{MM} + t_{WM} + t_{HW} \tag{5.2}$$

Figure 5.3 shows the signaling procedures between entities for each scheme.

Reactive and Partially Distributed (RP) Scheme

In this scheme, the disruption as consequence of the HO is due to the signaling exchanged between nMAR, pMAR and MMIS. Besides, the time for joining the multicast tree is due to the MLD Query and Reports (2× (t_{HW} +t_{WM})+ t_{MM}), plus the Aggregated MLD Report

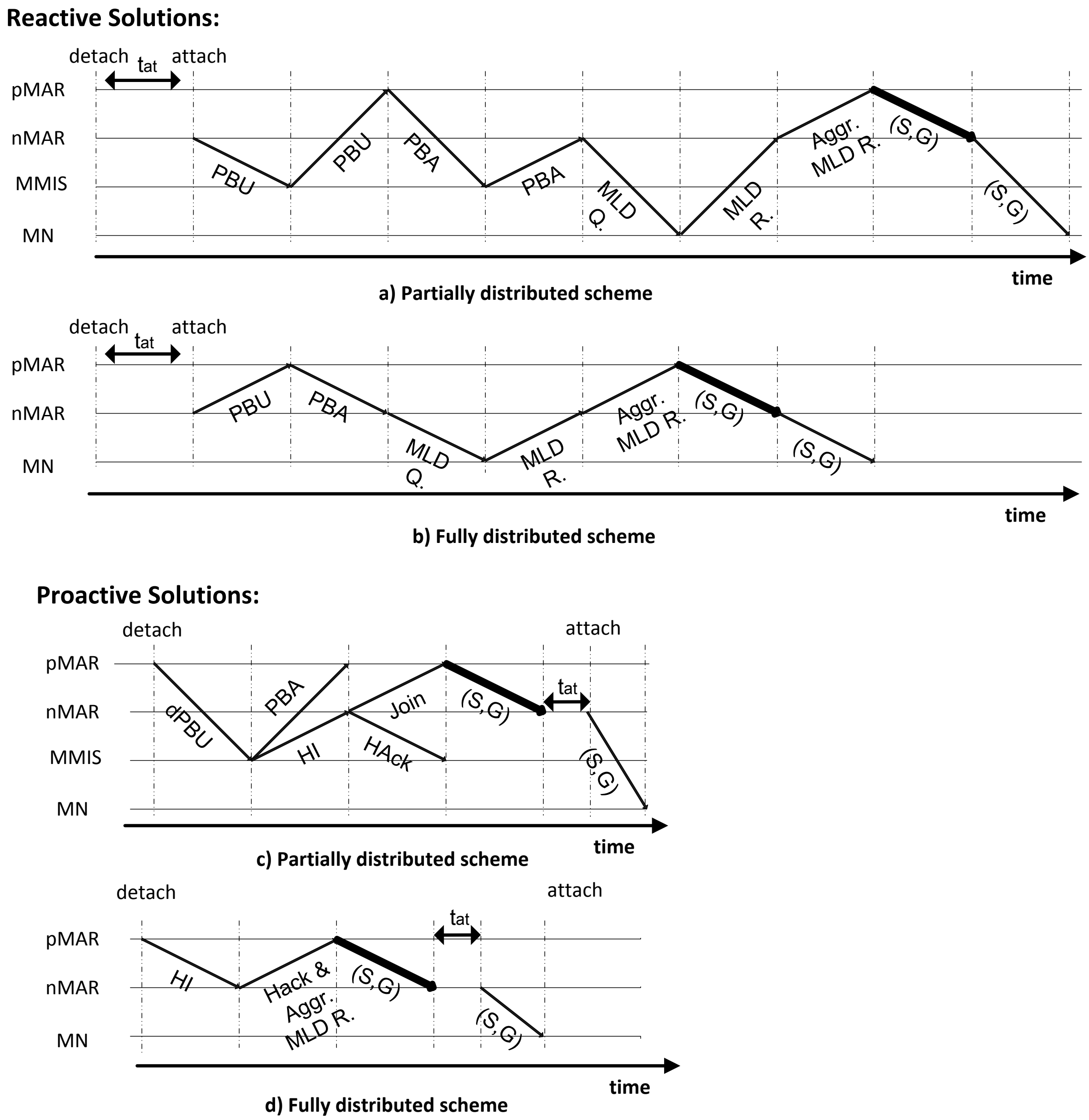


Figure 5.3: Signaling process for the different schemes

from the MLD Proxy. As such, the total disruption time is obtained as:

$$D_{TOTAL_{RP}} = 2t_{MM} + 4t_{MS} + 3t_{HW} + 3t_{WM} \tag{5.3}$$

Reactive and Fully Distributed Scheme

For this case, the largest part of messages is transferred between nMAR and pMAR. Moreover, D_{JOIN} is the same for both reactive schemes, which results in:

$$D_{TOTAL_{RF}} = 4t_{MM} + 3t_{HW} + 3t_{WM} \tag{5.4}$$

Proactive and Partially Distributed Scheme

In the current case the first message is towards the MMIS, being the reply from MMIS to nMAR (representing t_{MS}). Besides, for proactive schemes D_{JOIN} is simply the aggregated MLD Report, so:

$$D_{TOTAL_{PP}} = 2t_{MS} + 2t_{MM} + t_{WM} + t_{HW} \tag{5.5}$$

Reactive and Fully distributed Scheme

In this case, the HO time is simply the signaling for the tunnel establishment – HI and Hack -, being the subscription done immediately after. Therefore, the total service disruption time is:

$$D_{TOTAL_{PF}} = 3t_{MM} + t_{WM} + t_{HW} \tag{5.6}$$

5.2.4 Numerical Results

This section presents numerical results of average service disruption time. The experiments were done by varying: A) the multicast session to mobility ratio, for determining which schemes provide better resilience to mobility; B) MN's L2 re-attachment time t_{at} and C) the packet forwarding delay for traversing the tunnel between the MARs.

The following values were used for the numerical results. t_{HW} , t_{WM} , t_{MM} and t_{MS} are assumed to be 5ms, 2ms, 2ms, and 3ms, respectively, according to the literature [122]. And the values for t_{MM} and t_{MS} correspond to MAG-LMA and MAG-AAA delays, respectively.

5.2.4.1 Multicast Session to Mobility Ratio Impact

Figure 5.4 shows average service disruption time when SMR is varying. For values of SMR below 100, when the users have a more mobile profile, it can be seen that proactive schemes clearly provide a more robust service. For the sake of useful graphical information, SMR values above 30 are not shown in the figure.

5.2.4.2 Mobile Host re-attachment Time

In the previous results, MN re-attachment time (t_{at})was ignored, but this section intends to evaluate how the different schemes are impacted by it. For proactive cases its effect is reduced, as the HO process is triggered by the detachment detection, contrarily to reactive schemes where the process only starts after the MN attachment. As such, in proactive cases, if the MN attaches during the signaling process, the service disruption time is not affected;

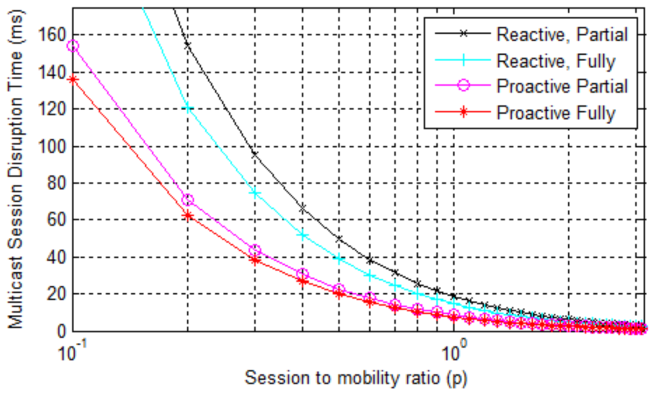


Figure 5.4: Service disruption time variation with SMR

therefore, it is expected to be larger in reactive cases. For the results calculation, the value of t at spans the interval [0.1, 100] ms, aiming to reflect as well worse wireless environments. From this point, the value of ρ is considered to be 0.1. Figure 5.5 shows that for reattachment time above 10ms, the service disruption increases significantly.

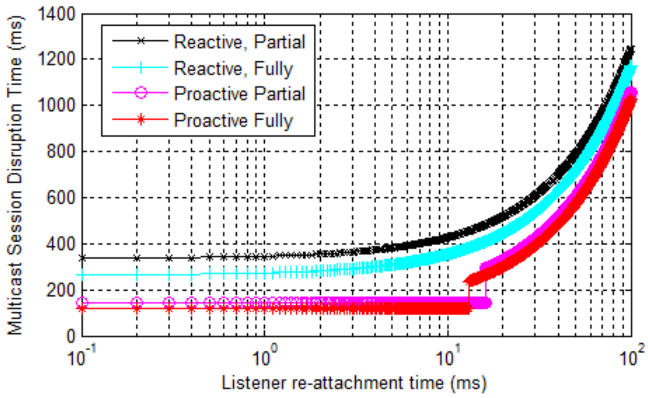


Figure 5.5: Service disruption time variation with host re-attachment time

5.2.4.3 Tunnel Delay Impact

The value for t_{MM} was varied over a range of 0.1 to 10ms. In Figure 5.6 it is visible how the schemes more dependent on the tunnel (i.e. fully distributed) may suffer in case it is non-optimal (e.g. ending in over-demanded or distant MARs).

5.2.5 Discussion

In this section, we provide a thorough analysis of each solution, inspecting their performance in aspects as ease of deployment, complexity, signaling and tunneling overhead. The section is organized according to the two previously referred parameters: distribution scheme and

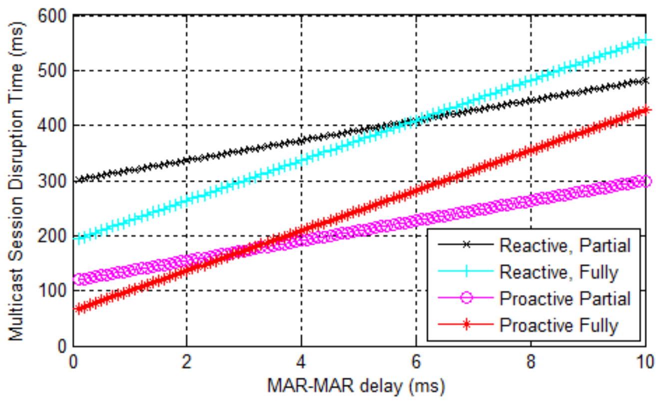


Figure 5.6: Service disruption time variation with tunnel traversal time

solution proactivity. Moreover, some considerations on the multicast functionality at the MAR (i.e. MLD Proxy vs Multicast Router usage) are done.

5.2.5.1 Distribution Degree

Partially distributed schemes, while allowing the distribution of the data plane among the access network, imply an extra number of signaling messages comparatively to the fully distributed approach, mainly due to the messages relaying. Additionally, it preserves some of the limitations of centralized mobility management schemes, namely the dependence on the central entity's proper operation, representing a single point of failure. The path travelled by the signaling information, as well as over-demand from the MMIS may lead to extra latency in finalizing the mobility process. Although these require more signaling exchange, they may be considered simpler, and with good degree of independence from other protocols.

On the other hand, a fully distributed scheme requires less signaling, but implies selecting a reliable underlying protocol / framework, such as 802.21 MIH, for accessing and exchanging information such as pMAR addresses(s), etc.

5.2.5.2 Solution Proactivity

Proactivity is an essential feature in applications which require seamless HO. In multicast applications, the HO impact includes the time required for the nMAR to subscribe the missing channel(s), thus it is useful to include multicast subscriptions within mobility signaling messages. Comparing Figure 5.3 (b) and Figure 5.3 (d), there is a clear trade-off between the number of signaling messages and the complexity of the solution (i.e. the required protocol extensions). The inclusion of destination MAR option requires additional intelligence at each MAR, while the multicast subscription option represents extra processing and signaling overhead.

5.2.5.3 Considerations on the usage of MLD Proxy

MLD Proxy usage is adopted in [57] as a mean to assure multicast listener support without changes to multicast and mobility protocols over PMIPv6. Although, several problems emerge when using a MLD Proxy in a DMM environment (Figure 5.7). First, such scenarios are prone to duplication and tunnel convergence problem. Duplication occurs whenever a MN without mobility status (i.e. no mobility anchor) subscribes to a channel already being received by a MN anchored at another MAR, leading to an unnecessary copy, while tunnel convergence will occur for MNs subscribing common channel(s) that are anchored at different MARs but currently located at the same one. This is depicted by MNs moving from MAR1 and MAR3, in 5.7 (b). A distinct resulting problem is that of non-optimal routing (Figure 5.7 (c)). If we consider a significantly large domain, there is the possibility for the tunnel to encompass a large distance, even if the current MAR is connected to the multicast infrastructure. This issue is a consequence of configuring the MLD Proxy upstream interface towards the anchor.

A possibility is to configure the upstream interface towards the multicast infrastructure [123], although this might imply unexpected delay due to multicast tree reconstruction. As such, and because a MN is expected to subscribe to different channels at different periods in time, an alternative for solving this problem is per-channel upstream configuration. This can be done either by enhancing MLD Proxy's for multiple upstream interfaces support (Figure 5.7 (d)), or by incorporating a decision entity in the operator network for deciding between local or remote subscription in a per-channel basi. This approach is explained in the next section.

5.3 Centralized Multicast channel management in DMM

The purpose of DMM is to mitigate traffic convergence to a single anchor, by distributing it to ARs. However, when simply applying a MLD Proxy in DMM routers, severe traffic problems appear, such as redundant multicast sessions, resulting in duplicate traffic, or the so called tunnel convergence problem. The magnitude of this problem in DMM is very different from that of PMIPv6, where tunnel convergence occurs for sessions arriving at a common MAG from distinct LMAs. LMA is an entity positioned at the upper-level of the hierarchy, which means that the the number of LMAs is typically very limited. However, in DMM scenarios, all MARs are access-level entities so it is expected that a MAR can have connections with all other MARs, at least within a domain. Consequently, the impact of the duplicate multicast traffic in DMM is potentially much higher than that of PMIPv6. Another performance problem of this approach is non-optimized tunnel path, resulting when MNs move away from their anchors while on a long mobile sessions. This may introduce

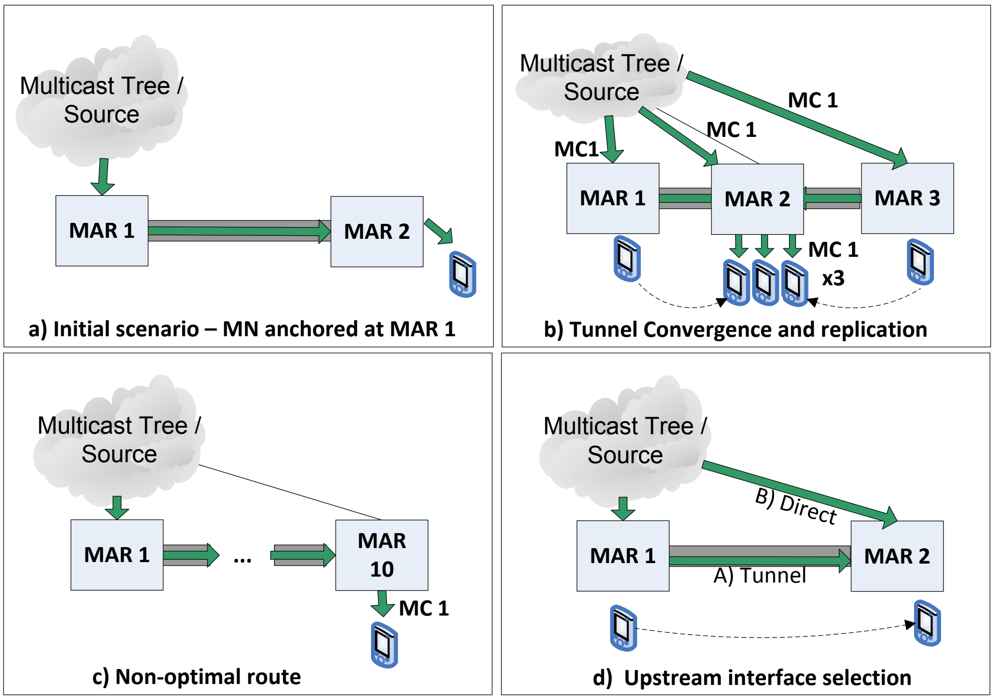


Figure 5.7: Issues with MLD Proxy in DMM scenarios

packet delivery latency and reduce session's liveliness in real-time multimedia broadcasting,

This work was a necessary step for better understanding the potential design possibilities for IP multicast support in DMM. One path to minimize the tunnel replication cost through flexible multicast channel management was explored with the author in [5]. The main concept is to apply centralized channel management policies, so that operators can flexibly adapt content distribution to several factors, such as popularity. This work, being the result from cooperation between the author of this Thesis and the author of the article, was a necessary step for better understanding the potential design possibilities for IP multicast support in DMM.

5.3.1 Reference Architecture and Operation

An IP Multicast framework applying centralized channel management was designed. The goal of such centralized management is facilitating efficient multicast traffic distribution, and to enable a policy-based channel management according to the operators' network environments. Towards this, a channel control server (CCS) is proposed while still considering the deployment of MLD Proxies on MAR. Besides, each MAR employs a channel enforcement function (CEF) which stores channel lists classified as 'L' or 'R', representing which channel should be locally or remotely supported. This classification is provided by CCS to the CEFs. The modules organization within the MAR is depicted in 5.8.

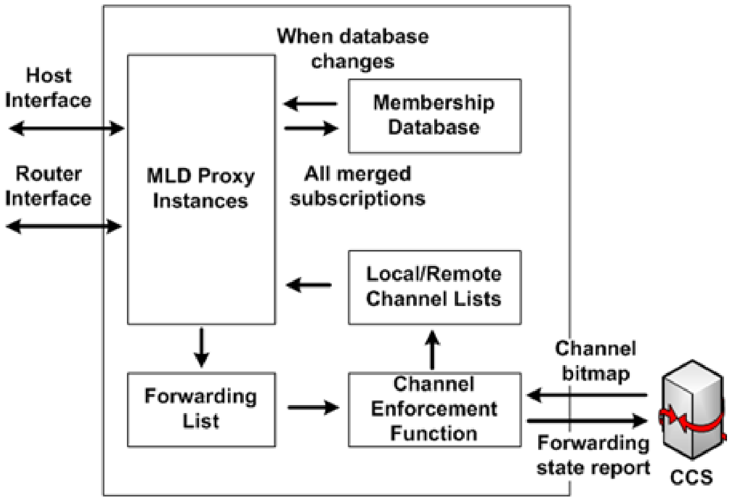


Figure 5.8: Multicast function structure on MAR

Once a MN attaches to a MAR, the direction of upstream interface is set based on the received policy from the CCS. Considering Figure 5.8, MN1 and MN2 were listening to CH1 and CH2 at MAR1 and MAR3, respectively, and both move to MAR2. CEF determines that CH1 is a local channel and CH2 is a remote channel. As a consequence of this channel policy, CH1 directly routed to an upstream IP multicast router, while a MLD Report message regarding CH2 is transmitted towards MAR3 in order to have MAR2 receiving the multicast packets through the tunnel towards MAR3. If MAR2 is asked to forward CH1 packets to other MARs, it adds a new downstream interface to the corresponding MLD Proxy instance towards requesting MAR. Thus, the meaning of “local channel” is limited to the reception of multicast packets.

By providing channel management policies, operators can flexibly adapt content distribution to several factors, such as popularity. For instance, popular sports games can be provided through direct routing (locally available) only, while less popular channels are allowed through both routing mechanisms, as the probability for a MAR to be subscribed to such a channel is reduced. This way, the efficient multicast traffic distribution is facilitated, and a policy-based channel management according to the operators' network environments is enabled.

The author [5] also evaluates the framework performance against direct application of MLD Proxy [4], and discusses in more detail CCS deployment considerations such as its correlation with eMBMS's BM-SC, as well as several performance aspects like the introduced overhead for CEF - CCS communication, or service delivery latency resulting from MARs lack of synchronization.

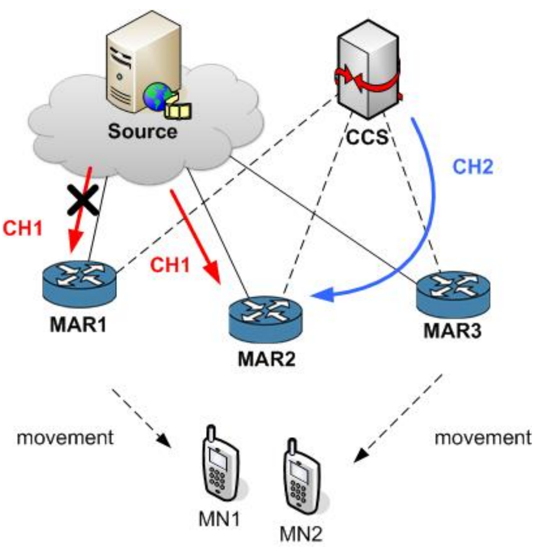


Figure 5.9: Multicast data forwarding per channel in CMM-DMM

5.4 Improving Efficiency through the usage of Multicast Routing

It was verified that the utilization of standard MLD Proxy within DMM environments may lead to the duplicate traffic issue, where a serving DMM router receives multiple copies of same multicast stream, one for each IGMP/MLD Proxy instance running on the router. Aiming to resolve this issue by tackling the lack of coordination between the IGMP/MLD Proxy instances and the serving router, a channel-manageable IP multicast architecture framework has been presented in [5]. It resolves the duplicate traffic issue by introducing a channel control server managing a multicast channel for a given serving router. However, due to its focus in assuring the IP multicast mobility service efficiency, service disruption while coordinating multiple IGMP/MLD Proxy instances may be a consequence. Besides, managing all the serving routers does not scale.

On the other hand, Protocol Independent Multicast Sparse Mode (PIM-SM) [40], being proven by production experience, is one widely accepted multicast routing protocol used to build multicast networks. Contrarily to MLD Proxies, it enables a multicast router to seamlessly manage multiple upstream interfaces; for such, Reverse Path Forwarding (RPF) is used to decide from which interface the multicast packet should be received among all the available routing interfaces. As such, it could be another option to be exploited over distributed mobility architecture.

In this section, a flexible architecture providing multicast mobility protocol solutions for DMM environments, dubbed Distributed Mobility Management and Multicast-enabled architecture (D3M)), is defined. It leverages on PIM-SM routing protocol, and enables different combinations based on two design criteria: multicast subscription discovery and the multicast packet resumption origin. The proposed D3M architecture is based on PMIPv6, and as such the MN is not involved in the mobility signaling process.

5.4.1 Reference Architecture and Operation

The proposed solution leverages on a network-based DMM protocol, re-using PMIPv6 functionalities and mobility signaling messages. The base entity is the MAR, merging the functionalities of a mobile access gateway (MAG) and a local mobility anchor (LMA)[3]. The MAR is classified into a serving MAR (S-MAR) acting as a MAG and an anchor MAR (A-MAR) acting as a LMA in a physical entity. The proposed architecture follows a partially distributed approach, where the data plane is distributed among the MARs, whereas the control plane is centralized at the multicast mobility information server (MMIS) [4], which acts as an anchor discovery proxy and a central mapping database between MN's prefix and the responsible anchor. By following a partially distributed design and not a fully distributed one, security and complexity issues are reduced. To refer an example, distributing the control plane implies the need for the new access router to identify the anchors using an external mechanism or protocol.

The typical mobility protocol operation when an MN moves is illustrated in Figure 5.10. In the scenario, MN initiated a flow at A-MAR. In case mobility occurs, the S-MAR will transmit a PBU to the MMIS after MN attachment, and the MMIS will forward it to the MN's previous MAR, which is now the A-MAR for the flow. Once the A-MAR receives PBU from MMIS, it then configures necessary routes for anchoring the traffic and sends back a PBA to MMIS and S-MAR. S-MAR will complete the tunnel configuration after reception of the PBA.

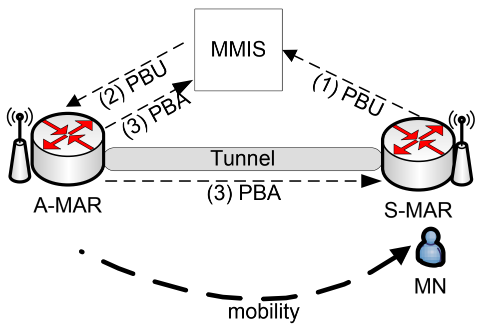


Figure 5.10: Reference DMM scenario

We extend the aforementioned DMM operation by enhancing each MAR with multicast forwarding and routing capabilities. With regards to the design of the IP multicast mo-

[3] This concept first appeared in MEDIEVAL Project, Deliverable D4.1: "Light IP Mobility architecture for Video Services: initial architecture", Jun. 2011

bility solution for DMM networks, there are two key factors taken into account; (i) how the MAR will learn about multicast subscription information of attached or incoming MNs and (ii) where from the router will receive multicast packets after mobility. Both considerations target seamless multicast service support. However, the former is associated with subscription information acquisition by S-MAR, while the latter defines how the IP multicast session subscription and reception are done after HO event. Taking these two design criteria into account, three D3M modes are defined, characterized by different combinations of multicast context discovery and packet delivery methods: native IP multicast (NIM), native IP multicast with subscription transfer (NIM-ST), and anchor-based multicast (ABM). NIM uses native IP multicast approach based on direct multicast routing and does not rely on any benefit of the proposed DMM support. Concretely, PIM-SM is installed at MARs and independently runs regardless of DMM protocol operation. NIM-ST mode also uses native IP multicast approach but additionally employs anchor-assisted channel subscription transfer aiming at fast subscription acquisition of incoming MNs, thus reducing service disruption latency due to mobility. Finally, ABM not only takes advantage of channel subscription transfer but also provides multicast packet anchoring, using the established tunnel to forward IP multicast from A-MAR to S-MAR. Each mode is independently available for facilitating IP multicast but it can be combined and used depending on the required level of performance of the underlying service(s). ABM has the highest degree of seamless service support among all D3M modes, whereas NIM has the lowest. However, this does not mean ABM is always the best among them, because the performance is highly dependent on various factors. Through the following sub-sections, we describe the detailed operation of the three modes and corresponding internal operation within the MAR, before assessing their deployment validity from an implementation point of view.

The generic attachment process of a MN in D3M is illustrated in Figure 5.11. For both NIM-ST and ABM, the multicast subscription event will lead to the storage of multicast context in the Binding Update List (BUL) entry, which does not occur in NIM. We will now present the operation for each of the proposed schemes.

5.4.1.1 Native IP Multicasting (NIM)

NIM operation is shown in Figure 5.12 (a). When the MN moves to a new S-MAR, the S-MAR will assign a new HNP to the MN and send several General MLD Queries. Once the MN receives the MLD Query message, it will send a MLD Report with designated QRI value. Hence, the new S-MAR will be informed about the channels information to which the MN has subscribed. This mode is the simplest approach for multicast mobility support in a distributed deployment environment, only relying on native IP multicast infrastructure and tuned behavior of the MLD operation for mobile scenarios, but regardless of IP mobility

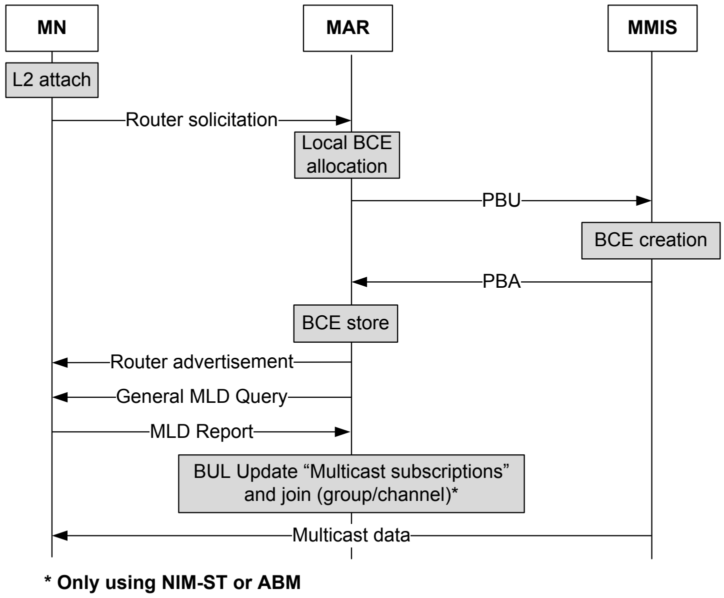


Figure 5.11: Initial attachment and multicast subscription in D3M

operations.

As a trade-off for the simplicity of the implementation, it requires MNs to endure significant delay in resuming existing multicast session after mobility. By tuning the QRI to a minimal value, it is possible to minimize the total service latency that the MN has to wait before sending the MLD Report to the new MAR, but may lead to severe signaling overhead and consequently higher loss probability over wireless mediums.

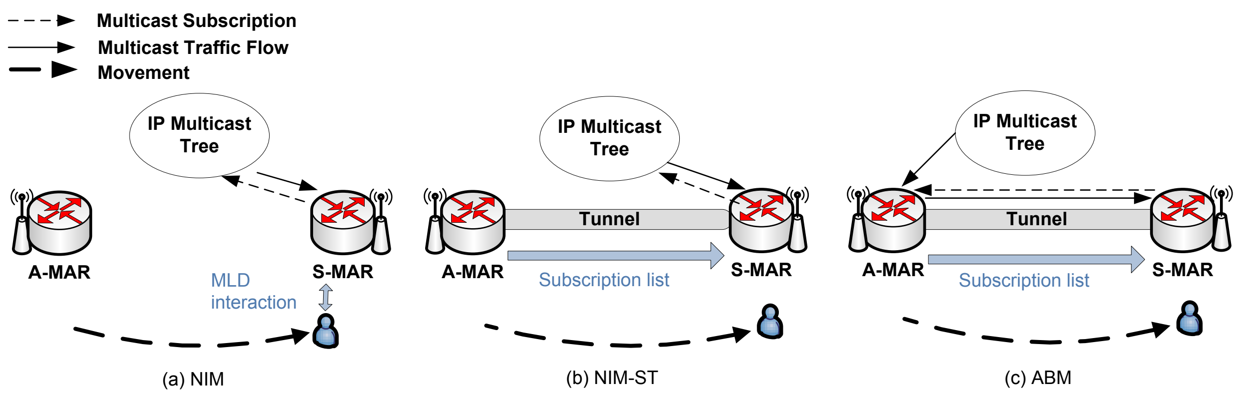


Figure 5.12: Operation comparison among three modes of proposed D3M

5.4.1.2 Native IP Multicasting with Subscription Transfer (NIM-ST)

NIM-ST enables the new S-MAR to be aware of channel subscription information by multicast subscription transfer from the previous MAR (A-MAR) to S-MAR, as mentioned before. This mode inherits from multicast fast HO [60] and can be activated by detecting the MN detachment or attachment. In the predictive mode, the HO is detected by the P-MAR, which will then contact MMIS to determine the target S-MAR while in the reactive mode, the S-MAR detects the attachment of the MN, initiating the signaling with a PBU message towards MMIS.

Suppose that a MN is subscribed to one or more multicast channels after initial attachment at a MAR, leading the MAR to update the subscription entry in the BUL. Following the reactive mode shown in Figure 5.13 (a), with the mobility of MN to a new MAR, the new S-MAR obtains the MN-identifier (MN-ID) and sends a PBU message to MMIS. As the anchor discovery proxy, MMIS forwards received PBU message to MN's A-MAR. A-MAR then retrieves the target S-MAR address, and sends an extended PBA message including registered multicast subscriptions to S-MAR, as well as a regular PBA message to MMIS to inform the completion of the mobility process. When the S-MAR receives the extended PBA message from A-MAR, it checks the multicast subscription from its MRIB to identify missing multicast subscriptions to be joined immediately. S-MAR fills the BUL entry corresponding to the MN, including the Multicast Records field, and sends the routing advertisement with the anchored HNP to the MN. The upstream router is chosen by standard PIM-SM operation, based on the MRIB.

In the predictive mode shown in Figure 5.13 (b), once A-MAR detects the detachment of the MN, a HI message containing the Subscription List option is sent towards MMIS, which forwards it to the S-MAR. When receiving the message, S-MAR checks which subscriptions are missing and transmits a HAck with the Subscription List option, potentially containing a sub-set of the initially received list corresponding to the missing ones. It will then join the corresponding sessions by PIM Join message.

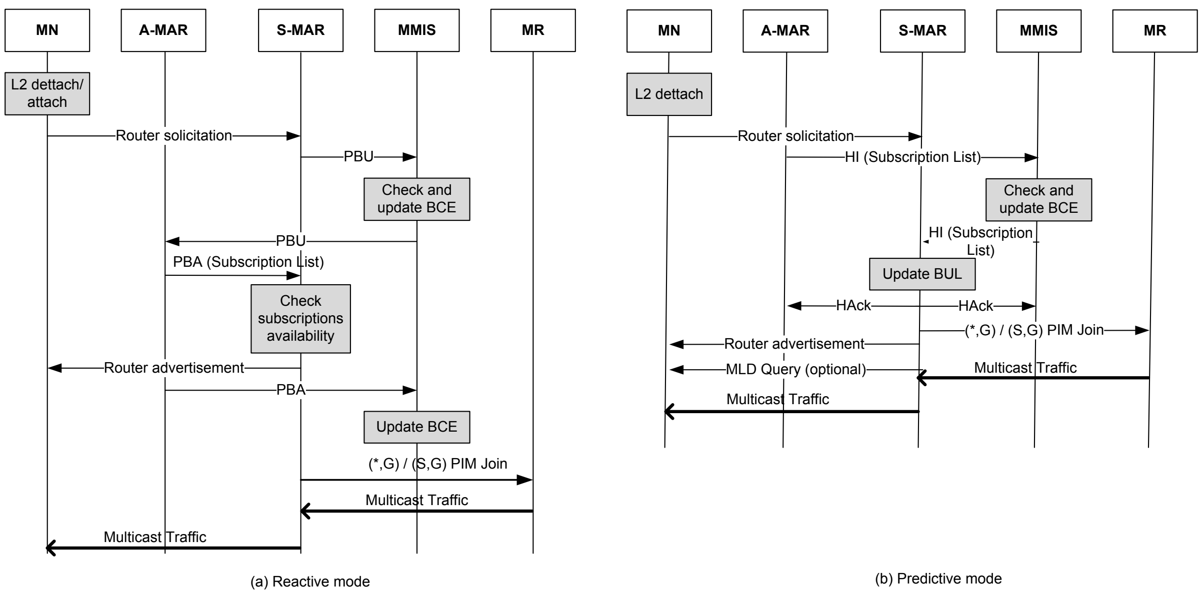


Figure 5.13: Signaling procedures in NIM-ST

5.4.1.3 Multicast Transport through the Anchor (ABM)

ABM mode has a similar control plane operation to NIM-ST, but the major difference is the creation and utilization of the mobility tunnel for forwarding multicast traffic. In reactive mode (Figure 5.14 (a)), when receiving the PBU, A-MAR configures its endpoint of the mobility tunnel with S-MAR. When S-MAR receives the PBA it then completes the tunnel establishment and begins receiving the forwarded traffic through the tunnel. The main advantage with this design is that it overcomes the usage of a multicast routing protocol and corresponding latency, allowing a fast HO by taking advantage of the mobility tunnel.

In predictive mode (Figure 5.14 (b)), A-MAR detects MN's detachment and then sends the HI with the Subscription List to MMIS, which is forwarded to A-MAR. Through this process, the tunnel creation is initiated by A-MAR. The reception of the HAck message at A-MAR completes the tunnel establishment and triggers the transmission of the multicast subscriptions of interest via the tunnel.

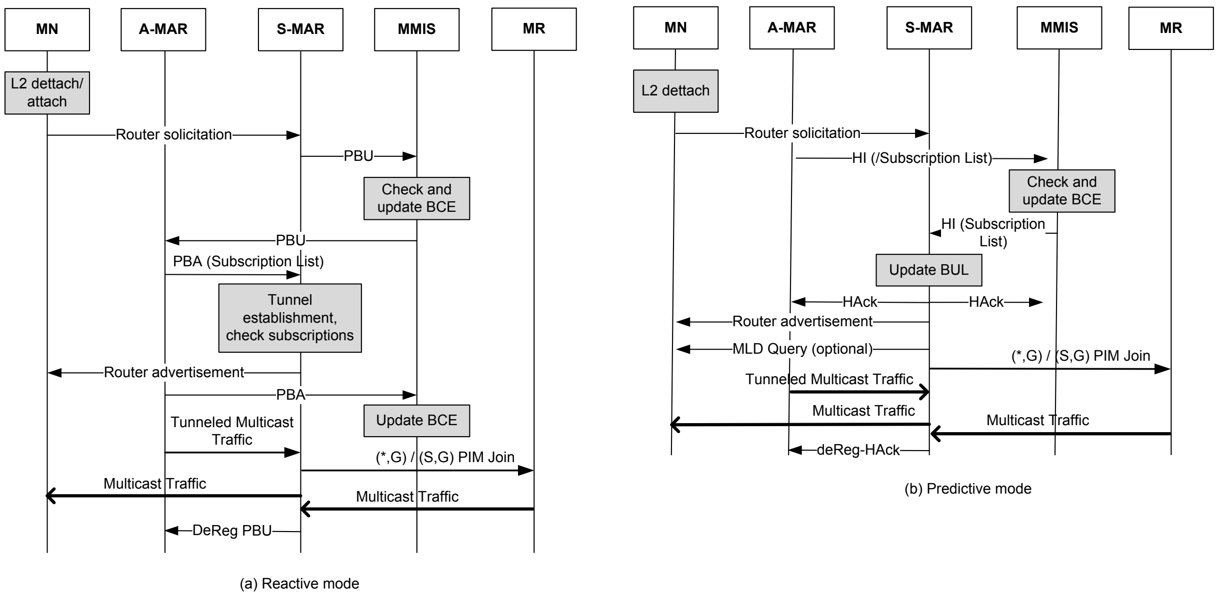


Figure 5.14: Signaling procedures in ABM

5.4.1.4 Coexistence of the different Schemes

As previously mentioned, network operators must design their network solutions taking into account distinct business or technical-related aspects. This fact demands flexible architectures leveraging on approaches beyond a "one-for-all" mobility solution. For instance, multicast mobility support of live Internet (e.g. YouTube) content may not be done in the same way as Operator-owned video services. The coexistence of the three D3M modes enables mobile operators to flexibly devise distinct multicast mobility management strategies.

The decision for selecting the mobility mode may be based on multiple factors such as the user profile, e.g. MN-ID, and the multicast flow properties, e.g. type of service. Relatively delay-tolerant services under a Basic Service profile may fit into NIM scheme, while the NIM-ST scheme could be suitable for users subscribed to a multimedia service like mobile IPTV. This second option implicitly assumes HO control by the operator for the subscribed channels being reliably stored in the BUL. Besides, NIM-ST can be applied as an intermediate approach, as it can deliver optimized HO disruption for networks, services or operators which do not have interest in or do not comply with the encapsulation of IP multicast traffic in general. For a delay-sensitive application like real-time video, and as long as the network infrastructure facilitates a tunneling mechanism assuring required QoS values, ABM could be used for achieving minimal IP multicast subscription and delivery latencies.

The utilization of proposed schemes can be considered in 3 main strategies:

1. Employ a single scheme for all multicast services and users (e.g. NIM-ST);
2. Employ a single scheme based on a specific metric / threshold;
3. Orchestrate all of the modes in a per-service / per-user basis. The application of approaches (2) and (3) requires a concrete selection algorithm.

The user profile and service information could be stored in a centralized database, or this information could be available at all MARs in the domain so the HO disruption is minimized. Taking the latter case into consideration, a use case, where the Operator combines three modes for three users having different user profiles, is depicted in Figure 5.15. Strategy (2) is used for all Base Service users (MN1). As such, scheme NIM is applied as soon as the operation cost is near the predefined threshold; otherwise, NIM-ST may be applied. Strategy (3) is applied in a per-user basis. In this case, all sessions from the operator's Silver Service – represented by MN3 - are supported by NIM-ST, while Gold Service users, willing to pay for having all its sessions, are supported using ABM – represented by MN2. The selection criterion employed in this use case is depicted in Fgure 5.16. Nevertheless, the comparison of distinct design criteria is out of scope of this work.

The three D3M modes can be coupled with an Explicit Tracking module [121] for efficiency-wise improved operation. While each BUL contains the subscriptions for one attached MN, the IGMP/MLD-based Explicit Membership Tracking would hold the aggregated subscription view from all attached MNs. Such could assure quick IP multicast group leaves after the sole listener of a channel moves from the MAR or unsubscribes the channel, which usually results in the so-called leave latency.

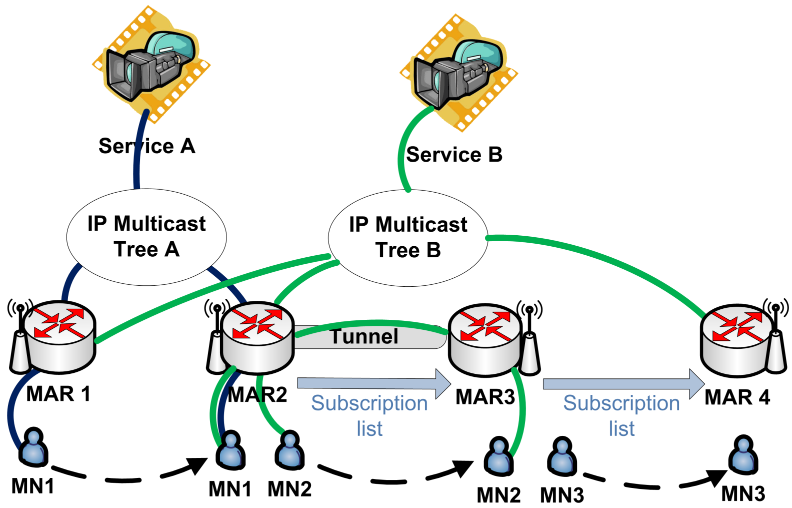


Figure 5.15: Use case combining the multiple modes of D3M

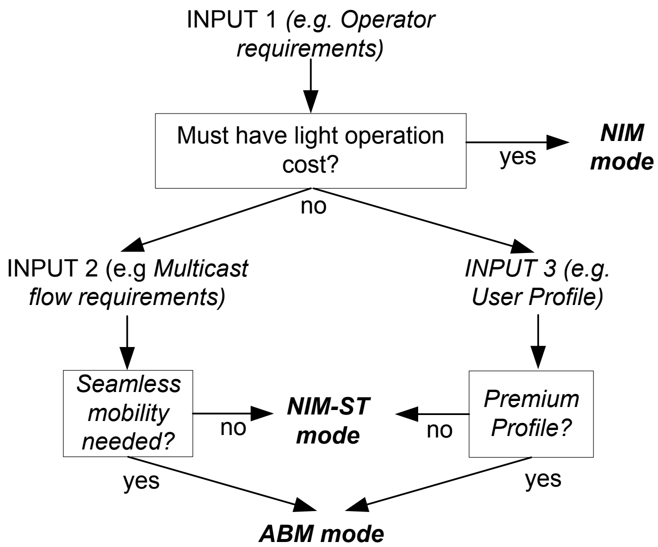


Figure 5.16: Decision flow of D3M mode selection

5.4.1.5 Internal Operation

Figure 5.17 shows the internal design of the MAR defined in our D3M scheme, focusing on interactions between the multicast and mobility planes, and processing control and data packets. Once a packet arrives at input queue ('1'), it is classified into different queues and handled according to its type. Multicast signaling and data packets are handled using standard PIM-SM operation: if a multicast data packet arrives through a mobility tunnel, it must be de-encapsulated ('2'). If the packet was received through a regular interface, a lookup is performed at the Multicast Forwarding Information Base (MFIB), determining the existence of an entry associated with the subscription ('3'). The RPF check verifies whether the packet arrived through the expected interface or not ('4'). If the matched incoming interface (IIF) is found in the RPF check, the packet is then additionally checked to identify the need for encapsulation. In this process, a proper outgoing interface is also found ('5'). Finally, the packet is forwarded through the selected interface(s) ('6').

Regarding signaling messages, the processing is done as follows. All PIM or MLD messages as well as internal control messages like cache miss or wrong incoming interface happening on the multicast plane are passed to the PIM-SM buffer ('7'). MLD Report or Done messages will affect the BUL's Multicast Records field corresponding to the originating MN ('8'). Concerning mobility signaling, the reception of an extended PBU or HAck messages is received in reactive or predictive modes, respectively, will lead the S-MAR to setup its tunnel endpoint, and join any missing subscription ('9').

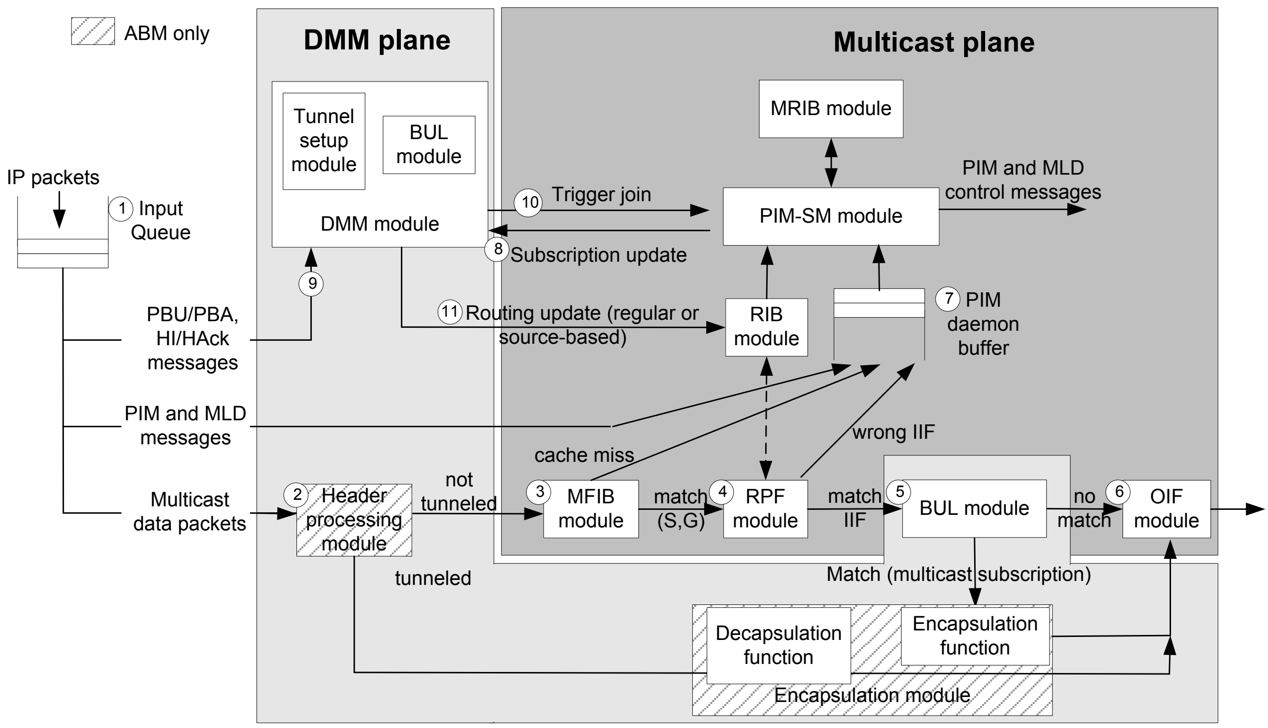


Figure 5.17: MAR internal operation

5.4.2 Quantitative Analysis

In this section we evaluate the proposed D3M solutions by mathematical analysis. The main goal is to identify the associated benefits of integrating multicast functions with the mobility protocol. To simplify, and because it is not our goal to evaluate the impact of prediction mechanisms, we opted by evaluating only reactive modes, which represent the typically considered scenario. We evaluate the schemes over two main stages. In the first scenario, we consider a single HO event, and evaluate the service disruption time and packet loss during HO. In the second scenario, packet delivery cost is calculated by taking several mobile multicast users into account and for multiple mobility events. The former is intended to show basic performance of IP multicast mobility from the QoS perspective, while the latter focuses on identifying the impact of the proposed D3M solutions in the data plane.

5.4.2.1 User Mobility and Traffic Models

Figure 5.18 shows the assumed network topology for performance analysis, where MARs are serially connected and hierarchically connected with upper multicast routers (MRs). The whole network is organized in a binary tree of h layers where MARs represent the last hop. Multicast traffic is sent from a single multicast Source, which corresponds to the root of the tree and is h hops away from any of the MARs. The transport delays employed in the mathematical equations are as follows:

- t_{at}: the layer-2 HO delay, i.e. the time taken from MN's link detachment from previous MAR to MN's link re-attachment at target MAR, including channel scan, authentication and link association;
- t_{UM}: the time taken to transmit a data or signaling packet between the MN and the MAR, including the time spent in the wireless access network;
- t_{MM}: the time taken to transmit a data or signaling packet between two directly connected MARs;
- t_{MS}: the time taken to transmit a data or signaling packet from a MAR to MMIS;
- t_{MI}: the time taken to transmit a data or signaling (such as PIM) packet through the IP multicast tree to a specific MAR. This value depends on the distance to the neighbor MR with the active subscription, and is only relevant for NIM and NIM-ST modes where traffic arrives from the IP multicast tree after mobility. In Figure 5.18, t_{MI} is depicted for a scenario where the MN moves from MAR 1 to MAR 2. If the MN moves to MAR M, t_{MI} would be the time that a packet takes to travel from the closest upstream routing member of the multicast tree to MAR M.

For the mobility model, we assume that N MNs are randomly distributed among R MARs, and move around the DMM network at each HO event, for a total of H HO events. The next S-MAR of each MN is determined among three options – the current MAR, a MAR on the left or on the right side – with equal probability (1/3). However, when the MN is in the first or last MAR, the next position of the MN will be bound to two options – current MAR or the other MAR – with half probability. In the particular case of ABM, it is assumed that a MAR can establish a bi-directional tunnel with any of the other MARs; as such, the

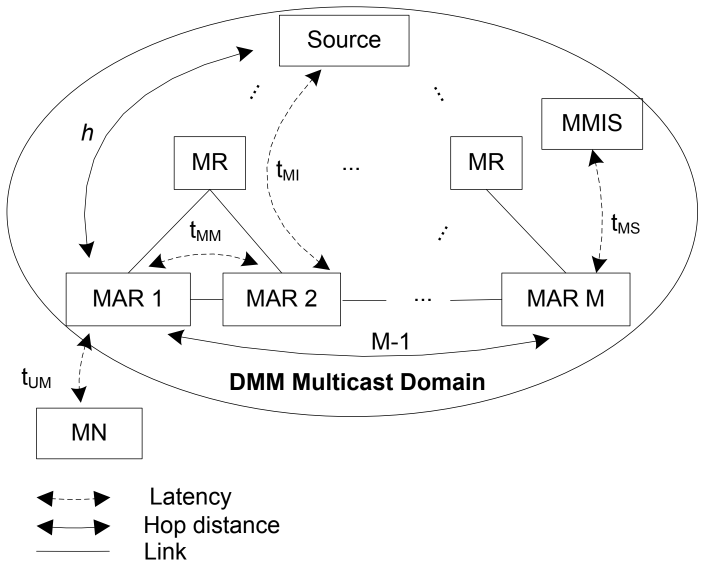


Figure 5.18: Network topology for performance analysis

maximum hop distance is equal to R-1, reflecting the case where MAR 1 tries to establish a tunnel with MAR M.

The multicast session inter-arrival time follows an exponential distribution with rate λ_I and its average time is E(I), expressed by $1/\lambda_I$. The subnet residence time also follows an exponential distribution with rate λ_S and average value denoted as E(S), or $1/\lambda_S$. We define SMR, denoted as ρ, as the ratio between the session arrival rate and the HO rate, given by $\rho = E(t_S)/E(t_I) = \lambda_I/\lambda_S$.

Mobility in ABM mode is modeled as follows. The tunneled multicast transport can be effective if no MNs are receiving the same group or channel at the target MAR during the HO. Otherwise, the subscription is received through the previous upstream interface selected from either a tunnel or a regular interface. When MNs from distinct MARs move to the same MAR which was previously non-occupied, only one tunnel will be used for multicast transport, following standard PIM-SM routing operation. In those cases, it is considered that only the tunnel corresponding to the MN with lower index is established.

5.4.2.2 Service Disruption Latency

The service disruption latency for multicast packet delivery is defined as the time since the MN detaches from the previous MAR until the MN receives a IP multicast packet in the new S-MAR, and for each X mode of D3M it is represented by D_X.

In NIM, the service disruption (Dt_{NIM}) consists of the HO latency (D_{HO}), the mobile multicast subscription (D_{MLD}), and consequent routing convergence and delivery latencies ($D_{DLV-NAT}$) by:

$$D_{NIM} = D_{HO} + D_{MLD} + D_{JOIN} + D_{DLV-NAT}, \tag{5.7}$$

where D_{HO} is given as the sum of layer-2 (t_{at}) and layer-3 attachment processes, which correspond to link scan, security check, and link association for the former, and movement detection, router discovery, mobility management and IP address configuration for the latter. D_{MLD} is the latency due to the MLD process at the new MAR, and $D_{DLV-NAT}$ is due to the PIM-SM routing convergence time. Besides, D_{HO} is given by:

$$D_{HO} = t_{at} + 2t_{UM}. \tag{5.8}$$

Moreover, D_{MLD} is computed as:

$$D_{MLD} = 2t_{UM} + t_{QRI}, \tag{5.9}$$

where t_{UM} represents the transmission time of MLD Query and Report messages, and t_{QRI} the Query Response Interval. It is assumed that the MLD General Query is immediately sent after layer-2 attachment and IPv6 address configuration, meaning that Startup Query Interval is 0.D_{JOIN} denotes latency due to the joining process for transmission of the necessary PIM Join message(s), and $D_{DLV-NAT}$ denotes the routing convergence time ($t_{CONV-NAT}$) plus the time that MAR takes to send the traffic to the MN.

$$D_{JOIN} = t_{MI}, \tag{5.10}$$

$$D_{DLV-NAT} = t_{CONV-NAT} + t_{UM}. \tag{5.11}$$

In NIM-ST and ABM schemes, disruption during mobility is affected by the D3M mode-specific signaling procedure, which is expressed as D_{HO-MOB}. Similarly to NIM, the total disruption in NIM-ST considers the PIM join transmission latency and the delivery latency of the first packet from the multicast tree to the MN. The total service disruption latency for NIM-ST and ABM are given by:

$$D_{NIM-ST} = D_{HO-MOB} + D_{JOIN} + D_{DLV-NAT}, \tag{5.12}$$

$$D_{ABM} = D_{HO-MOB} + D_{DLV-TUN}. \tag{5.13}$$

D_{HO-MOB} includes the time for receiving the RS but not RA, because the latter is considered to be sent after the HO signaling and in parallel to the multicast join and delivery process. Thus, D_{HO-MOB} for both NIM-ST and ABM is defined by:

$$D_{HO-MOB} = t_{at} + t_{UM} + 2t_{MS} + t_{MM}. \tag{5.14}$$

In ABM, the multicast traffic will flow through the tunnel as soon as it is established. Thus, the multicast packet delivery time translates the time it takes for the multicast traffic to reach the MAR using the tunnel ($T_{CONV-TUN}$), plus the time that MAR takes to send the traffic to the MN (T_{UM}):

$$D_{DLV-TUN} = t_{CONV-TUN} + t_{UM.} \tag{5.15}$$

Figure 5.19 shows a timing diagram comparing service disruption latencies among the three D3M modes. Concerning the proportions used for each of the factors representation, we grouped the latency factors in three different groups according to the expected weight: Low, Medium and Large latency. Within the first group, we include T_{UM}, T_{MM}, T_{MI} and $T_{CONV-TUN}$, which corresponds to T_{MM} - such values are in the order of milliseconds under the literature, as will be shown in Section 5.4.3; classified under the second group are re-attachment latency (D_{L2}) and T_{MS} factors; to finalize, within the latter group, both $T_{CONV-NAT}$ and QRI interval are included, as they are typically in the order of seconds.

It can be observed that the performance difference between NIM-ST and ABM is associated with the upstream router target: NIM-ST depends primarily upon the routing convergence time of the multicast tree, associated to the multicast infrastructure topology, while ABM depends upon the mobility tunnel conditions.

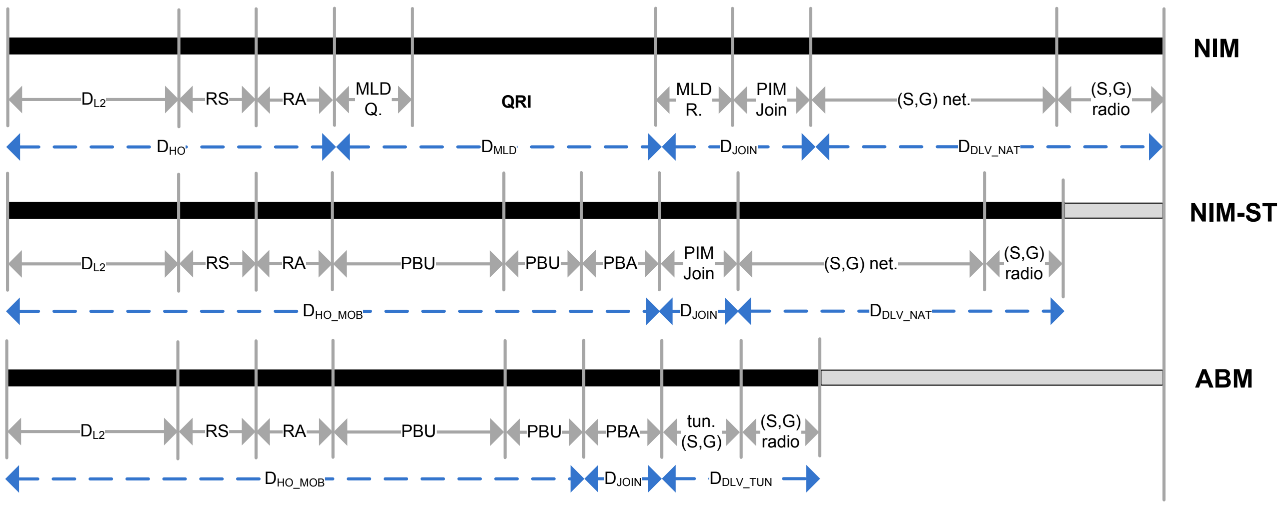


Figure 5.19: Timing diagram for service disruption latencies among D3M modes

5.4.2.3 Packet Delivery Cost

The total packet delivery cost C is defined as the total hop count required for multicast packet delivery from the multicast source to all the mobile multicast listeners, and the cost for each mode X is represented as C_X. In case IP mobility tunnel is not activated (NIM or NIM-ST modes), the cost is computed by the routing cost C_R, which denotes how many routing hops a multicast packet went through to reach all MARs occupied by at least one MN.

In case of IP mobility tunnel activation (ABM), C_T is computed as the sum of C_R and the cost for tunnel forwarding (C_F), where C_F denotes the cost associated to traffic which went through the total n active mobility tunnels used for multicast packet delivery, as shown in equations (5.16) and (5.17):

$$C_{ABM} = C_R + C_F. \tag{5.16}$$

$$C_F = \sum_{i=1}^{n} C_{A-MAR_i->S-MAR_j} \tag{5.17}$$

The packet transmission cost over wireless media from MARs to MNs is neglected, since our main goal is committed to identify the reduction in backhaul cost and the improvement in multicast service performance during mobility. In the point-to-point link model used in PMIPv6, the total link transmission cost would be the same for all the schemes, and equal to the number of receiving MNs.

5.4.2.4 Packet Loss

The average packet loss (L) during HO can be obtained in a straight-forward way, based on packet arrival ratio and the corresponding disruption time:

$$L = \lambda_I \times T \tag{5.18}$$

5.4.3 Numerical Results

Based on the analysis of the previous section, we performed some numerical analysis, taking reasonable values for the multiple parameters at stake. Some of the used parameters are derived from the literature. For the latency values, it is considered that t_{UM} = 7 ms and t_{MS} = 10 ms by taking real values in 3G and UMTS into consideration [124] and the total delay contribution of each routing hop is 1 ms [125]. Regarding the network dimension, we consider h = 3, which results in t_{MI} = 3ms. By default, we use the distance value between anchor and serving MAR is equal to 1, resulting in a delay t_{MM} = 1 ms. Finally, λ_I is considered to be 100 packet/s, and t at approximated to 50 ms as considered in [126]. In the first part, service disruption time is calculated for a given topology through mathematical calculation. In the second part, packet delivery cost was achieved through MATLAB using a custom simulation script.

5.4.3.1 Service Disruption Time

The service disruption latency is evaluated with the impact of three different factors, organized in three corresponding stages: Stage (i) SMR; Stage (ii) the inter-MAR latency (t_{MM}); and Stage (iii) additional multicast tree join latency. The variation of SMR permits assessing the variation of disruption latency for different rates between mobility and session duration. t_{MM} reflects delay impact due to subscription transfer in NIM-ST and the packet transport through the mobility tunnel in ABM. The introduced join latency translates a variation in the multicast routing hop count from the closest router subscribing the multicast traffic – referred as "source MR" – to the target MAR, and is used to study how the multicast infrastructure size affects the HO latency in the different schemes, in particular NIM / NIM-ST modes.

SMR was assigned a value of 1, except in Stage (i), where it spans the [0.1; 1] range. As in [127], we consider a configurable parameter for additional join latency, derived from the multicast routing tree convergence time. In [128], the authors consider it to fit in the range [0, 5] s. In this paper we consider such value to be 1s by default, except for Stage (iii) where it spans from 0 to 3 s.

The results regarding Stage (i), when varying SMR values, are depicted in Figure 5.20 (a), where it is observed that NIM and NIM-ST present increased disruption time with increased user mobility. For SMR values closer to 1, where the HO rate is equal to the session inter-arrival rate, the differential disruption between NIM / NIM- ST schemes and ABM is reduced to a difference of about 1s; this means the major factor in the HO disruption in NIM-based schemes is due to the multicast join process. On the other hand, the disruption suffered using ABM scheme is below 100 ms for SMR values close to 1.

The results for stage (ii) obtained by varying t_{MM} are shown in 5.20 (b). It is observed that even for larger values of t_{MM}, ABM outperforms the other schemes. Even with ABM performance being increasingly degraded for higher values, which can either correspond to a mobility tunnel spanning a high hop number of routers or highly congested or unplanned mobile network topology, the improvement when using ABM is around 1.1s.

Finally, the results for stage (iii) concerning different multicast join latency values are shown in 5.20 (c). It is shown that the variable latency can significantly impact the service during HO, leading to the highest service disruption in NIM / NIM-ST schemes. Even for an ideal (and unrealistic) join latency value of 0s, ABM outperforms the other schemes.

The latter scheme, which does not depend on the multicast infrastructure, shows a stable value below 100 ms. As a brief conclusion, during HO the subscription resume using the multicast infrastructure will typically lead to higher service disruption than using an anchored subscription.

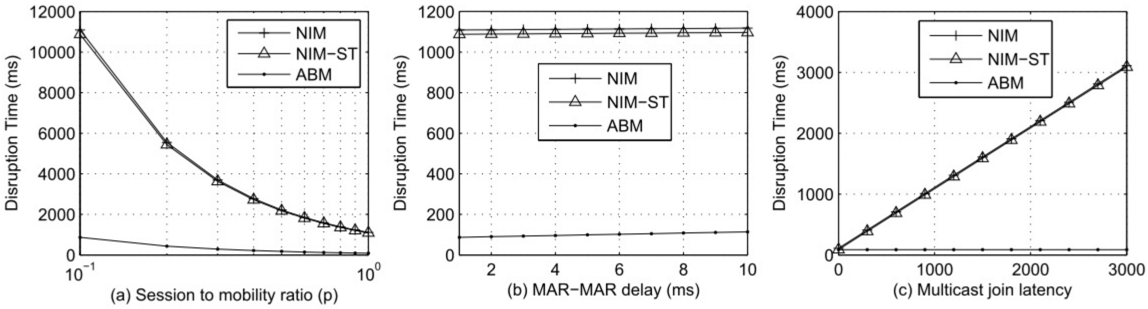


Figure 5.20: Service disruption time as function of (a) SMR, (b) inter-MAR latency and (c) additional join latency

5.4.3.2 Packet Loss

Given that inter-arrival rate is equal to 100 packet/s, the resulting packet loss, for each of the previous calculations is depicted in Figure 5.21. As known, the packet loss is directly proportional to the service disruption time experience during the HO; consequently, the visual differences are quite similar to the previous section. Thus, NIM and NIM-ST schemes present the larger packet loss values, especially for larger SMR and multicast join latencies. Additionally, the impact of the considered t_{MM} range, while increasing the packet loss for all schemes, is not significant.

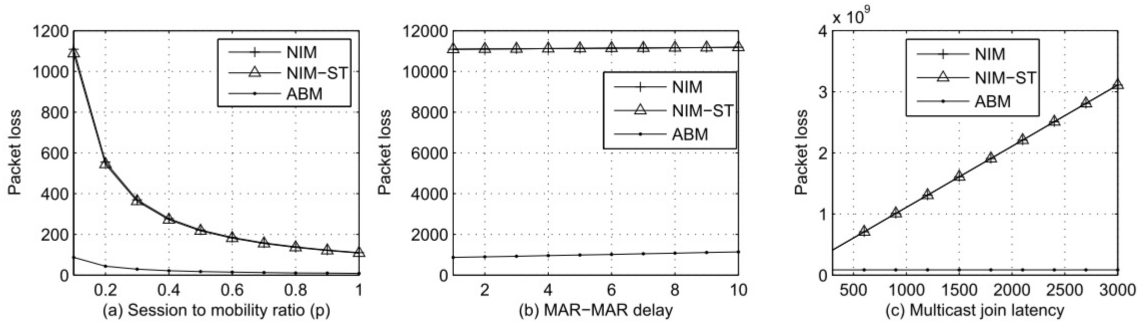


Figure 5.21: Average packet loss as a function of (a) SMR, (b) inter-MAR latency and (c) additional join latency

5.4.3.3 Packet Delivery Cost

The different mobility modes were compared over the mobility scenario described in subsection A, and the mathematical results were achieved over MATLAB using a custom script.

It is again highlighted that the mobility to a MAR with active multicast state means that no subscription is needed, independently of the used scheme. As such, a tunnel is only created and used when the multicast subscription is not yet being received at the target MAR, otherwise the traffic will continue flowing through the same upstream interface as before the mobility process. Two different factors were extensively examined: the initial user / MAR density, and the overall number of users.

Initial User Density Impact

The following values are used throughout the scenarios: N and R are 32 and 16, respectively; user density (ð) is a function of N and takes three different values: ð = 2 (N = 32), 1 (N = 16) and 0.5 (N = 8) – corresponding to 2, 1 and 0.5 MNs per MAR. Besides, h = 3, and T = 20.

For ð = 2, the corresponding packet delivery cost over time for the three schemes is depicted in Figure 5.22 (a). It is verified that the utilization of tunneling in ABM does not introduce a meaningful cost (Ct_F) beyond NIM-ST or ABM. This means that the cost driven from the tunneling was not significant. In order to assess the exact origin of the extra delivery cost in ABM, the total number of tunnels at each event t, as well as the corresponding average cost, was analyzed, as depicted in Figure 5.22 (b). Initially, no tunnels are created, as all the MARs are populated. With the progress in time, some MARs leave the multicast tree due to the absence of MNs, and the arrival of other MNs leads to the tunnel creation. It can be observed that the number of tunnels at each instant was never more than 3, and the average cost per tunnel was 1, meaning that the users either moved to a MAR where the subscription already existed or to a neighbor MAR without the subscription.

The obtained results for ð =1 and ð = 0.5 are depicted in figures 5.22 (c) to 5.22 (f). For both cases, the number of tunnels varied between 1 and 5, and the average tunnel cost was always 1. It was observed that part of the cost was transferred from C_R to C_F , due to the higher probability to move to a MAR not yet receiving the session.

Overall Number of Users

The overall impact derived from the variation in the number of users over a DMM domain was analyzed in more detail. For such, we repeated the simulations considering the users' initial position is now randomly assigned, i.e. without a fixed initial user per MAR density as in previous scenarios. We considered the total number of users for each scenario as follows: 5, 20, 40 and 75.

The decrease in the number of tunnels with the increase in the number of users is observable when N is over 40. Besides, the average number of tunnels increases when N is between 5 or 20; for higher user populations, both the number of tunnels and its average

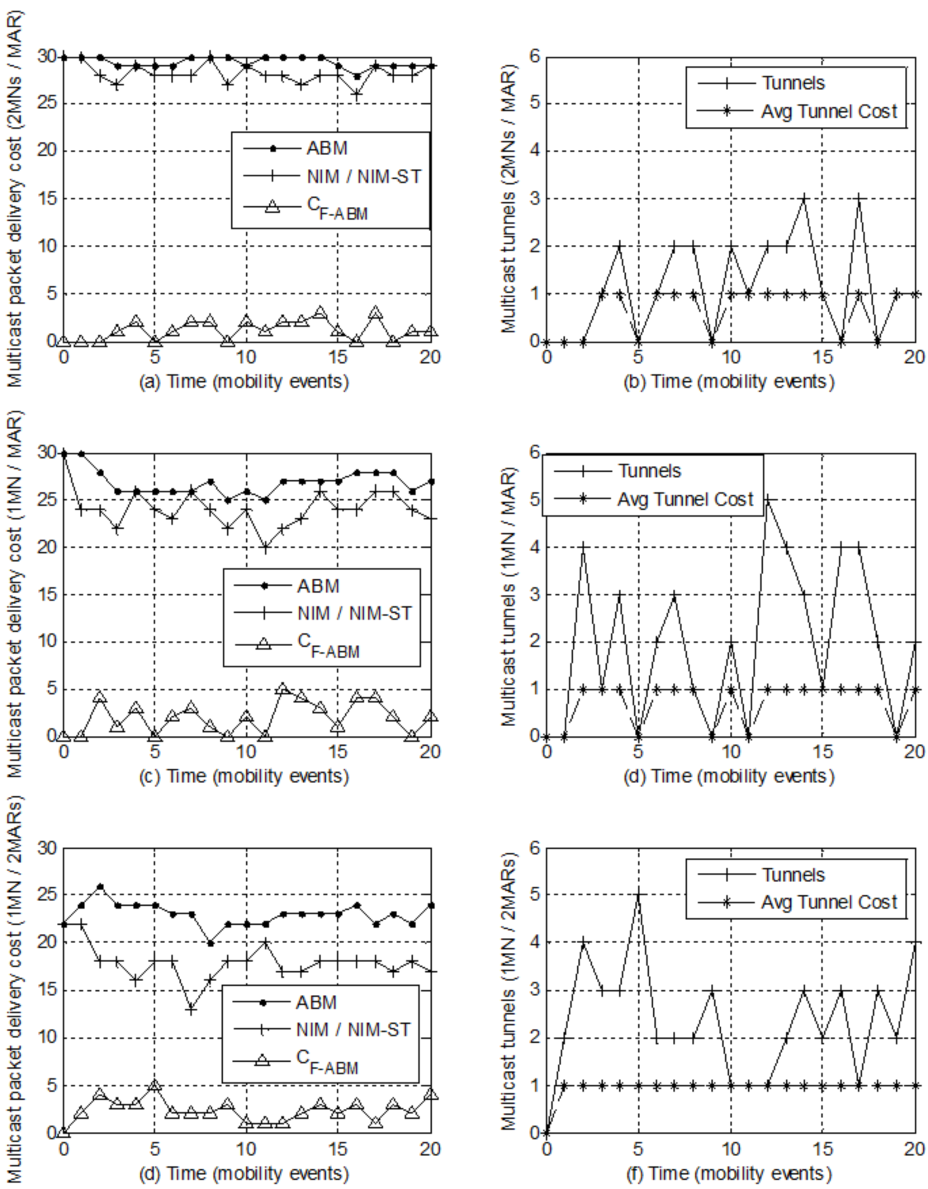


Figure 5.22: Multicast packet delivery cost over time in terms of total cost, number of multicast tunnels and average tunnelling cost (N = 32)

cost decrease, due to the higher probability to move to an already active MAR. The overall packet delivery cost increases with the user population, but its value tends to stabilize close to 30, as a consequence of the higher amount of MARs receiving multicast traffic natively and not via a tunnel: this corresponds to the maximum packet delivery cost when all MARs subscribe to the traffic natively. Considering that the maximum number of tunnels within a DMM domain is equal to M $\times$ (M-1) / 2 = 120, the amount of tunnels effectively required for multicast mobility of the considered subscription was considerably small independently of the amount of users. Additionally, the number of tunnels was inversely proportional to the average MAR occupation (Table 5.1).

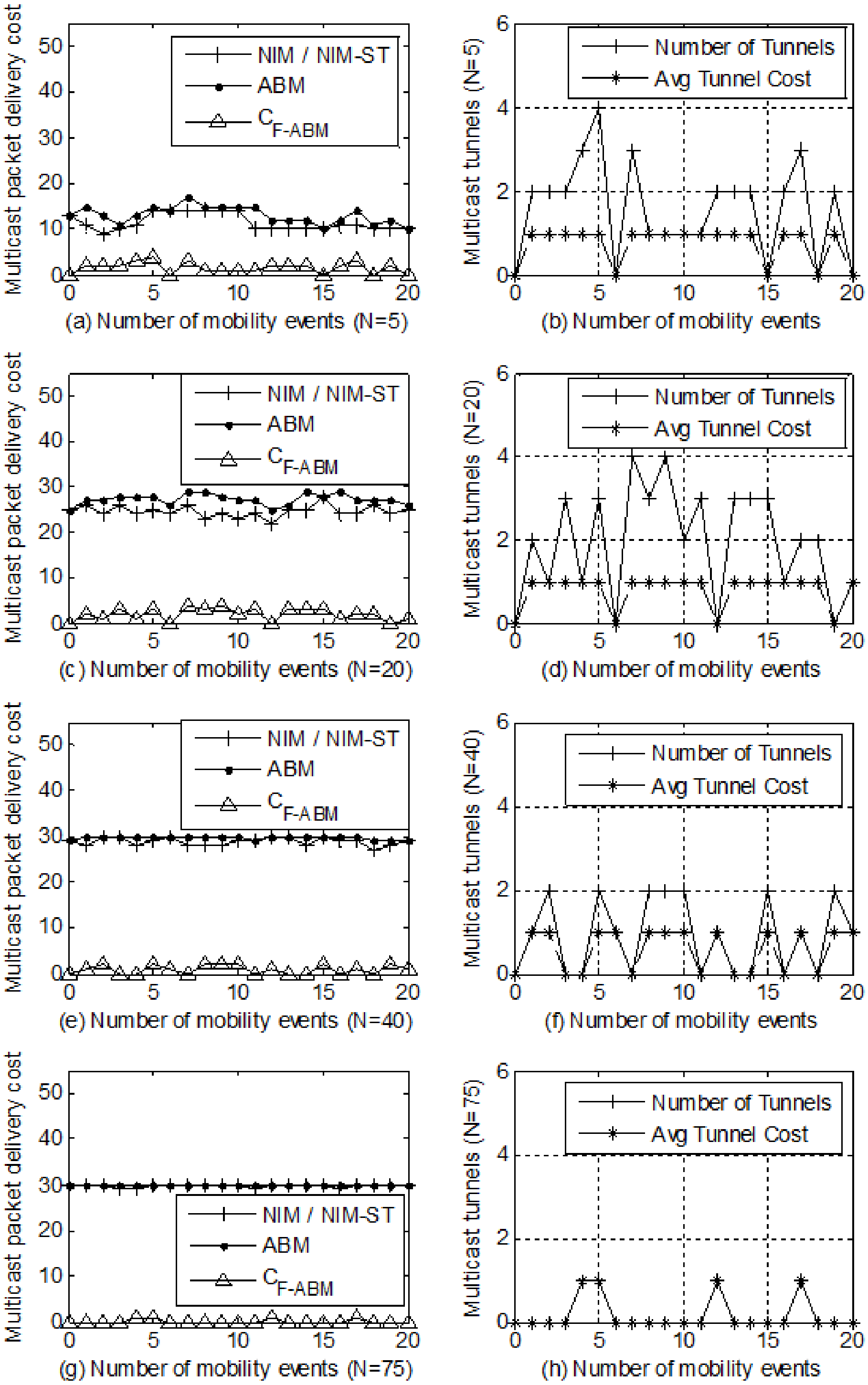


Figure 5.23: Total packet delivery cost and tunnel properties for N = 5, N = 15, N = 40 and N = 75

Average MAR occupation rate (%)			
N = 5	N = 20	N = 40	N = 75
25	68.75	87.5	93.75

Table 5.1: Occupation ratio variation with the number of users

5.4.3.4 Packet and Signalling Overhead

In this section we compare each of the schemes according to the signaling overhead and packet transmission overhead introduced by the multicast mobility protocol. Regarding packet overhead, when using generic tunneling, each packet is added 40 bytes due to the encapsulation header. ABM introduces additional overhead against NIM and NIM-ST, which transport multicast natively. For a typical Maximum Transmission Unit (MTU) of 1500 bytes, the overhead corresponds to an additional 2.7% of bytes. It is worth noting that this calculation does not take into account potential losses resulting from the tunnel MTU.

As for the signaling overhead, both NIM-ST and ABM have similar additional signaling overhead due to the exchanged mobility messages. The size of PBU and PBA messages depend on several factors, such as the embedded addresses (number of HNPs, subscriptions, etc.). Thus, the total mobility management protocol header overhead will vary. Assuming 20 bytes are driven from the information regarding a single multicast subscription after encoding, the Multicast Subscription option has a size of 28 bytes including the header. We estimate a PBA message would in such conditions have a total size of about 76+28 = 104 bytes. Consequently, the total overhead due to the PMIPv6-based signaling would be 104+76 = 180 bytes. Based on this assumption, from the resulting overhead, only 20 bytes in 180 (11%) are exclusively related to multicast operation, resulting in a low rate of information dedicated to multicast mobility. Considering the same size, if two or three subscriptions were included, the rate would be higher (20% and 27%, respectively).

5.5 Vertical IP Multicast Mobility support in DMM

Two important changes are occurring in current mobile networking. First, heterogeneous access networks are converging in all-IP architectures; second, the hierarchically centralized nature of mobile architectures is shifting into a flatter architecture, supported by research efforts such as DMM. In such a network environment, IP multicasting persists as a key enabler for efficient multimedia delivery. Its support is problematic though, due to the lack of schemes combining fast technology-agnostic HO with fast acquisition of channel subscription of mobile users over heterogeneous mobile wireless networks. Applications relying on IP multicast must use specific service interface calls whose listening state is both socket- and interface-specific. In inter-tech HOs, a different service interface must be used after HO.

Thus, without “intelligence”, the application cannot invoke the subscription on the target interface and receive the multicast channel(s), even if the target multicast network has already the subscription(s) of interest.

In this section, a solution for enhancing multicast receiver mobility in DMM is presented. The solution is empowered by a cross-layer design leveraging on IEEE 802.21 Media-Independent Handover (MIH) standard [28] and Context Transfer Protocol [93] for fast multicast HO over heterogeneous mobile wireless networks. IEEE 802.21 MIH is used as the main enabler for delivering multicast HO-related information between different access networks. Context Transfer Protocol is extended for network-side transfer of multicast subscription context, significantly reducing service disruption for mobile users.The next section presents the designed architecture, which integrated MEDIEVAL's wider-scoped architecture. MEDIEVAL high level architecture is later summarized in Section 6.2.

5.5.1 Reference Architecture and Operation

We employ a distributed mobility architecture consisting of multicast-enabled MARs, which provide access and anchoring functionalities to attached mobile users – following the previously presented DMM paradigm [37] – and multicast routing and context transfer functions - which replaces IP tunneling mechanisms for providing IP multicast mobility.

The solution consists of two main entities: Flow Manager (FM) and Connection Manager (CM), as shown in Figure 5.24. FM is responsible for managing the network-side resources for preserving the video session during mobility (i.e. adequate mobility schemes, network selection, activation of radio resources, etc), It resides at the MAR and is responsible for the management of data flows, being mobility management applied on a per-flow basis. The main focus of the FM within the mobility framework is to keep track of all data flows that traverse it and, either according to events from network entities, or implemented policies, manage those flows to provide the mobile user with the best possible service. For achieving this purpose, the FM leverages on its central position on the MAR, where it has a complete perspective of both the access network as well as the infrastructure near the access, enabling it to gather information from both perspectives to provide better decisions. All of its operation related to IP multicast are supported by the Multicast Mobility Module (MUME) - later described in more detail.

The other entity, Connection Manager (CM), manages radio interfaces and service interface calls during HO. It resides in the mobile terminal and is responsible for managing all required connectivity actions . The CM is a MIH user that interacts with the wireless access networks using 802.21 primitives in order to implement mobility, routing and flow handling. CM implements access network policies, selecting the preferred access interface to use or splitting the traffic along the multiple access networks available, when the terminal is able to use them simultaneously. These policies can be provisioned on the CM by multiple sources, namely CM GUIs, applications and operators (e.g., ANDSF).

This solution is tailored for multicast video services but can be applied to other services based on IP multicast requiring seamless mobility.

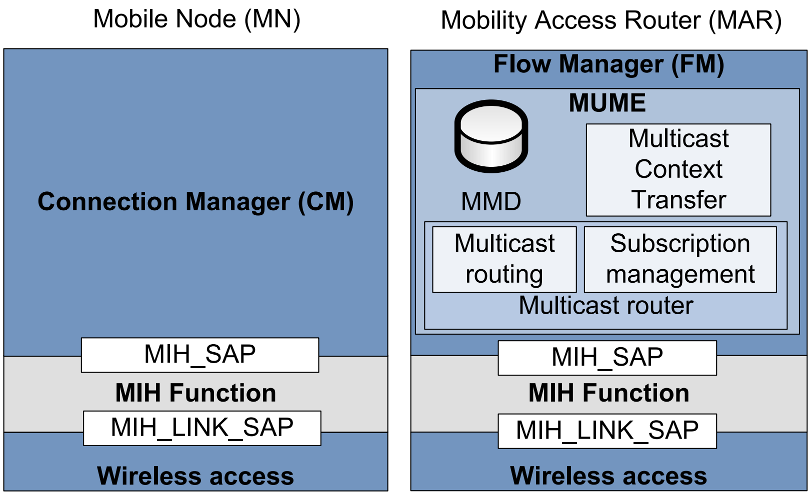


Figure 5.24: Functional modules in MN and MAR

Figure 5.25 shows the signaling procedure for the proposed multicast mobility solution. The IEEE 802.21 signaling is summarized by providing brief descriptions. In the depicted procedure, the target MAR (nMAR) is triggered either by the MN or the network (mobile / network-initiated HO) to activate multicast context transfer (MCXT) – which is extended from context transfer protocol for allowing multicast router to join the missing subscription proactively. CM is triggered to activate and use the new interface for the subscription of the multicast session, seamlessly receiving the session packets from the new interface.

5.5.1.1 Multicast Mobility Engine (MUME)

MUME module was defined within MEDIEVAL's architecture, which is later presented in Section 6.2. It is collocated within the MAR, as shown in Figure 5.24, and is the core piece responsible for managing mobility of IP multicast flows for terminals or services which depend on it, providing mobility solutions for those multicast flows which require it. For doing so, it effectively depends on four main functions: i) multicast group management function, ii) multicast routing function, iii) mobility management function and iv) context transfer function. The first function refers to the multicast group management operations and information storage, realized using MLDv2 router-side functions (through MLDv2 Queries) with the mobile hosts. The multicast routing function corresponds to the multicast routing protocol stack of the node, which in the considered scenarios will be PIM family protocols. As for the mobility management function, it resembles the mobility protocol stack, which

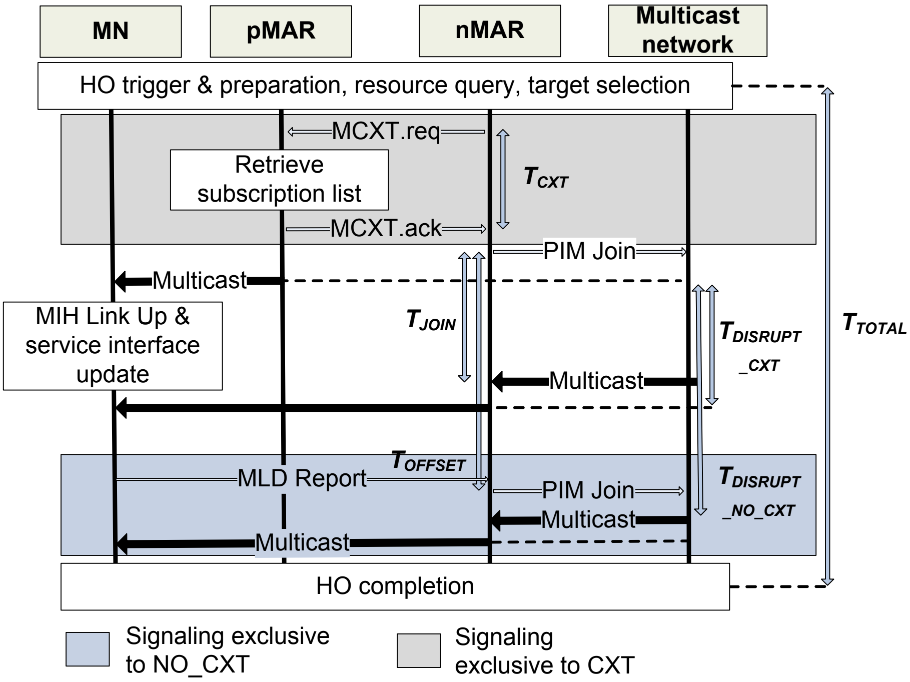


Figure 5.25: Proposed signaling

corresponds to different functionality depending on the role of the multicast host. Finally, the context transfer function is responsible for preemptively modifying the multicast tree in cases of multicast receiver mobility.

The main operations supported by MUME can thus be listed as:

- Multicast explicit tracking function: MUME keeps per-MN information regarding their multicast subscriptions.

- Multicast context transfer: By making use of the explicit tracking function, MUME stores always up-to-date subscription information, which is the basis for correct responses provided by the queried MUME (from previous MAR), and allows tunnel-free multicast mobility – not needing MIPv6-related stack.

- Tunnel-based receiver mobility: In some scenarios, the network operator may wish to take advantage of the DMM mobility tunnels used for unicast traffic, like the solution presented in Section 5.3. In such cases, MUME allows the setup of the MRIB entries based on the tunnel's endpoints addresses.

- Multicast source mobility: MUME enables transparent source mobility, achieved by preemptive dissemination and configuration of the RP address - which acts simultaneously as the mobility tunnel endpoint - at each MAR to which the mobile source moves. The MAR assigned as RP acts analogously to the anchoring MAR in unicast DMM mobility operation. The referred preemptiveness is consequence of enhancements to regular MIH operation, speeding up the multicast mobility process. This and other schemes enabling multicast source mobility can be referred to in Section 5.6.

Data Structures and Messages

MUME decisions are done based on the following information contained in Multicast Mobility Database (MMD):

1 – Mobile Terminal ID: this is necessary by MFM for properly identifying the MN during mobility processes;

2 – Mobile Terminal Address: required for storing updated mapping between terminals and associated subscriptions;

3 – Multicast subscriptions (aligned with the structure of MLD multicast information);

4 - Counter with the number of listeners per IP multicast channel: For supporting the explicit tracking function, enabling e.g. the identification of the last subscriber of a group. This information is in particular essential for proper multicast context transfer operation. MUME module core operation is the control of the Multicast Router.

The interactions between MUME and the Multicast Router and with other MUMEs may be split as client-based and server-based functionality. As a client, its main functions include:

- Request multicast context transfer from other MUMEs;

- Request multicast tree updates to the local MR: joining of multicast trees (in case of MN multicast service initiation or arrival due to mobility) and departure of multicast trees (in case MN moves to another MAR;

- Request tunnel based multicast mobility establishment through another MUME.

Its server-based functions include:

- Reply to the requesting MUME with multicast context;

- Complete mobility tunnel establishment for multicast content reception.

5.5.2 Quantitative Analysis

The current section mainly serves the purpose of evaluating MUME performance for the goal of achieving improved vertical IP mobility of multicast receivers in DMM scenarios. Figure 5.26 depicts the evaluation scenario. In the experiment, a multicast receiver roams from a contention-based wireless access (WLAN) to a emulated coordination-based access (LTE) while playing a video received through IP multicast. In the considered scenario, the multicast receiver mobility is supported by a multicast context transfer process. Thus, this is the main focus in the undertaken evaluation of MUME. For a clearer observation of the achieved performance, we evaluated the proposed scheme (dubbed CXT) against an identical one where multicast context transfer is not applied, denoted as $NO-CXT$. Basically, this alternative scheme also leverages on the operation between CM and FM but follows standard MLD signaling procedure. Notice that this approach is already an improvement over the default one, where IEEE 802.21 mechanisms are not provided. The corresponding signaling is also depicted in Figure 5.25. The results were achieved by averaging a 10 times run, and using the following open-source softwares:

- MRD6[4] : provides multicast routing and subscription functions (PIM and MLDv2).
- ODTONE[5] : an open-source implementation of IEEE 802.21.
- OpenAirInterface[6] : an open-source software used for emulating the LTE access.

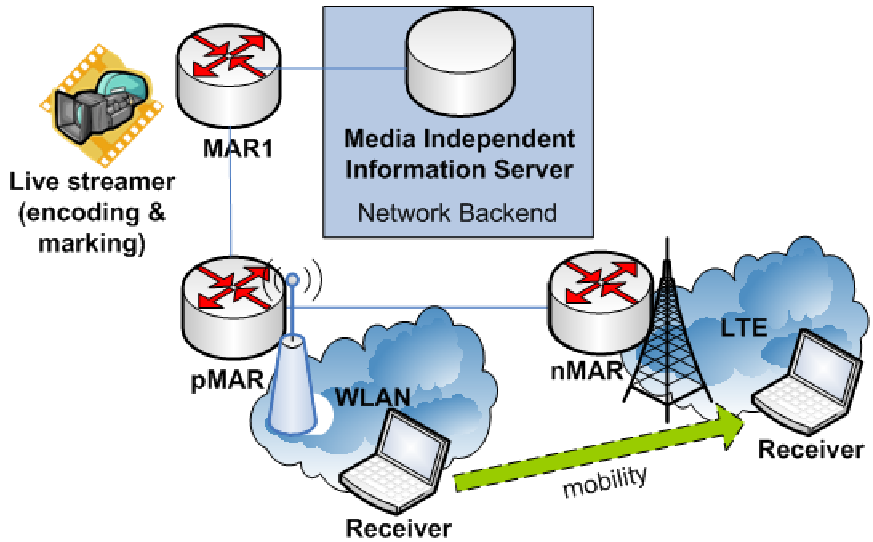


Figure 5.26: Handover scenario

5.5.3 Performance Evaluation

The HO latency of the solution was evaluated, and is broken according to relevant parameters; T_{D-CXT} is the time between the reception of the last and first data packets over the previous and new links, respectively. $T_{D-NO-CXT}$ is the equivalent to the latter but applied to $NO-CXT$ solution. T_{CXT} is the time spent in the multicast context transfer signaling. T_{JOIN} is the time taken for receiving the first IP multicast packet at the router after sending the PIM Join. In the scenario, the upstream multicast router from the MARs (i.e. the nearest router subscribing the multicast channel) is one hop away. T_{OFFSET} is the time between the transmission of the PIM Join in $MCXT$ and $NO-CXT$. Additionally, T_{TOTAL} is defined as the total latency required for the IEEE 802.21 signaling.

[4]https://fivebits.net/proj/mrd6/
[5]http://helios.av.it.pt/projects/opmip/
[6]http://atnog.av.it.pt/odtone/

Table 5.2 shows the measured latencies corresponding to all defined delay factors, which are as follows. $T_{D-NO-CXT}$ was larger than 1s for all experiments and significantly larger than T_{D-CXT}. The improvement of using $MCXT$ over MLD is further shown by T_{OFFSET}, which was near 0.5s. The time that nMAR took to join the multicast tree (T_{JOIN}) was 12.8ms.

Delay factor	**Measured value (ms)**	**Std Deviation (ms)**
T_{JOIN}	12.80	5.10
T_{TOTAL}	590.46	0.03
Specific to MCXT		
T_{CXT}	2.34	0.89
T_{D-CXT}	127.10	2.70
Specific to NO_CXT		
$T_{D-NO-CXT}$	553.67	570.54
T_{OFFSET}	429.22	230.01

Table 5.2: Handover latencies for vertical multicast HO

5.6 Multicast Source Mobility in DMM

We have witnessed a paradigm shift with regards to content transport over the Internet; while previously most of user connectivity was dedicated to data receival, now a significant portion of online tasks include the dissemination of private data, such as photos and video. With this new trend, challenges with respect to the support of mobile upload gain particular attention, ranging from the uplink properties to the application behavior. The wireless transmission of IP multicast towards the network adds further requirements, such as those referred in Section 2.4.3.2.

This section presents an analysis of use cases and potential solutions considering mobile multicast source and associated options. Similarly to the work on receiver mobility, this section identifies two main decisions towards source mobility support: the multicast functions employed at the MAR and the delivery method. The different options are well scrutinized in the referred use cases, and characterized in terms of advantages and disadvantages.

5.6.1 Multicast Operation at the MAR

5.6.1.1 MLD Proxy Deployment at MAR

Source mobility support is known to lead to service disruption problems impacting all the multicast tree, in particular if SPT is active. The utilization of MLD Proxy in PMIPv6 environments is proposed in [94], being the upstream interface always configured towards the fixed anchoring entity: the LMA. The utilization of MLD Proxy carries the previously referred advantages, as ease of deployment and operation lightness.

To allow the source to transmit multicast content to the multicast tree in a DMM framework, the MLD Proxy should configure its upstream interface towards a router from the multicast infrastructure. Depending on the network topology, it may also be configured towards a MLD Proxy placed on a neighbor MAR. In case of mobility of the multicast source (Figure 5.27), the MLD Proxy may operate similarly to the receiver mobility case. Concretely, as the traffic from the multicast source arrives through one of MLD Proxy's downstream interfaces, the traffic is forwarded through the uplink interface towards the anchoring MAR.

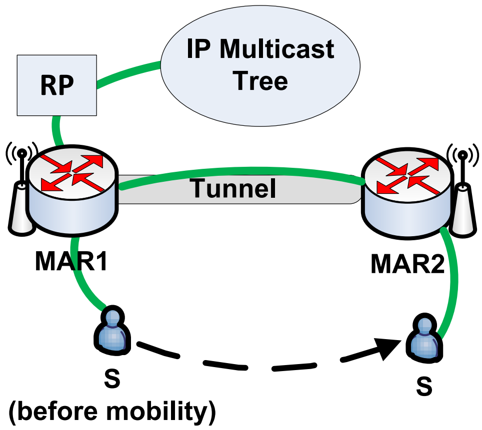


Figure 5.27: Multicast source mobility

When a source moves to new MAR while keeping a multicast session, multicast data will be sent through the mobility tunnel between the two MARs (Figure 5.28 (a)). If a receiver R attaches to the same MAR (MAR2), it will receive the multicast data through multicast infrastructure, following the configuration of MLD Proxy. Hence, the multicast data is routed non-optimally between the source and receiver, going from the current MAR to the anchoring one, to the multicast routing tree, and then back to current MAR again before reaching the receiver.

A similar problem occurs in the opposite process, i.e. if a multicast source starts transmitting multicast content at a MAR, and a receiver moves to the same MAR while receiving the source's content (Figure 5.28 (b)).

Although, when the multicast source does not move, and the receiver is within the same MAR (MAR2), the traffic will be optimally sent to the receiver without the need to go through native multicast infrastructure. As the traffic reaches the MLD Proxy via the downstream interface to which the source is attached, it will be sent through the downstream interface to which the receiver sent the MLD Report. However, if the source and the receiver move to different MARs, the traffic will traverse the following non-optimal path, even though they share a common anchor: Source -> MAR1 -> MAR2 -> Multicast Tree -> MAR2 ->

MAR3 -> Receiver. This problem is depicted in Figure 5.28 (c).

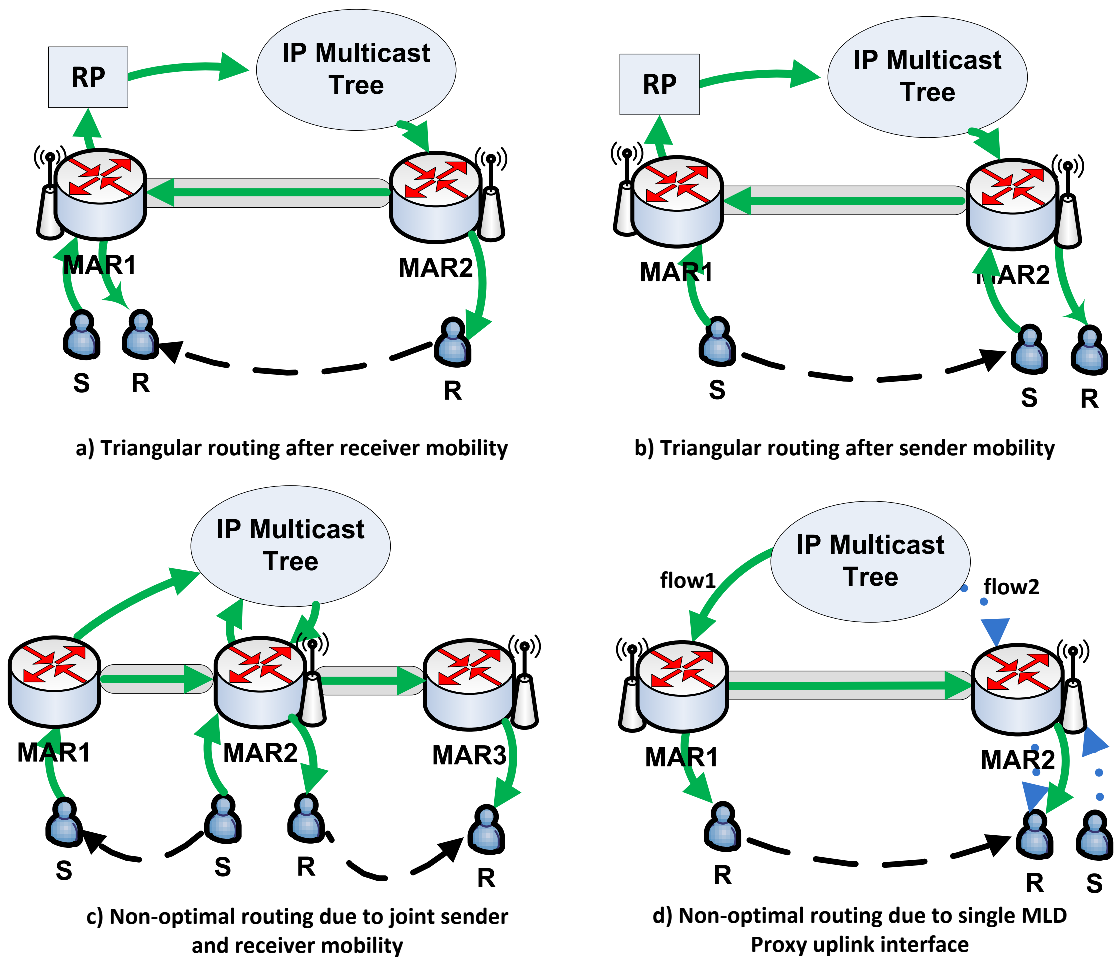


Figure 5.28: Several cases of non-optimal routing while applying MLD Proxy

Requirement 1 (REQ1) from [6] refers that "IP mobility, network access solutions, and forwarding solutions provided by DMM MUST enable traffic to avoid traversing a single mobility anchor far from the optimal route." Applying MLD Proxy, when a MN subscribes to a new multicast session with existing multicast mobility session, the Aggregated MLD Report containing all the MN's multicast subscriptions will be sent from the current MLD Proxy through the same uplink interface, i.e. towards a single multicast mobility anchor. This results in some of previously identified issues, such as non-optimality in the path that both the subscription and multicast traffic traverse. As previously seen for mobile multicast receiver use cases, it can be stated that the MLD Proxy nature doesn't comply with the aforementioned requirement, leading to the subscription of any multicast flow using the same multicast mobility data path.

This problem is depicted in Figure 5.28 (d), where both multicast flow 1 and flow 2 (dashed line) reach MAR2 from MAR1, being flow 2's optimal routing path affected by the mobility status of the MN, and in particular by the order in which the multicast flows were subscribed. While this issue is not exclusively related to mobile multicast sources, its' impact in the routing is more obvious when considering one.

5.6.1.2 Multicast Router Deployment at MAR

Considering that full multicast routing stack is deployed at MARs, when a source starts transmitting multicast traffic, the content will be encapsulated in PIM-Register messages, and sent towards a RP - statically configured or discovered through a Bootstrapping Router (BSR). In DMM, the RP can be a core MR or a MAR, including the anchoring MAR. The RP's SPT and each of the DR's SPTs may then be created. When the source moves, the MR of the new MAR (N-MAR) will create the state for the new multicast group, and the traffic will be forwarded using the tunnel to the previous MAR (P-MAR) until reaching the RP - unless the RP is actually the P-MAR -, and is then sent down the RPT. Again, the creation of the SPTs will typically be triggered following PIM-SM regular operation.

In case the RP's SPT is built before the mobility process, it will be destroyed due to mobility, and the tree construction process will be reinitiated at the new MAR. Also, in case the SPT between the listener's DRs and the source's DR is being used, mobility will reset the PIM process to the RPT stage. This means that each source mobility event results in increased signaling overhead and delay, as consequence of the multicast routing convergence (i.e. Phase 2 and Phase 3 from PIM-SM operation). Moreover, non-optimal routing occurs when the RPT is used. When a source moves to a MAR where multicast receivers are subscribing its channel, the multicast traffic will always reach the N- MAR by going through the RP, just like in the MLD Proxy case (Figure 5.28 (c)).

Using PIM-SM in DMM scenarios there is a trade-off between the routing non-optimality of RPT and the non-efficient consequences of frequent SPT establishment. It is important to note that this impact is magnified the more receiver's DRs are receiving the multicast channel(s).

5.6.2 Subscription and Routing Origin

A high level description of different alternatives for multicast source mobility support when applying multicast router at the MAR is herein provided. The core design option relates to the way multicast operation interacts with mobility-related functions. When applying a tunnel-based scheme, multicast mobility support is handled similarly to unicast mobility support, while by using direct routing multicast traffic is operated separately, relying on the multicast native infrastructure. The different ways that multicast traffic flows for the considered solutions space are numbered from 1 to 4 in Figure 5.29, where the application of RPT relies on a RP, while SPT refers to the shortest multicast routing path.

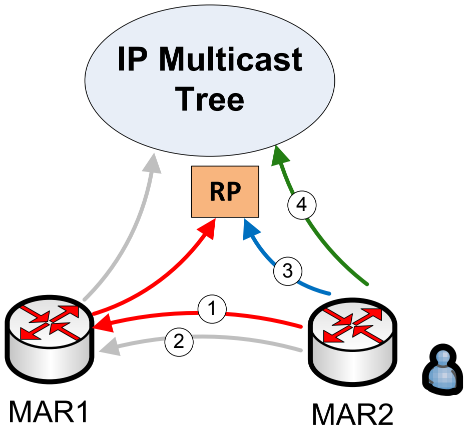


Figure 5.29: Multicast delivery options in source mobility

Option 1 - Tunnel-based Scheme and RPT

This approach allows the preservation of the RPT, and the global source address to be kept constant during the mobility of the MN. Depending on the domain's size and the distance the MN moves from its anchoring MAR - a function of the Localized Mobility Domain (LMD) size -, it may lead to non-optimal routes. Comparatively to PMIPv6, in DMM the tunnel's concentration per anchor is decreased, just like in unicast mobility support, avoiding the single-point-of-failure and overhead problems. When the anchoring MAR starts receiving the multicast content, it should build a (S,G) source-specific entry for that content, not going through the (*,G) PIM Register and PIM Stop phases which would happen when sending the content through a DR before arriving to the RP. After the MN starts moving, the S-MAR, adds a new downstream link with up-link to the A-MAR.

Option 2 - Tunnel-based Scheme and SPT

Just like Option 1, the traffic flows through a common point, the A-MAR. This means that the theoretically main advantage relatively to the use of RPT, the forwarding path optimization, is virtually eliminated. This happens because each (S, G) Join message arrives to the RP, the first router having the (S,G) state, resulting in the same end-to-end path as with the RPT. Not only is this advantage eliminated, but it would also bring unnecessary signaling proportionally to the amount of receiver's DR, corresponding to the PIM-Register / Stops they would need to send.

Option 3 - Direct routing Scheme and RPT

This method allows traffic to be transported using optimal routes from the source to the multicast tree, and avoids unnecessary replication independently of the mobility support scheme at the receivers. Although, it imposes rapid discovery and routing to RP, so that the arrival time of the first data packet from the current AR to the RP does not result in session discontinuity.

Option 4 - Direct routing Scheme and SPT

The application of this scheme in mobility scenarios causes severe problems to the multicast session, particularly due to the constant SPT reconstruction for high-mobility source multicast nodes, as the source IP address changes with the MN address reconfiguration. This can be considered as the baseline solution, where no mobility operation occurs, meaning that a new HNP is used when sending traffic at the new MAR. This implies that the receivers must subscribe to the new channel at each source mobility event. It may therefore be stated that both direct routing schemes are better aimed for applications without IP address continuity requirements.

5.6.3 Trade-offs between MLD Proxy and Multicast Router

This subsection intends to present an overview of the differences between the two described multicast functions. Table 5.3 summarizes the previous analysis, globally depicting the differences between MLD Proxy and MR over DMM. Note that this comparison includes both the multicast source and receiver perspectives. The meaning of each of the analyzed characteristics is now listed:

- Lightweight: this entry reflects whether the deployed multicast feature has a resources-wise lightweight operation.
- Optimal routing: this entry reflects whether optimal routing is assured.
- Efficient distribution: this entry reflects vulnerability to multicast traffic replication.
- Distributed anchoring: this entry assesses whether for a single MN, different multicast streams can be anchored at different mobility anchors or not.
- Seamless mobility (receiver-only): This entry reflects whether IP mobility is seamless from the point of view of the mobile receiver's application.

- Signaling overhead: This entry assesses the amount of signaling that the IP mobility of a MN represents. This signaling may be relative to the mobility protocol or the general signaling, such as that resulting from the multicast routing convergence.

Feature / Function	MLD Proxy	Multicast Router
Lightweight	Yes	No
Optimal routing	No	Yes (using SPT)
Efficient distribution	No	Yes
Distributed anchoring	No	Yes
Seamless mobility	No	No
Signaling overhead	Low	Average

Table 5.3: Comparison between MLD Proxy and MR deployment at the MAR

5.7 Concluding Remarks

The early consideration of IP multicast in Distributed Mobility Management design is mandatory, as reflected in RFC7333 [6]. This chapter started by exploring possible use cases, by deriving from Base solution supporting IP multicast in PMIPv6, being the primary objective to assess potential issues, how they relate or differ from previous mobility management approaches. Most of the same problems were observed, although at a different scale, as a result from the mobility anchoring distribution. One of the main remarks was thus the observation of MLD Proxy limitations, and its lack of compliance with DMM scenarios, as a result from its "single upstream interface" design. The utilization of Multicast Routing was thus explored, which led to the design and evaluation of a novel framework (D3M), addressing all of the aforementioned issues through the availability of multiple modes.

Chapter 6

Beyond layer 3 Optimizations for Mobile Multicast Video Delivery

The mechanisms explored in the previous chapter were mainly focused in Layer 3 optimizations or in interactions between external entities and network modules for triggering L3 operations. Although, technologies from other layers must be improved for several reasons, spanning from optimized radio technologies to applications more actively taking advantage of IP multicast benefits. Namely, the adoption of a cross-layer design implies the coordination and extensions to the multiple involved layers, improving feedback capabilities to respond to the multiple events (e.g. mobility, congestion, bitrate adaptation). Most of the contents of this Section relate to concepts researched within MEDIEVAL and its achievements.

6.1 Introduction

Previous chapters focused on enhancing IP multicast and IP mobility, two key mechanisms from TCP/IP communications, and bringing both mechanisms together towards the provision of improved video delivery in mobile environments, with a particular focus on HO events. Thus, the employment of IEEE 802.21 as a middleware enabling cross-layer interactions was prospected as a possibility to improve HO performance of multicast users, allowing better informed decisions, aware to service and access technology type, as well as other information. Nevertheless, other advances related to the base support of multicasting and broadcasting mechanisms are also relevant and necessary. While some parameters such as scalability may be subject to further improvement, other challenges such as improved multicast rate transmission over wireless remain unaddressed. We define a list of requirements to be fulfilled by potential mechanisms as follows:

- **R1**: improved scalability of video delivery, taking into account the diversity of devices and user profiles;

- **R2**: improved bitrate performance of multicast traffic in WLAN and cellular scenarios;
- **R3**: efficient mobility signaling in scenarios taking advantage of IP multicast, such as group-based mobility

As referred, this chapter explores additional mechanisms which may be coupled with work from previous sections on IP multicast mobility under DMM, both including orthogonal mechanisms and protocols associated to the other communication layers. The considered mechanisms were defined within MEDIEVAL architecture, which is finally introduced in this chapter. Summarizing, the chapter is dedicated at detailing such mechanisms and present scenarios which demonstrate how they resolve the aforementioned requirements. Sections 6.3 and 6.4 describes two solutions addressed cooperatively in MEDIEVAL's architecture, while Section 6.5 is the direct result of a co-authored publication, although not included in MEDIEVAL's specification.

Chapter Contents

- **Section 6.2** presents MEDIEVAL architecture, a solution tailoring mobile networks for enhanced mobile video delivery by means of a cross-layer design, where relevant network modules directly exchange information for the sake of the QoS and QoE of the video services.
- **Section 6.3** concerns requirement **R1**, and explores the usage of SVC within multicast DMM scenarios, presenting conceptual approaches.
- **Section 6.4** describes how requirement **R2** may be answered by means of IEEE 802.11aa, which is especially devised for improving the performance of multicast services in 802.11 protocol
- **Section 6.5** proposes the extension of IEEE 802.21 signaling to resolve requirement **R3**, by providing the signaling of a group of users by means of a minimal number of messages.
- **Section 6.6** describes different use cases where the aforementioned optimizations are utilized, either separately or combined. This section intends to demonstrate how the proposed extensions would apply in concrete scenarios.

6.2 The MEDIEVAL High level Architecture

MEDIEVAL follows a vision where the future Internet architecture should be tailored to efficiently support the requirements of video traffic, and that specific enhancements for video should be introduced at all layers of the protocol stack where needed. The proposed architecture follows a cross-layer design that, by exploiting the interaction between layers, can raise performance to values unattainable with individual developments. The technology developed by the project takes into account the requirements of network operators for commercial deployment, and aims at improving the QoE for users while reducing the costs for operators. MEDIEVAL technology was developed in a testbed that serves as a proof of concept of the project results, and was the basis for some commercial deployments. The architecture consists of the following sub-systems: Video Services Control, Wireless Access, Mobility Management and Transport Optimizations.

6.2.1 Video Services Control

The video service control subsystem is responsible for linking the services and the underlying network delivery entities. It aims at enabling a reliable video delivery over an evolved mobile network, offering improved resources utilization and an enhanced user experience using a cross-layer set of interfaces from Video Service Control to the other sub-systems. This subsystem also proposes a set of innovative service controllers to support new video applications leveraged by the social networking trend, hiding the service management issues from the multimedia applications, with QoS support and while improving resource utilization and application flexibility. Last, but not least, the subsystem provides reliable and adaptive content delivery in inherently unreliable networks, maximizing the users' quality of experience, taking into account the network dynamics as well as other potential factors such as monetisation schemes or user differentiation, for the diversity of video-rich applications.

The video service control is mainly responsible for:

- Service provisioning which is further segmented into services, contents and user attributes.
- Session management and network monitoring, from sessions initiation to provision of ongoing measurements of the underlying networks conditions.
- Control of video content generation and delivery, based on session measurements and network events, like HOs or resource changes in the network. It is also responsible for providing the network with sensitivity graphs, to allow network adaptation, such as resource allocation to different flows.
- Content adaptation, content protection and packet marking, in order to signal the underlying networks about packet prioritization

6.2.2 Wireless Access

The main reference model of the project consists in an operator supporting connectivity through heterogeneous access technologies. Thus, the objective of the wireless access sub-system is to describe the architectural solutions envisioned to provide enhanced video delivery in the last (wireless) hop, mainly focusing on novel access techniques. According to how they make use of the wireless medium, we can classify access techniques into contention-based, such as the IEEE 802.11 standard for WLANs, and coordination-based, e.g., LTE-A. For each access category, the project aims at developing novel mechanisms to enhance video transmission over these wireless accesses, providing a satisfactory QoE and enabling cross-layer optimizations in the interaction with upper layers. In order to encompass this optimization, cross-layer signaling is implemented between the lower layers of the wireless access and the video application and services, as well as with mobility services. This is accomplished by the definition of an abstraction layer and its associated functions, together with some ad-hoc features designed to further enhance the video flow transfer over the air.

6.2.3 Mobility Management

The Mobility Management sub-system employs a DMM approach, where the anchors are at the very edge of the network. The architecture provides a hybrid operation, where: network-based mobility management (i.e., PMIPv6-alike) is used when possible, and client-based mobility management is used otherwise (e.g., between different domains). Moreover, the DMM solution chosen for network-based mobility management is classified under the Fully distributed category, which consists on removing any central anchor both for the data and control planes. The critical point in the fully distributed approach is that the MARs need a mechanism to learn about MN's movements to address the mobility update to the correct MAR. Differently from PMIPv6, where a database containing the users' mobility sessions is stored in the LMA, in the fully distributed approach each MAR stores only the database's part related to the currently attached MNs, thus this information may be incomplete if the MN has been roaming among the access networks.

The access network is organized in Localized Mobility Domains (LMDs) in which a network-based scheme is applied. Users are expected to be most of the time roaming within a single LMD, but, for those cases where this is not possible (e.g., roaming to a network owned by a different operator or running a different mobility support scheme), a host-based DMM approach is followed. MAR was introduced in order to integrate both approaches, enabling MNs to simultaneously have sessions managed by the two methods. MAR is thus a network entity able to play the role of plain AR, home agent, local mobility anchor and mobile access gateway on a per-address basis.

MEDIEVAL also supports Network Mobility (NEMO) solutions, i.e., a MAR moving within an LMD. This entity is called Mobile MAR (mMAR) and it is supposed to gain connectivity from a fixed MAR. The MARs need to collect pieces of information about MNs by contacting formerly visited MARs. To face this issue, a solution following a Make-Before-Break approach for the HO operation was developed; this solutions integrates Layer-2 and Layer-3 mobility procedures within the same framework to assist and drive the HO, being IEEE 802.21 for Media Independent Handover the chosen protocol for this purpose. Due to the video-centric nature of the project, multicast traffic delivery and content distribution aspects are fully supported and integrated in the same mobility management solution. As presented in previous chapter, MUME module is the entity designed for providing IP multicast mobility support.

6.2.4 Transport Optimization

The Transport Optimisation subsystem provides optimized video traffic in the mobile operator's core network through intelligent caching and cross-layer interactions. The main objective is two-fold: i) reduce the load on the operator's backbone, ii) while still providing a satisfactory QoE to the users. The first goal is addressed by establishing a mobile CDN, with a special focus on the selection of optimal cache locations and node selection based on costs like "network distance". This means that MEDIEVAL aims at service placement (i.e., finding optimal locations for deploying the CDN nodes considering, e.g., various cost metrics, the design of the core network and operator policies), content placement (i.e., the optimal distribution of content among the CDN nodes), and content routing (i.e., choosing from the set of CDN nodes, providing the desired content, the node or subset of nodes that minimizes streaming costs). The second goal is addressed by providing proper optimized resource allocation and traffic engineering techniques in order to increase as much as possible the user perceived quality (QoE) within the given resources in the network. Therefore, the system performance is evaluated in a network-wide context using cross-layer optimization techniques. Information is collected from the other MEDIEVAL subsystems, like MAC and buffer states from the Wireless Access, QoE-based data about video sensitivity from the Video Services, and HO candidates from the Mobility subsystem.

The specified architecture and its modules is depicted in Figure 6.1.

6.3 Achieving Scalability with Multicast Video Layering

Previous works have explored SVC as a method to deliver layered multicast video [129], with its advantages over simulcast delivery coming from the protocol inherent scalability options (in terms of spatial, temporal and quality-driven levels), which is turned into efficiency. A

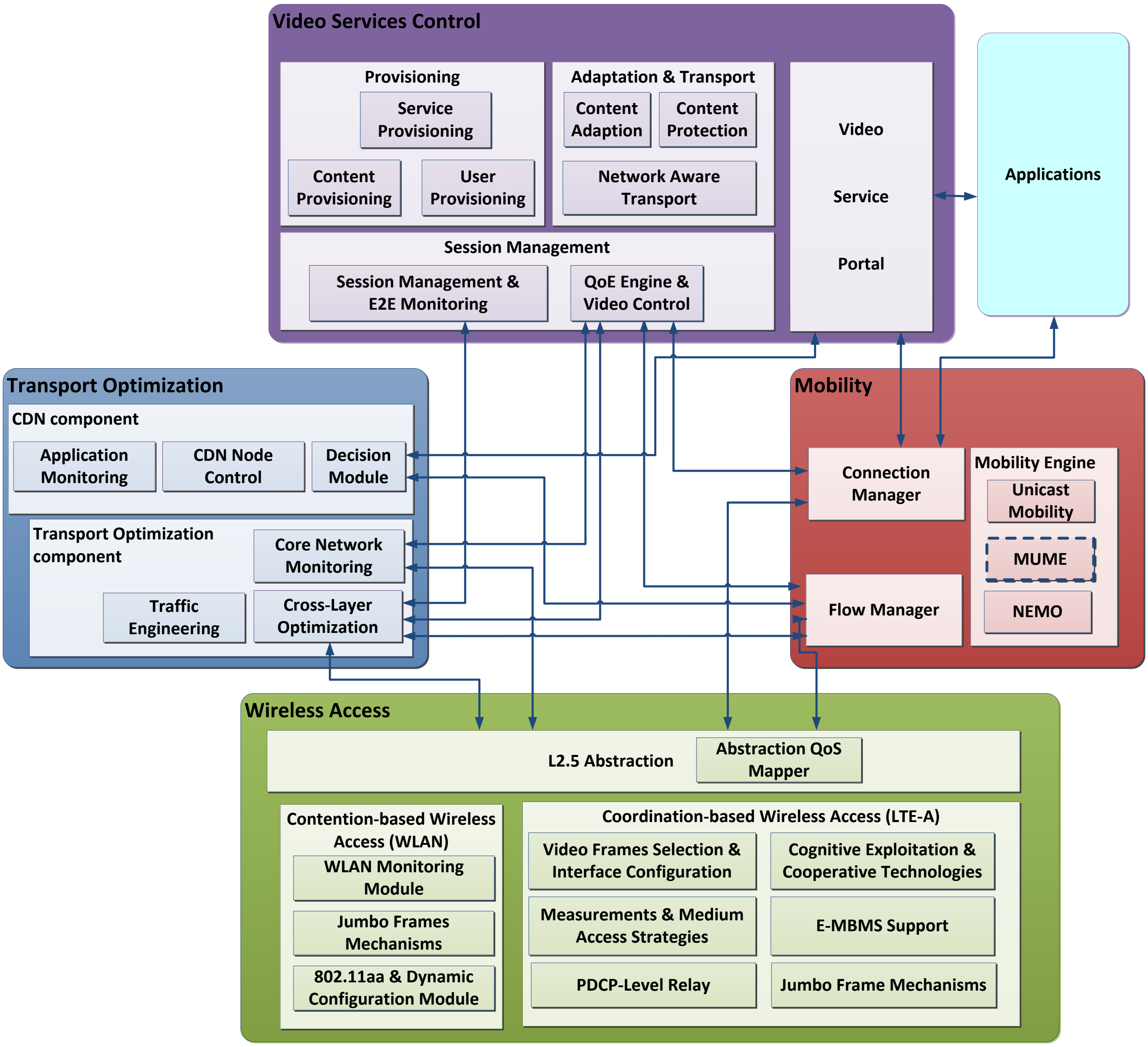


Figure 6.1: MEDIEVAL Architecture

SVC stream is constructed from NALUs (Network Abstraction Layer Units) to represent a part of the picture's encoded bit stream, which belong to a single layer. The stream is constructed from a basic layer which is not dependent on any other layers, and enhancements layers, dependent on lower layers. Due to this scalable property, the SVC encoding is an ideal technique for providing multimedia multicasting to heterogeneous networks and devices as explained here and in section 3. In order to transport multicast SVC in RTP packets over heterogeneous environments, Multi-Session Transmission (MST) was specified in [130]. Using this mode, multiple RTP sessions are used to carry the SVC data. Albeit, depending on the application requirements, this may translate into transporting one layer per RTP session or encapsulating multiple layers in one RTP session, by using a Media-Aware Network Element for aggregating RTP sessions into a single RTP stream for each client (unicast) or group (multicast). Besides, different layer combinations (base layer only, enhancement layer(s) only or base and enhancement layer(s)) are possible. Additionally, distinct packetization modes exist.

6.4 High-Performance delivery of IP Multicast with IEEE 802.11aa

Regarding contention-based access, the goal is to compute the MAC parameters which achieve the optimal performance taking into account the cross-layer packet marking for every video flow. However, current IEEE 802.11 standard [131] does not allow for intra-flow differentiation (e.g., prioritize an SVC video layer over another), and multicast transmission, namely No Ack/No Retry, offers poor performance. It imposes low rates and provides reduced reliability against collisions or interferences due to the lack of MAC-level recovery procedures (see Figure 6.2 (a).

To address the previous limitations, IEEE 802.11aa Task Group [1] has: (i) defined the Groupcast with Retries (GCR) service which increases the reliability of group addressed frames (multicast groups) by employing Unsolicited Retry (UR) or the extension of Block Acknowledgment mechanism defined in IEEE 802.11e for multicast; (ii) adapted the already existing Directed Multicast Service (DMS) defined by IEEE 802.11v to group addressed frames; and (iii) defined a Stream Classification Service (SCS) which enables classification using L2 and/or L3 signaling (hence leveraging MEDIEVAL cross-layer packet marking) and allows for intra-access category Traffic Stream (TS) prioritization.

GCR UR preemptively retransmits a frame one or more times (up to a certain lifetime limit), to increase the delivery probability at the stations without introducing the associated overhead of an acknowledgment mechanism (see Figure 6.2 (b). DMS consists on the multicast to unicast conversion (as illustrated in Figure 6.2 (c) for two group members). Hence, those frames transmitted to a multicast address are individually transmitted to each of the associated stations that joined the multicast group up to a retransmission limit. This mechanism provides high reliability but it has large scalability constrains as the required throughput increases with the number of group members.

GCR Block Ack transmits bursts of frames to a group address and sends BlockAck Request frames in turns to each GCR group member which replies with BlockAck frames. There are two possible GCR Block Ack mechanisms: Immediate Block Ack in which the recipient of a BlockAck Request replies immediately with a BlockAck frame (Figure 6.2 (d)),

[1]Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications - Amendment 3: MAC Enhancements for Robust Audio Video Streaming, IEEE Amendment 802.11aa, 2012

and Delayed Block Ack, in which after receiving a BlockAck Request the recipient starts a backoff process before sending the BlockAck frame. With the Delayed Block Ack, BlockAck management frames are acknowledged with ACK frames (Figure 6.2 (e)).

The performance optimization for contention-based wireless access also implies the design of an algorithm to choose the most appropriate multicast mechanism for a given scenario.

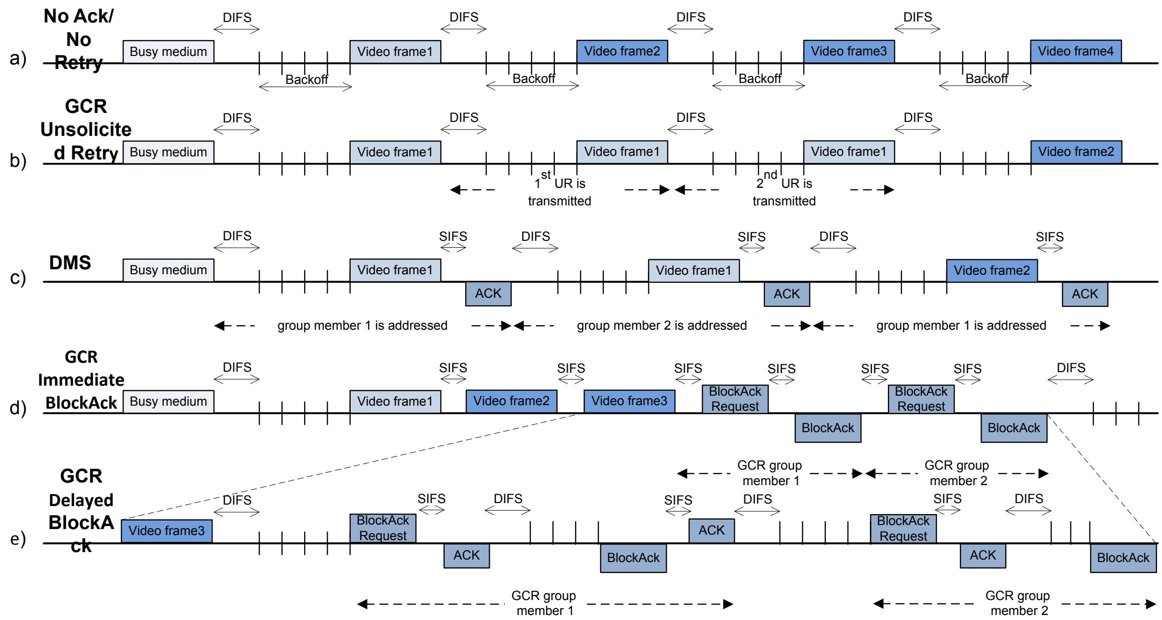


Figure 6.2: Example of video frames exchange for different group addressed frame MAC mechanisms

6.5 Extending IEEE 802.21 Signaling for Group Mobility

The usage of MIH signaling generates overhead due to the required information in the MIH frame. Message fields such as source and destination MIHF ID, service ID and others need to be present in every message. Also, the message exchange mechanism assumes a request/response method, further increasing the amount of data flowing in the network. This study, realized out of the scope of MEDIEVAL architecture, intends to improve the signaling efficiency in scenarios of group mobility.

Our study considers groups of users, connected to the same or nearby access networks, accessing broadcast or multicast video services. We argue that, when network conditions change due to the same phenomena (i.e., network congestion, servicing, or environmental causes) and affect a video feed received by several users nearby, it affects not just a single user, but blocks of users. In traditional MIH signaling, each single user would be the subject of an independent MIH signaling transaction. In this work, we aim not only to extend the core 802.21 mechanisms to support video-enhancing events and commands, but also to take advantage of the underlying multicast and broadcast framework, enabling the provision of 802.21 signaling via multicast.

The concept is shown in Figure 6.3. When the Network Decision Point (NDP) needs to send 802.21 messages affecting all nodes at the PoA, if it supports multicast 802.21, a single message is required. However, if there is no multicast 802.21 support, one message per terminal is required. To achieve the intended new feature over 802.21, four key interventions to the 802.21 mechanisms must be done:

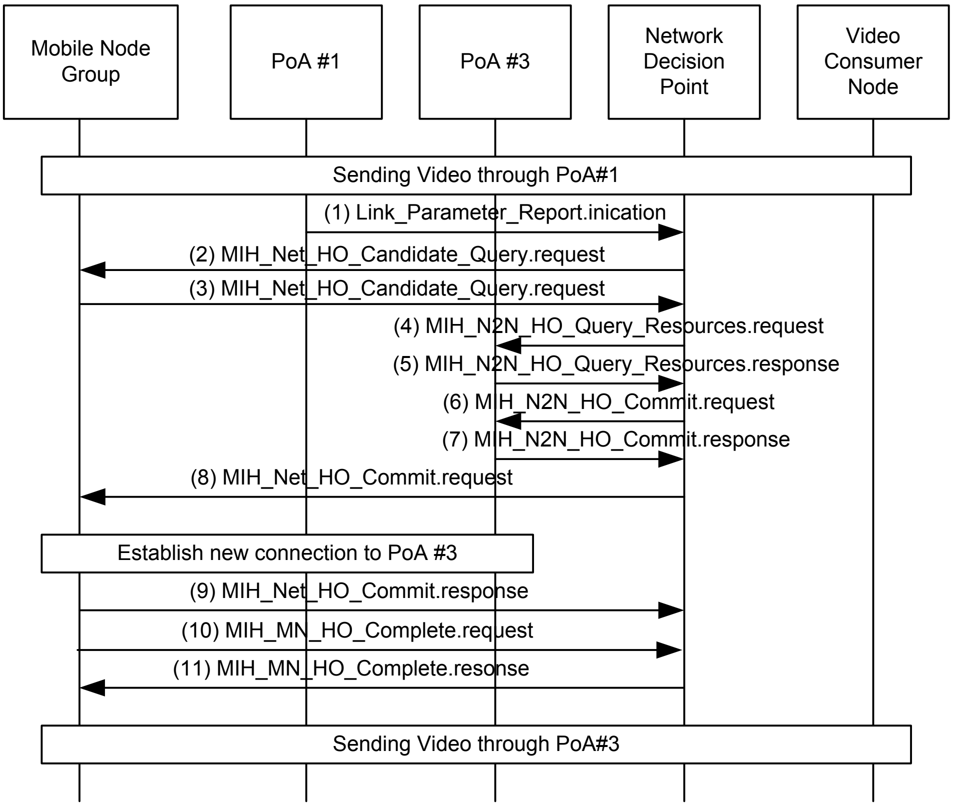


Figure 6.3: Multicast 802.21 HO scenario

6.5.1 Discovery and Capabilities Discovery

MIH nodes are able to discover each other and exchange information regarding supported services, using a MIH_Capability_Discover.request/response exchange, in a solicited or unsolicited way. In the first case, when the address of a node is already known, the message is issued with that address as a target. In the second, the node broadcasts the message and collects responses from nodes which have received it. The message contains the parameters presented in Table 6.1.

The SupportedTransportList parameter is a 16bit map, with two defined values (i.e., '0' for UDP and '1' for TCP) and the rest reserved. We added value '3' indicating "Multicast

Name	Description
SourceIdentifier	The invoker MIHF ID
LinkAddressList	An optional list of link addresses and types supported by the node
SupportedMIHEventList	Optional list of supported events
SupportedMIHCommandList	Optional list of supported commands
SupportedISQueryTypeList	Optional list of supported MIIS query types
SupportedTransportList	Optional list of supported transport types
MBBHandoverSupport	Optional list to indicate if a make before break HO is supported

Table 6.1: MIH_Capability_Discover.request parameters

support". We have also proposed a new optional parameter, "MulticastAddress" indicating the multicast address of that operator, over which multicast signaling is sent. This address can either be in IPv4 or IPv6, and is used by terminals to subscribe to the multicast group, and to indicate to 802.21-enabled network management entities their multicast support.

6.5.2 New Information Elements for MIIS

The MIIS provides standard IEs, which can be queried by terminal or network nodes, in order to obtain information about PoAs. IEs related to PoAs are presented in Table 6.2:

Information Element	Description
IE_POA_LINK_ADDR	Link address of this PoA
IE_POA_LOCATION	Geo-location of the PoA
IE_POA_CHANNEL_RANGE	Supported channel range
IE_POA_SYSTEM_INFO	System information supported by the PoA
IE_POA_SUBNET_INFO	Information about supported subnets
IE_POA_IP_ADDR	IP Address of PoA
Vendor Specific PoA IE	Vendor specific IEs

Table 6.2: PoA Information Elements

Two new items were added: IE_MULTICAST_SUPPORT, which indicates if this PoA supports multicast, and IE_MULTICAST_ADDRESS, which indicates the multicast address pertaining to the group of this PoA. These two new IEs assist in identifying PoAs with multicast support, which can have impact in HO candidate decision.

6.5.3 Multicast MIHF Identifier

The issued 802.21 remote commands and events must contain the source and destination MIHF identifiers. While using multicast 802.21 signaling, a new destination identifier has to be defined, which represents not one but all the nodes involved in the multicast group. In this case, the DESTINATION MIHF ID will be replaced with the IP multicast address, identifying the destination multicast group. Upon the creation of such message, this parameter will be evaluated by the MIHF and be sent as a multicast message towards the designated multicast group.

6.5.4 NET_SAP

Name	Description
TransportType	The transport protocol to be used.
SourceAddress	The source transport protocol address
DestinationAddress	The destination transport protocol address
ReliableDeliveryFlag	Usage of message reliability
MIHProtocolPDU	The MIH PDU to be sent

Table 6.3: MIH_TP_DATA.request parameters

For this matter the TransportType parameter was extended to support a 8 bit map, where the option 'Multicast' could be added to the other two (i.e., 'L2' and 'L3') - refer to Table 6.3. Upon the reception of this primitive with the 'Multicast' parameter, the transport services of the node interface with a multicast protocol to send the frame.

6.5.5 Integration with Multicast Group Management Protocol

In order to update the multicast tree, a core extension needs to be done to the MIHF. An MIH-User was created which was able to interface with a multicast group management protocol (i.e., IGMPv3 for IPv4 or MLDv2 for IPv6). Whenever a multicast 802.21-enabled node starts the discovery and capability procedures, and exchanges MIH_Capability_Discovery.request / response messages, the MIH-User interfacing with the group management protocol is fed with the multicast address provided by the capability message exchange (i.e., the new MulticastAddress parameter). With this multicast address, the MIH-User is able to initiate IGMP or MLD procedures, and thus the node is announced to the multicast router, which is now able to update the multicast tree.

6.5.6 eMBMS Enhancements for IP Mobility

The eMBMS, described in section 2.4.2, is an enhancement of the Evolved Packet System (EPS) which provides a point to multipoint capability for broadcast or multicast services, allowing resources to be shared in the network. In the eMBMS version of the EPS, the broadcast mode is provided by tightly synchronized cells organized in semi-static MBSFN areas. User mobility is ensured by the synchronization of the cells, with potential data loss being recovered by the applications. Its reference architecture was previously depicted in Figure 2.5. Multicast mode is not supported, which prevents from benefiting in a dynamic way of the resource sharing for sessions received by a reduced, but yet large enough, group. In this proposal, eMBMS is extended to provide several levels of QoS and improve its efficiency for the transfer of the video frames. Whenever possible, the opportunity of transferring the flow in a multicast bearer is exploited. The Session Start procedure is optimized to convey the maximum amount of information at once and reduce the number of steps needed for its completion, in particular by introducing some cross-layer parameters exchange to flatten the procedure at start-up or HO.

The eMBMS model has been integrated in the global MEDIEVAL architecture. The eNB is considered as the LTE PoA while the WLAN is seen as a trusted non-3GPP access. The session start and resource setup procedures at eNB are executed when receiving the request from the FM. The control plane functions for the communication between the eUTRAN and the MBMS-GW, collocated with the Mobility Management Entity (MME), are handled in the MAR. So far, the eMBMS does not really consider seamless mobility, which gives all freedom for a flexible solution. If the core network is multicast enabled, the multicast mobility procedures are executed. Within MEDIEVAL context, where this work was proposed, the MBMS-GW operates as a MR and is mapped to a combination of the FM and the MUME - both entities previously described. If the network is not multicast enabled, the multicast tree starts at the MBMS-GW, linked with the existing functions for unicast mobility located in the MAR. In both cases, the final hop is multicast on the wireless link, as defined in the settings of the flow description. The BM-SC functions are located inside the Core Network. Multicast/unicast decision based on network conditions should be part of the transport optimization sub-system. In this work, this decision is made based on the service type (multicast is default for Mobile TV and PBS services). User service provisioning and announcement are handled by the Video Services Control which takes care of the streaming functions.

6.6 Use Case Scenarios

In this section, two distinct scenarios - one with single user mobility, and other with group mobility - are described with the purpose of demonstrating the advantages of the previously introduced extensions.

6.6.1 Single Receiver Multicast Mobility Scenario

The chosen scenario (depicted in Figure 6.4) is intended to showcase the previously defined multicast mobility management mechanisms, for both source and receiver, and, to show in real-life scenario how such mechanisms can benefit from already existing technologies like SVC video.

The scenario is as follows: Mike has just started his own e-club channel for broadcasting video. Using this service, he shares with his subscribers the latest news and events in his town, thus sometimes needing to record on the move. Anne, one of his subscribers, has a mobile phone enabled with the introduced features (Stage 3). As such, she wants to take the most of it, watching videos with high level of quality even when moving. The MAR where she started viewing the live video has already some subscribers requesting the same video (stream S1 in the figure), but with inferior level of quality associated to their profiles, either because of their terminal characteristics or due their service level agreement (SLA) with the mobile operator. At a later time, Anne associates to a new MAR due to mobility; nevertheless the service is not interrupted.

We dwell into some of the crucial cross-layer communications that underpin the network behavior throughout the previous scenario, which was described in [10]. Relying on the previously presented mobility modules[132] is aware that entities acting as MIH users, both in the user terminal (CM) and access network (FM), are able to act towards the selection of the best available access network, not only taking into account radio properties but also information such as the availability of multicast routing capabilities. Besides, IEEE 802.21 is used for tasks such as notifying a MAR which router is supposed to act as the A-MAR, required for the tunnel establishment. It is also involved in informing the wireless access layers of resources to be established, enabling multicast session and SVC frames priorities.

In order to support a node acting as multicast source, at the service request and registration the uplink provisioning must be initiated. As such, Mike's application requests the terminal properties and network conditions. When Mike's terminal is associated to a new MAR (Stage 2), its session starts to be tunneled from the S-MAR to the A-MAR, which was configured as the multicast RP for the session. Besides, a vital interaction occurs between the terminal's CM and the Content Adaptation function of the network. Basically, this Content Adaption function receives input relative to network conditions, and in this particular case, is required for preparing the uplink before and after the occurrence of a HO operation. Another introduced feature is the interaction between transport-aware entities and mobility-related ones. This interface avoids mobility operations towards congested access points, due to a candidate network weighting process, and other intelligent decisions.

A possibility of the introduced mechanisms is to use a different multicast group for each expected quality (temporal, spatial, Signal to Noise ratio) set. This can be seen as a hybrid simulcast-layered solution, splitting the pros and cons of each transport mechanism. We foresee that in real networks the number of deployed layers per video will typically be low (e.g. four), and therefore the replication of information is bearable, and inferior to current simulcast proposals. On the other hand, most terminals will be spared from the SVC decoding effort, having an IP multicast session addressed to it. As such, when Anne starts receiving the video, MAR1 subscribes the missing layers, aggregates them in a single RTP session (represented as S2 in the figure) adapted to her terminal's needs.

In order to reduce the packet loss and delay during the mobility process, a multicast context transfer process takes place when Anne moves from MAR1 to MAR2 (Stage 4). As the new MAR is informed by the Decision Module (DM) that there is limited bandwidth available in its upstream link to the core, it doesn't subscribe to all the layers corresponding to the expected quality by Anne, aggregating a lower-quality version in a new RTP session to the same multicast destination address. Such mechanism is analogous to the SVC layer drop that may occur at wireless transmission, but takes place before the last hop. This means that, having two versions of the same video, V1 and V2, where V1 has more enhancement layers than V2, at some point in time V1 may actually be delivered with the same quality as V2, by network decision. Regarding the example, when the context transfer takes place between MAR1 and MAR2, the DM, which establishes a mapping between the multicast group being requested and the effective layers to be requested, is responsible for informing the FM placed the MAR2 about the subscriptions to be made for this multicast session. In practice, it leads to MAR2 not joining all the multicast groups (layers) of the video during the congestion period. This same kind of content adaptation may also be done as a consequence of a different trigger, such as QoE level decrease.

At that point there would be a single user requesting that stream in MAR2, so an adaptation of the scheme in [133], in the case of LTE, or DMS mode in the case of 802.11, may be used, increasing transmission reliability. As soon as MAR2 verifies it is able to support a better (and more bandwidth-demanding) version of the video, informed by DM, it subscribes to the missing channels and aggregates them transparently to Anne's mobile device.

6.6.2 Group of Receivers Mobility Scenario

To showcase the usefulness of using multicast 802.21 signaling, the following scenario is considered, applying the signaling from Figure 6.3. In this scenario, a group of users is attending a press conference and connected to a Wi-Fi hotspot. Consider a scenario where the users receive e.g. a high-quality PBS stream through broadcast, which quickly stresses available resources at that hotspot. Considering this, the NDP needs to move a block of users to another hotspot, for load balancing. Using the presented multicast 802.21 signaling, a single signaling action is required per block of users, instead of per specific user. The corresponding scenario is represented in Figure6.5, showing only remote 802.21 signaling.

When the PoA that is serving the Mobile Node Group (MNG) detects that network

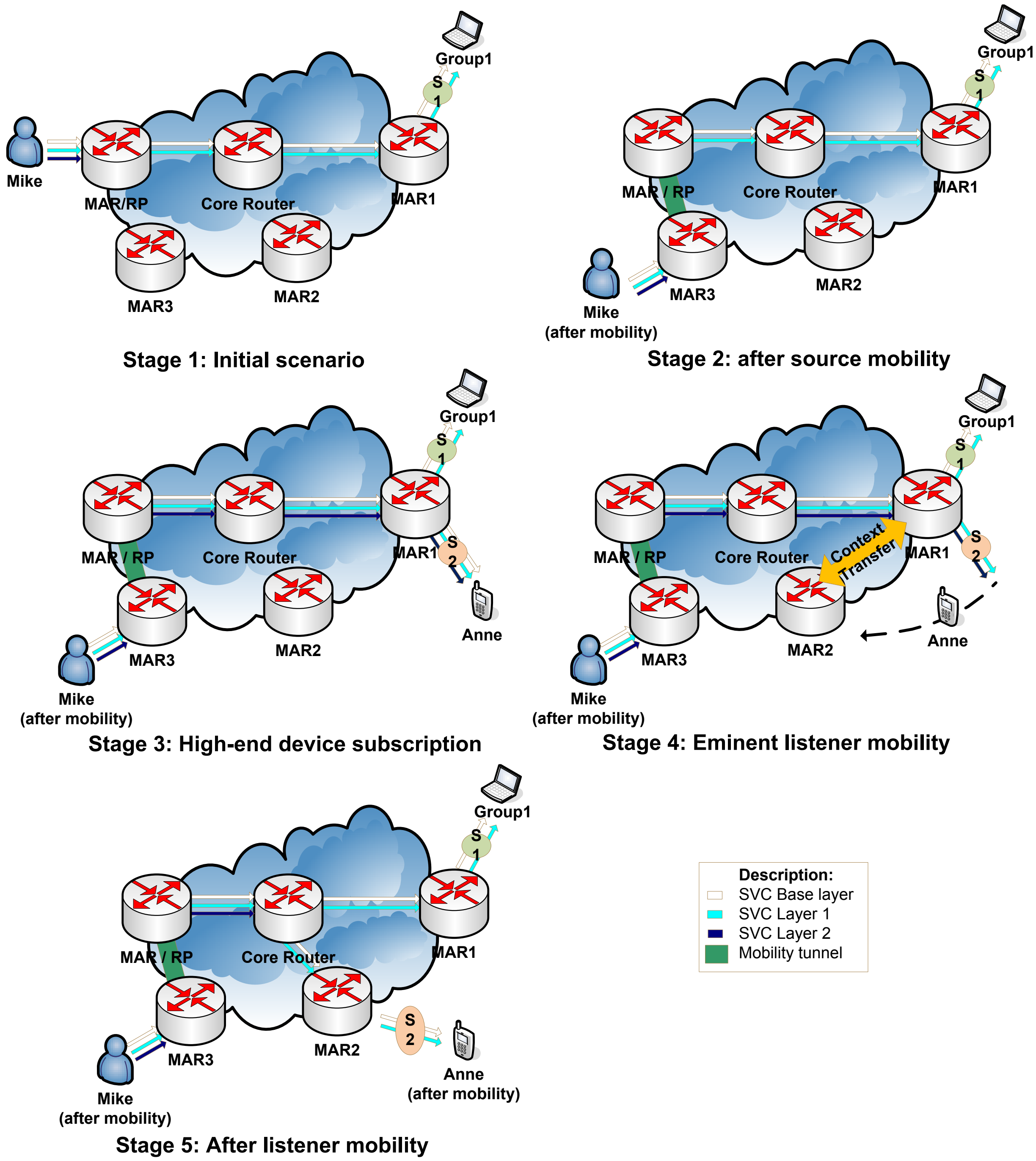


Figure 6.4: SVC multicast mobility scenario

conditions are decreasing, it generates a report event (1) towards the NDP, which then sends (2) towards the MNG in order to evaluate which other PoAs are within range. Notice that the message sent by the NDP is transported in multicast, but the answers are received independently, and thus it is able to evaluate for a common PoA within range of all nodes belonging to that block. The NDP selects PoA#3 as the HO candidate and sends (4) to query resources. Upon receiving the answer and verifying that PoA#3 is able to accommodate the user block, is commits those resources via (6), and commands the MNG to start HO procedures with (8). When this message is received, nodes are able to execute the L2 attachment at PoA#3 and report its result via (9). At this point, the MNG can initiate L3 mobility procedures if required after which (10) is sent to the NDP, which can trigger other procedures such as clearing resources at the old PoA. Finishing the signaling, the terminals at the MNG are now able to send video through PoA#3.

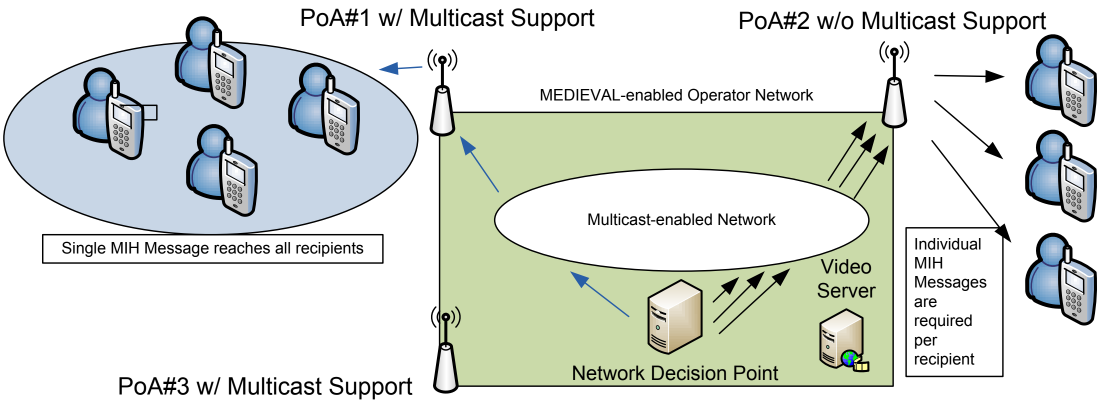


Figure 6.5: Multicast 802.21 HO scenario

6.7 Concluding Remarks

Mechanisms for enabling efficient and reliable IP multicast video delivery, focusing on optimizations suitable to mobile environments, were presented. Targeting three base requirements - scalability, improved rate transmission, and efficient signaling -, the proposals are described at a conceptual level and shown in two different scenarios, highlighting potential use cases. The advantages associated to the presented mechanisms depend on the provision of the right information to the suitable network decision entity / middleware - a logical next step in the validation of the introduced concepts would be the specification of this operation within the context of a 3GPP architecture, inserted for instance in the User Plane Congestion Management (UPCON[2]) work item.

[2]UPCON is targetted for 3GPP's Release 13

Chapter 7

Conclusions and Future Directions

7.1 Review of Achievements

The ubiquitous availability of Internet results in the shift of consumption habits to mobile, with users demanding immediate connectivity everywhere they go for fulfilling communication and social needs: receiving people's calls and sharing their own mood and status, photos, videos, etc. In particular, the reduction in the size and cost of mobile hardware is expected to enable "killer applications" such as Personal Broadcasting, where e.g. GoPro-sized cameras transmit recorded video using cellular or other wireless networks.

The convergence of billions of video-capable devices in a single network is thus a huge challenge to the underlying infrastructure. All generated traffic must be handled and transported using mobile networks, motivating mobile operators to react not only by improving access data rates but also by embedding non-3GPP accesses for offloading purposes, as well as through the updated architecture and protocols redesign, for the sake of robustness of their networks and satisfaction of their users.

Under these constrains, operators aim to benefit from novel interesting services leveraging on video, hoping to answer and even exceed user's expectations. In order to magnify introduced video services' attractiveness, operators need to take advantage of their strategic position and resources. They are best positioned to assure QoS and QoE throughout the whole sessions, and to deploy cost-efficient mechanisms to support these services, ultimately differentiating themselves from competition, namely Over-the-top (OTT) services. IP multicast and broadcasting mechanisms, for instance, are a naturally efficient offloading tool for several over-ther-air live transmission services. We have seen that solutions previous to this work still lack reliable mobility management approaches for preserving QoS and QoE throughout the whole sessions duration, and against distinct events, calling for intelligently designed cross-layer approaches which don't represent significant additional overhead.

The challenge presented to the author at the beginning of this work was to develop generic mobility solutions adapted to novel communication models and scenarios which result from the dominance of video, with a strong focus on efficiency and taking advantage of cross-layer interactions. From this point, the research work evolved towards optimizing mobile architectures and its modules for reliable and efficient video-aware transport in group-based video scenarios. One of the initially considered scenarios was that of group mobility, which enabled to identify an efficiency gap in IEEE 802.21's signaling in such scenarios. The results of this work were presented to IEEE 802.21 group, and partially triggered the currently being defined standard, IEEE 802.21d.

As such, the presented work started by designing a network architecture which reuses a widely accepted - though not widely deployed - mobility management solutions (PMIPv6) for fully supporting both mobile broadcasters or regular consumers. One of the first observations was that the network-localized mobility management nature of PMIPv6 and the unicast nature of the network stack fabric of operating systems require adaptations to enable vertical mobility in multicast environments. This motivated the consideration and adaptation of IEEE 802.21 as a multicast mobility enabler: through the strict coordination between hosts and network, enabling close-to-seamless transmission during intra- or vertical-HOs at the cost of minimal additional signaling overhead. Moreover, given the identification of a diversity of multicast receiver mobility solutions in PMIPv6, having in common the adoption of a multicast context transfer-based mechanism, a detailed comparison of the solution space was realized. With such work, it was possible to single-handedly tackle both receiver mobility and source mobility, which posed very distinct challenges with regards to the multicast session preservation and quality.

The predominance of video affects all network topology sections, from the network access to the core, in particular when considering current hierarchical schemes relying on the concept of fixed convergence points. While a significant effort is placed at the periphery of the network, further stressed by the constraints of wireless properties, there must be coordinated research in all fronts, and simultaneously take into account the evolution in multiple technologies such as mobile CDN systems. This motivated the research of optimizations for mobility management solutions, whose centralized nature converges most traffic at anchor points (PGW, LMA) and leads to backhaul congestion, which is in scope with the distributed and dynamic mobility paradigms. The yet-to-be-fully-defined Distributed Mobility Management protocol is expected to provide changes in the mobility management plane, the mobility functions placement, and the exposition of address properties to upper layers, i.e. applications. The involvement of applications in mobility decisions is expected to increase even while still applying network-based mobility management, as such involvement is needed for leading to higher customization of IP mobility services which ultimately lead to more efficient network operation [11].

The early consideration of IP multicast in currently investigated mobility management schemes is a necessary step to improve its successful integration probability, by extending its advantages in regards to network's efficiency maximization. Part of the developed work was dedicated to the identification of: 1) use cases where IP multicast support applies, such as receiver and source mobility while applying different technical approaches, (application of MLD Proxy or multicast router, etc); 2) technical limitations in such scenarios; 3) requirements of multicast mobility solutions in order to avoid or minimize identified limitations. This effort was reflected in IETF activities, with the contribution to the requirements of upcoming distributed mobility management solutions [6].

While the need for optimizing mobility architectures through dynamic and distributed functions is clear, it is natural that different Operators, and even different networks from a same Operator will present different characteristics - and needs -, such as the average number of users, type of traffic, access technologies, and so on. Thus, part of the developed work concerned the identification and classification of different architectural variables - 1) the multicast function deployed at the mobility management entity; 2) the subscription discovery method; 3) the upstream router selection method -; as well as their detailed evaluation. Such task provides Operators with a range of solutions and facilitates the evolution decision process of their existing networks. The evaluation of these options resulted in several observations, such as the fact that during HO the multicast re-subscription using the multicast infrastructure usually originates an higher service disruption than using an anchored subscription, and that leveraging on mobility tunneling for multicast support during HO does not introduce significant overhead.

The realized work was evaluated in prototypes developed by the author under the scope of MEDIEVAL project, with MUME module being a key concept for future MARs. MUME is presented here as a concept of what the author believes should be the functions delivered for flexible support of IP multicast mobility in a network-based mobility management protocol, both capable of tunnel-based or native multicast subscription and transmission. With variable network characteristics and needs, operators need to be able to dynamically adapt content transport to current status (e.g. according to content availability, congestion). MUME module was developed with this scenario in mind, and provided a core piece of the architecture.

A substantial part of the realized work impacted large network concepts, namely those related to multicast mobility support in emerging mobility management environments (and disseminated through several published articles referenced throughout the text).

7.2 Future Directions

As we've seen, the availability of novelty-enabler technologies leads users to apply it in refreshing - and more demanding for the network - ways. If we consider the "Internet of Things" trend, it is not difficult to imagine future scenarios such as video-monitoring / video analytics sensor networks, where sensors are enriched with mobility capabilities, such as so-called drones.

This huge shift from consumer to prosumer needs to be specifically tackled, with operators needing to find profitable business models to counter the sudden rise in uplink traffic. Technical adaptations to eMBMS in order to take into account PBS / UGC scenarios are one possibility.

Operators pursuit for efficient delivery of traffic is one far from completion. and those issues already observed (i.e. explosion in mobile multimedia traffic) will potentially be magnified. Several of existing paradigms will as such need a reevaluation. The centralized processing - distributed functionality dichotomy, or cloud vs flat architectures, will be one decision to pay attention at. While keeping the complexity at the core of the network reduces costs, there are also advantages in placing the intelligence - as well as the content - at the network's edges. The right balance between the two options will be a key discussion during the next years, which will certainly determine the design of video delivery systems, IP mobility management, and IP multicast (e.g. CCN) and their integration.

References

[1] S. Figueiredo, C. Guimaraes, D. Corujo, R. L. Aguiar, MI3M: A framework for media independent multicast mobility management, Computer Networks 70 (0) (2014) 345 – 365. doi:http://dx.doi.org/10.1016/j.comnet.2014.06.006.
URL http://www.sciencedirect.com/science/article/pii/S1389128614002370

[2] S. Jeon, S. Figueiredo, R. Aguiar, On the impacts of distributed and dynamic mobility management strategy: A simulation study, in: Wireless Days (WD), 2013 IFIP, 2013, pp. 1–6. doi:10.1109/WD.2013.6686509.

[3] S. Figueiredo, S. Jeon, R. L. Aguiar, IP Multicast Use Cases and Analysis over Distributed Mobility Management, Internet-Draft draft-sfigueiredo-multimob-use-case-dmm-03.txt, IETF Secretariat (2012).

[4] S. Figueiredo, S. Jeon, R. L. Aguiar, Use-cases analysis for multicast listener support in network-based distributed mobility management, in: Personal Indoor and Mobile Radio Communications (PIMRC), 2012 IEEE 23rd International Symposium on, 2012, pp. 1478–1483. doi:10.1109/PIMRC.2012.6362581.

[5] S. Jeon, S. Figueiredo, R. Aguiar, A channel-manageable ip multicast support framework for distributed mobility management, in: Wireless Days (WD), 2012 IFIP, 2012, pp. 1–6. doi:10.1109/WD.2012.6402867.

[6] H. Chan, D. Liu, P. Seite, H. Yokota, J. Korhonen, Requirements for Distributed Mobility Management, RFC 7333 (Informational) (Aug. 2014).
URL http://www.ietf.org/rfc/rfc7333.txt

[7] S. Figueiredo, C. Guimaraes, R. Aguiar, T.-T. Nguyen, L. Yadin, N. Carapeto, C. Parada, Broadcasting user content over novel mobile networks, in: Communications (ICC), 2013 IEEE International Conference on, 2013, pp. 3613–3618. doi:10.1109/ICC.2013.6655113.

[8] S. Figueiredo, S. Jeon, D. Gomes, R. L. Aguiar, D3M: multicast listener mobility support mechanisms over distributed mobility anchoring architectures, Journal of Network and Computer Applicationsdoi:http://dx.doi.org/10.1016/j.jnca.2015.02.006.

[9] S. Figueiredo, S. Jeon, R. L. Aguiar, Empowering ip multicast for multimedia delivery over heterogeneous mobile wireless networks, in: Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on, 2014, pp. 211–212. doi:10.1109/INFCOMW.2014.6849230.

[10] S. Figueiredo, M. Wetterwald, T.-T. Nguyen, L. Eznarriaga, N. Amram, R. L. Aguiar, SVC multicast video mobility support in MEDIEVAL Project, in: Future Network Mobile Summit (FutureNetw), 2012, 2012, pp. 1–8.

[11] S. Jeon, S. Figueiredo, Y. Kim, Use Cases and API Extension for Source IP address selection, Internet-Draft draft-sijeon-dmm-use-cases-api-source-00.txt, IETF Secretariat.

[12] D. Corujo, S. Figueiredo, R. L. Aguiar, Media-independent multicast signalling for enhanced video performance in the MEDIEVAL Project, in: Future Network Mobile Summit (FutureNetw), 2011, 2011, pp. 1–9.

[13] J. P. Barraca, D. Gomes, R. L. Aguiar, Amazing - advanced mobile wireless playground, in: T. Magedanz, A. Gavras, N. Thanh, J. Chase (Eds.), Testbeds and Research Infrastructures. Development of Networks and Communities, Vol. 46 of Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Springer Berlin Heidelberg, 2011, pp. 219–230. doi:10.1007/978-3-642-17851-1_18.
URL http://dx.doi.org/10.1007/978-3-642-17851-1_18

[14] I. Al-Surmi, M. Othman, B. Mohd Ali, Review: Mobility management for ip-based next generation mobile networks: Review, challenge and perspective, J. Netw. Comput. Appl. 35 (1) (2012) 295–315. doi:10.1016/j.jnca.2011.09.001.
URL http://dx.doi.org/10.1016/j.jnca.2011.09.001

[15] C. Perkins, IP Mobility Support for IPv4, RFC 3344 (Proposed Standard), obsoleted by RFC 5944, updated by RFCs 4636, 4721 (Aug. 2002).
URL http://www.ietf.org/rfc/rfc3344.txt

[16] D. Johnson, C. Perkins, J. Arkko, Mobility Support in IPv6, RFC 3775 (Proposed Standard), obsoleted by RFC 6275 (Jun. 2004).
URL http://www.ietf.org/rfc/rfc3775.txt

[17] R. Koodli, Mobile IPv6 Fast Handovers, RFC 5568 (Proposed Standard) (Jul. 2009). URL `http://www.ietf.org/rfc/rfc5568.txt`

[18] H. Soliman, C. Castelluccia, K. ElMalki, L. Bellier, Hierarchical Mobile IPv6 (HMIPv6) Mobility Management, RFC 5380 (Proposed Standard) (Oct. 2008). URL `http://www.ietf.org/rfc/rfc5380.txt`

[19] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, B. Patil, Proxy Mobile IPv6, RFC 5213 (Proposed Standard), updated by RFC 6543 (Aug. 2008). URL `http://www.ietf.org/rfc/rfc5213.txt`

[20] 3GPP, Architecture enhancements for non-3GPP accesses, TS 23.402, 3rd Generation Partnership Project (3GPP) (Sep. 2008). URL `http://www.3gpp.org/ftp/Specs/html-info/23402.htm`

[21] W. F. N. W. Group, et al., Wimax forum network architecture - stage 3: Detailed protocols and procedures - release 1, version 1.2, in: WiMAX Forum, January, 2008.

[22] C. Perkins, D. Johnson, J. Arkko, Mobility Support in IPv6, RFC 6275 (Proposed Standard) (Jul. 2011). URL `http://www.ietf.org/rfc/rfc6275.txt`

[23] H. Yokota, K. Chowdhury, R. Koodli, B. Patil, F. Xia, Fast Handovers for Proxy Mobile IPv6, RFC 5949 (Proposed Standard) (Sep. 2010). URL `http://www.ietf.org/rfc/rfc5949.txt`

[24] G. Bollano, D. Panno, F. Ricciato, M. Turolla, N. Vaccaro, D. Ettorre, Enhanced android connection manager: An application-based solution to manage mobile data traffic, in: World Telecommunications Congress (WTC), 2012, 2012, pp. 1–6.

[25] P. Seite, G. Feige, T. Melia, J. Zuniga, Connection manager requirements, Internet-Draft draft-seite-mif-connection-manager-02.txt, IETF Secretariat (2011).

[26] M. Blanchet, P. Seite, Multiple Interfaces and Provisioning Domains Problem Statement, RFC 6418 (Informational) (Nov. 2011). URL `http://www.ietf.org/rfc/rfc6418.txt`

[27] K. Taniuchi, Y. Ohba, V. Fajardo, S. Das, M. Tauil, Y.-H. Cheng, A. Dutta, D. Baker, M. Yajnik, D. Famolari, Ieee 802.21: Media independent handover: Features, applicability, and realization, Communications Magazine, IEEE 47 (1) (2009) 112–120. `doi:10.1109/MCOM.2009.4752687`.

[28] LAN/MAN Committee of the IEEE Computer Society, IEEE Std 802.21-2008, Standards for Local and Metropolitan Area - Part 21: Media Independent Handover Services, 2008.

[29] 3GPP, Generic Access Network (GAN); Stage 2, TS 43.318, 3rd Generation Partnership Project (3GPP) (Sep. 2008).
URL `http://www.3gpp.org/ftp/Specs/html-info/43318.htm`

[30] 3GPP, Requirements on 3GPP system to Wireless Local Area Network (WLAN) interworking, TS 22.234, 3rd Generation Partnership Project (3GPP) (Jun. 2007).
URL `http://www.3gpp.org/ftp/Specs/html-info/22234.htm`

[31] 3GPP, Direct tunnel deployment guideline, TR 23.919, 3rd Generation Partnership Project (3GPP) (Jun. 2007).
URL `http://www.3gpp.org/ftp/Specs/html-info/23919.htm`

[32] 3GPP, Service requirements for Home NodeBs and Home eNodeBs, TS 22.220, 3rd Generation Partnership Project (3GPP).
URL `http://www.3gpp.org/ftp/Specs/html-info/22220.htm`

[33] K. Samdanis, T. Taleb, S. Schmid, Traffic offload enhancements for eutran, Communications Surveys Tutorials, IEEE 14 (3) (2012) 884–896. `doi:10.1109/SURV.2011.072711.00168`.

[34] 3GPP, Study on S2a Mobility based on GPRS Tunnelling Protocol (GTP) and Wireless Local Area Network (WLAN) access to the Enhanced Packet Core (EPC) network (SaMOG); Stage 2, TR 23.852, 3rd Generation Partnership Project (3GPP) (2013).
URL `http://www.3gpp.org/ftp/Specs/html-info/23852.htm`

[35] 3GPP, Local ip access (lipa) mobility and selected ip traffic offload (sipto) at the local network, TR 23.859, 3rd Generation Partnership Project (3GPP) (Apr. 2013).
URL `http://www.3gpp.org/ftp/Specs/html-info/23852.htm`

[36] 3GPP, Access Network Discovery and Selection Function (ANDSF) Management Object (MO), TS 24.312, 3rd Generation Partnership Project (3GPP) (Dec. 2010).
URL `http://www.3gpp.org/ftp/Specs/html-info/24312.htm`

[37] J. Zuniga, C. Bernardos, A. de la Oliva, T. Melia, R. Costa, A. Reznik, Distributed mobility management: A standards landscape, Communications Magazine, IEEE 51 (3) (2013) 80–87. `doi:10.1109/MCOM.2013.6476870`.

[38] H. Holbrook, B. Cain, B. Haberman, Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast, RFC 4604 (Proposed Standard) (Aug. 2006).
URL `http://www.ietf.org/rfc/rfc4604.txt`

[39] R. Vida, L. Costa, Multicast Listener Discovery Version 2 (MLDv2) for IPv6, RFC 3810 (Proposed Standard), updated by RFC 4604 (Jun. 2004).
URL `http://www.ietf.org/rfc/rfc3810.txt`

[40] B. Fenner, M. Handley, H. Holbrook, I. Kouvelas, Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised), RFC 4601 (Proposed Standard), updated by RFCs 5059, 5796, 6226 (Aug. 2006).
URL `http://www.ietf.org/rfc/rfc4601.txt`

[41] 3GPP, Typical examples of Radio Access Bearers (RABs) and Radio Bearers (RBs) supported by Universal Terrestrial Radio Access (UTRA), TR 25.993, 3rd Generation Partnership Project (3GPP) (Sep. 2008).
URL `http://www.3gpp.org/ftp/Specs/html-info/25993.htm`

[42] 3GPP, Feasibility study on improvement of the Multimedia Broadcast / Multicast Service (MBMS) in UTRAN, TR 25.905, 3rd Generation Partnership Project (3GPP) (Jan. 2008).
URL `http://www.3gpp.org/ftp/Specs/html-info/25905.htm`

[43] M. Gruber, D. Zeller, Multimedia broadcast multicast service: new transmission schemes and related challenges, Communications Magazine, IEEE 49 (12) (2011) 176–181. `doi:10.1109/MCOM.2011.6094023`.

[44] 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRAN); Overall description; Stage 2, TS 36.300v8.9.0, 3rd Generation Partnership Project (3GPP) (Jun. 2009).
URL `http://www.3gpp.org/ftp/Specs/html-info/36300.htm`

[45] 3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRAN); Overall description; Stage 2, TS 36.300v10.3.0, 3rd Generation Partnership Project (3GPP) (Sep. 2012).
URL `http://www.3gpp.org/ftp/Specs/html-info/36300.htm`

[46] D. Lecompte, F. Gabin, Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: overview and Rel-11 enhancements, Communications Magazine, IEEE 50 (11) (2012) 68 –74. `doi:10.1109/MCOM.2012.6353684`.

[47] R. Pantos, W. May, HTTP Live Streaming, Internet-Draft draft-pantos-http-live-streaming-13.txt, IETF Secretariat (2014).

[48] 3GPP, Transparent end-to-end Packet-switched Streaming Service (PSS); Protocols and codecs, TS 26.234, 3rd Generation Partnership Project (3GPP).
URL `http://www.3gpp.org/ftp/Specs/html-info/26234.htm`

[49] F. Gabin, M. Kampmann, T. Lohmar, C. Priddle, 3gpp mobile multimedia streaming standards [standards in a nutshell], Signal Processing Magazine, IEEE 27 (6) (2010) 134–138. doi:10.1109/MSP.2010.938087.

[50] ISO/IEC, ISO/IEC 23009-1:2012. Dynamic adaptive streaming over HTTP (DASH), ISO/IEC, 2012.

[51] 3GPP, Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH), TS 26.247, 3rd Generation Partnership Project (3GPP) (Jun. 2011).
URL `http://www.3gpp.org/ftp/Specs/html-info/26247.htm`

[52] 3GPP, Study on Personal Broadcast Service (PBS), TR 22.947, 3rd Generation Partnership Project (3GPP) (Mar. 2011).
URL `http://www.3gpp.org/ftp/Specs/html-info/22947.htm`

[53] I. Romdhani, M. Kellil, H.-Y. Lach, A. Bouabdallah, H. Bettahar, IP mobile multicast: Challenges and solutions, Communications Surveys Tutorials, IEEE 6 (1) (2004) 18 –41. doi:10.1109/COMST.2004.5342232.

[54] T. Schmidt, M. Waehlisch, G. Fairhurst, Multicast Mobility in Mobile IP Version 6 (MIPv6): Problem Statement and Brief Survey, RFC 5757 (Informational) (Feb. 2010).
URL `http://www.ietf.org/rfc/rfc5757.txt`

[55] S. Deering, W. Fenner, B. Haberman, Multicast Listener Discovery (MLD) for IPv6, RFC 2710 (Proposed Standard), updated by RFCs 3590, 3810 (Oct. 1999).
URL `http://www.ietf.org/rfc/rfc2710.txt`

[56] H. Asaeda, H. Liu, Q. Wu, Tuning the Behavior of the Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) for Routers in Mobile and Wireless Networks, RFC 6636 (Informational) (May 2012).
URL `http://www.ietf.org/rfc/rfc6636.txt`

[57] T. Schmidt, M. Waehlisch, S. Krishnan, Base Deployment for Multicast Listener Support in Proxy Mobile IPv6 (PMIPv6) Domains, RFC 6224 (Informational) (Apr. 2011).
URL `http://www.ietf.org/rfc/rfc6224.txt`

[58] J. Zuniga, L. Contreras, C. Bernardos, S. Jeon, Y. Kim, Multicast Mobility Routing Optimizations for Proxy Mobile IPv6, RFC 7028 (Experimental) (Sep. 2013). URL http://www.ietf.org/rfc/rfc7028.txt

[59] L. M. Contreras, C. J. Bernardos, I. Soto, PMIPv6 multicast handover optimization by the Subscription Information Acquisition through the LMA (SIAL), MULTIMOB WG document (Dec. 2013).

[60] T. C. Schmidt, M. Waehlisch, R. Koodli, G. Fairhurst, Multicast Listener Extensions for MIPv6 and PMIPv6 Fast Handovers, Internet-Draft draft-ietf-multimob-fmipv6-pfmipv6-multicast-09, IETF Secretariat (2014).

[61] T.-T. Nguyen, C. Bonnet, Performance optimization of multicast content delivery in a mobile environment based on pmipv6, in: Wireless Communications and Networking Conference (WCNC), 2013 IEEE, 2013, pp. 1249–1254. doi:10.1109/WCNC.2013.6554743.

[62] O. Christ, T. Schmidt, M. Wahlisch, Towards Seamless Handovers in SSM Source Mobility An Evaluation of the Tree Morphing Protocol, in: Next Generation Mobile Applications, Services and Technologies, 2008. NGMAST '08. The Second International Conference on, 2008, pp. 456 –461. doi:10.1109/NGMAST.2008.82.

[63] T. C. Schmidt, S. Gao, H. Zhang, M. Waehlisch, Mobile Multicast Sender Support in Proxy Mobile IPv6 (PMIPv6) Domains, Internet-Draft draft-ietf-multimob-pmipv6-source-09, IETF Secretariat (2014).

[64] L. Wang, S. Gao, H. Zhang, T. C. Schmidt, J. Guan, Multicast Source Mobility Support Schemes in PMIPv6 Networks, in: Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th, 2013, pp. 1–7. doi:10.1109/VTCFall.2013.6692291.

[65] T. Melia, S. Gundavelli, Logical Interface Support for multi-mode (IP) Hosts, NETEXT WG document (Mar. 2014).

[66] L. A. Mphatsi, O. E. Falowo, Achieving efficient network resources utilization in bicasting Proxy Mobile IPv6, in: Wireless and Mobile Computing, Networking and Communications (WiMob), 2012 IEEE 8th International Conference on, 2012, pp. 55 –60. doi:10.1109/WiMOB.2012.6379131.

[67] I.-S. Jang, W.-T. Kim, Y.-J. Park, An efficient mobile multicast mechanism based on media independent handover, in: Communications (MICC), 2009 IEEE 9th Malaysia International Conference on, 2009, pp. 501 –505. doi:10.1109/MICC.2009.5431610.

[68] H.-F. Lo, S. Guizani, T.-Y. Wu, W.-T. Lee, H.-W. Lin, IEEE 802.21-based seamless multicast streaming with dynamic playback control, Computer Communications 35 (2) (2012) 179 – 187. doi:10.1016/j.comcom.2011.09.002.
URL http://www.sciencedirect.com/science/article/pii/S0140366411002726

[69] K. Ma, R. Bartos, S. Bhatia, R. Nair, Mobile video delivery with http, Communications Magazine, IEEE 49 (4) (2011) 166–175. doi:10.1109/MCOM.2011.5741161.

[70] Y. Liang, J. Apostolopoulos, B. Girod, Analysis of packet loss for compressed video: Effect of burst losses and correlation between error frames, Circuits and Systems for Video Technology, IEEE Transactions on 18 (7) (2008) 861–874. doi:10.1109/TCSVT.2008.923139.

[71] Y. Liang, B. Girod, Network-adaptive low-latency video communication over best-effort networks, Circuits and Systems for Video Technology, IEEE Transactions on 16 (1) (2006) 72–81. doi:10.1109/TCSVT.2005.856919.

[72] Y. Li, A. Markopoulou, J. Apostolopoulos, N. Bambos, Content-aware playout and packet scheduling for video streaming over wireless links, Multimedia, IEEE Transactions on 10 (5) (2008) 885–895. doi:10.1109/TMM.2008.922860.

[73] Y. Lim, K. Park, J. Y. Lee, S. Aoki, G. Fernando, Mmt: An emerging mpeg standard for multimedia delivery over the internet, MultiMedia, IEEE 20 (1) (2013) 80–85. doi:10.1109/MMUL.2013.7.

[74] J. Widmer, M. Handley, TCP-Friendly Multicast Congestion Control (TFMCC): Protocol Specification, RFC 4654 (Experimental) (Aug. 2006).
URL http://www.ietf.org/rfc/rfc4654.txt

[75] H. Yousefi'zadeh, H. Jafarkhani, Achieving inter-receiver fairness utilizing layered media multicast control (lmmc), in: Communications, 2004 IEEE International Conference on, Vol. 2, 2004, pp. 1151–1155 Vol.2. doi:10.1109/ICC.2004.1312680.

[76] C. Bouras, A. Gkamas, G. Kioumourtzis, Adaptive smooth simulcast protocol (assp) for video applications: Description and performance evaluation, J. Netw. Syst. Manage. 19 (2) (2011) 143–177. doi:10.1007/s10922-010-9159-8.
URL http://dx.doi.org/10.1007/s10922-010-9159-8

[77] Y.-M. Hsiao, J.-F. Lee, J.-S. Chen, Y.-S. Chu, Review: H.264 video transmissions over wireless networks: Challenges and solutions, Comput. Commun. 34 (14) (2011) 1661–1672. doi:10.1016/j.comcom.2011.03.016.
URL http://dx.doi.org/10.1016/j.comcom.2011.03.016

[78] P. McDonagh, C. Vallati, A. Pande, P. Mohapatra, P. Perry, E. Mingozzi, Investigation of scalable video delivery using h.264 svc on an lte network, in: Wireless Personal Multimedia Communications (WPMC), 2011 14th International Symposium on, 2011, pp. 1–5.

[79] J. Nightingale, Q. Wang, C. Grecos, Hevstream: a framework for streaming and evaluation of high efficiency video coding (hevc) content in loss-prone networks, Consumer Electronics, IEEE Transactions on 58 (2) (2012) 404–412. doi:10.1109/TCE.2012.6227440.

[80] J. Roberts, The clean-slate approach to future internet design: a survey of research initiatives, annals of telecommunications - annales des tÃ©lÃ©communications 64 (5-6) (2009) 271–276. doi:10.1007/s12243-009-0109-y.
URL http://dx.doi.org/10.1007/s12243-009-0109-y

[81] G. Paravati, F. Lamberti, A. Sanna, C. Celozzi, High-performance solutions for adaptive and customizable streaming of interactive content to mobile devices, in: A. S. Sarmiento, M. Lopez (Eds.), Multimedia Services and Streaming for Mobile Devices: Challenges and Innovations, IGI Global, 2012, pp. 154–180.

[82] F. Foukalas, V. Gazis, N. Alonistioti, Cross-layer design proposals for wireless mobile networks: A survey and taxonomy, Communications Surveys Tutorials, IEEE 10 (1) (2008) 70–85. doi:10.1109/COMST.2008.4483671.

[83] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, W. Kellerer, Application-driven cross-layer optimization for video streaming over wireless networks, Communications Magazine, IEEE 44 (1) (2006) 122–130. doi:10.1109/MCOM.2006.1580942.

[84] H. Jiang, W. Zhuang, X. Shen, Cross-layer design for resource allocation in 3g wireless networks and beyond, Communications Magazine, IEEE 43 (12) (2005) 120–126. doi:10.1109/MCOM.2005.1561929.

[85] Y. Falik, A. Averbuch, U. Yechiali, Transmission algorithm for video streaming over cellular networks, Wirel. Netw. 16 (5) (2010) 1459–1475. doi:10.1007/s11276-009-0214-x.
URL http://dx.doi.org/10.1007/s11276-009-0214-x

[86] M. van der Schaar, D. Turaga, R. Wong, Classification-based system for cross-layer optimized wireless video transmission, Multimedia, IEEE Transactions on 8 (5) (2006) 1082–1095. doi:10.1109/TMM.2006.879827.

[87] S. Cicalo, V. Tralli, Distortion-fair cross-layer resource allocation for scalable video transmission in ofdma wireless networks, Multimedia, IEEE Transactions on 16 (3) (2014) 848–863. doi:10.1109/TMM.2014.2300442.

[88] H. Nishiyama, H. Yamada, H. Yoshino, N. Kato, A cooperative user-system approach for optimizing performance in content distribution/delivery networks, Selected Areas in Communications, IEEE Journal on 30 (2) (2012) 476–483. doi:10.1109/JSAC.2012.120228.

[89] X. Liu, F. Dobrian, H. Milner, J. Jiang, V. Sekar, I. Stoica, H. Zhang, A case for a coordinated internet video control plane, in: Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, SIGCOMM '12, ACM, New York, NY, USA, 2012, pp. 359–370. doi:10.1145/2342356.2342431.
URL http://doi.acm.org/10.1145/2342356.2342431

[90] C. Atici, M. Sunay, Improving the performance of wireless h.264 video broadcasting through a cross-layer design, in: Broadband Multimedia Systems and Broadcasting, 2009. BMSB '09. IEEE International Symposium on, 2009, pp. 1–6. doi:10.1109/ISBMSB.2009.5133837.

[91] C. Singhal, S. De, R. Trestian, G.-M. Muntean, Joint optimization of user-experience and energy-efficiency in wireless multimedia broadcast, Mobile Computing, IEEE Transactions on 13 (7) (2014) 1522–1535. doi:10.1109/TMC.2013.138.

[92] B. Fenner, H. He, B. Haberman, H. Sandick, Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying"), RFC 4605 (Proposed Standard) (Aug. 2006).
URL http://www.ietf.org/rfc/rfc4605.txt

[93] J. Loughney, M. Nakhjiri, C. Perkins, R. Koodli, Context Transfer Protocol (CXTP), RFC 4067 (Experimental) (Jul. 2005).
URL http://www.ietf.org/rfc/rfc4067.txt

[94] T. Schmidt, S. Gao, H. Zhang, M. Waehlisch, Mobile Multicast Sender Support in Proxy Mobile IPv6 (PMIPv6) Domains, RFC 7287 (Experimental) (Jun. 2014).
URL http://www.ietf.org/rfc/rfc7287.txt

[95] R. Koodli, Fast Handovers for Mobile IPv6, RFC 4068 (Experimental), obsoleted by RFC 5268 (Jul. 2005).
URL http://www.ietf.org/rfc/rfc4068.txt

[96] T. Melia, D. Corujo, A. de la Oliva, A. Vidal, R. Aguiar, I. Soto, Impact of heterogeneous network controlled handovers on multi-mode mobile device design, in: Wireless Communications and Networking Conference, 2007.WCNC 2007. IEEE, 2007, pp. 3884 –3889. doi:10.1109/WCNC.2007.710.

[97] D. Corujo, C. Guimaraes, B. Santos, R. Aguiar, Using an open-source IEEE 802.21 implementation for network-based localized mobility management, Communications Magazine, IEEE 49 (9) (2011) 114 –123. doi:10.1109/MCOM.2011.6011742.

[98] J. Baek, S. Seo, F. Shi, J. Song, A novel pre-authentication scheme based on fast channel switching in IEEE 802.11 WLANs, EURASIP Journal on Wireless Communications and Networking 2012 (1) (2012) 1–15. doi:10.1186/1687-1499-2012-39. URL http://dx.doi.org/10.1186/1687-1499-2012-39

[99] B. Santos, Localized mobility management protocol implementation using PMIPv6, Master Thesis, Universidade de Aveiro (2011).

[100] K. Samdanis, T. Taleb, S. Schmid, Traffic offload enhancements for eutran, Communications Surveys Tutorials, IEEE 14 (3) (2012) 884–896. doi:10.1109/SURV.2011.072711.00168.

[101] J.-H. Lee, J.-M. Bonnin, P. Seite, H. Chan, Distributed ip mobility management from the perspective of the ietf: motivations, requirements, approaches, comparison, and challenges, Wireless Communications, IEEE 20 (5) (2013) 159–168. doi:10.1109/MWC.2013.6664487.

[102] D.-H. Shin, D. Moses, M. Venkatachalam, S. Bagchi, Distributed mobility management for efficient video delivery over all-ip mobile networks: Competing approaches, Network, IEEE 27 (2) (2013) 28–33. doi:10.1109/MNET.2013.6485093.

[103] K. Pentikousis, P. Bertin, Mobility management in infrastructure networks, Internet Computing, IEEE 17 (5) (2013) 74–79. doi:10.1109/MIC.2013.98.

[104] J.-H. Lee, J. Bonnin, X. Lagrange, Host-based distributed mobility management support protocol for ipv6 mobile networks, in: Wireless and Mobile Computing, Networking and Communications (WiMob), 2012 IEEE 8th International Conference on, 2012, pp. 61–68. doi:10.1109/WiMOB.2012.6379140.

[105] H. Chan, Proxy mobile ip with distributed mobility anchors, in: GLOBECOM Workshops (GC Wkshps), 2010 IEEE, 2010, pp. 16–20. doi:10.1109/GLOCOMW.2010.5700231.

[106] P. Bertin, S. Bonjour, J. Bonnin, Distributed or centralized mobility?, in: Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE, 2009, pp. 1–6. doi:10.1109/GLOCOM.2009.5426302.

[107] P. Bertin, S. Bonjour, J. Bonnin, An evaluation of dynamic mobility anchoring, in: Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th, 2009, pp. 1–5. doi:10.1109/VETECF.2009.5378906.

[108] H. Ali-Ahmad, M. Ouzzif, P. Bertin, X. Lagrange, Comparative performance analysis on dynamic mobility anchoring and proxy mobile ipv6, in: Wireless Personal Multimedia Communications (WPMC), 2012 15th International Symposium on, 2012, pp. 653–657.

[109] H. Ali-Ahmad, M. Ouzzif, P. Bertin, X. Lagrange, Distributed mobility management: Approaches and analysis, in: Communications Workshops (ICC), 2013 IEEE International Conference on, 2013, pp. 1297–1302. doi:10.1109/ICCW.2013.6649437.

[110] H. Ali-Ahmad, D. Moses, H. Moustafa, P. Seite, T. Condeixa, Mobility Anchor Selection in DMM: Use-case Scenarios, Internet-Draft draft-yegin-dmm-cnet-homing-02.txt, IETF Secretariat (2013).

[111] C. Bernardos, A. de la Oliva, F. Giust, A PMIPv6-based solution for Distributed Mobility Management, Internet-Draft draft-bernardos-dmm-pmip-03.txt, IETF Secretariat (May 2014).

[112] P. Ernest, H. Chan, O. Falowo, Distributed mobility management scheme with mobility routing function at the gateways, in: Global Communications Conference (GLOBECOM), 2012 IEEE, 2012, pp. 5254–5259. doi:10.1109/GLOCOM.2012.6503955.

[113] A. Yegin, K. Kweon, J. Lee, J. Park, Corresponding Network Homing, Internet-Draft draft-yegin-dmm-cnet-homing-02.txt, IETF Secretariat (2014).

[114] D. Thaler, R. Draves, A. Matsumoto, T. Chown, Default Address Selection for Internet Protocol Version 6 (IPv6), RFC 6724 (Proposed Standard) (Sep. 2012).
URL http://www.ietf.org/rfc/rfc6724.txt

[115] A. Yegin, K. Kweon, J. Lee, J. Park, On Demand Mobility Management, Internet-Draft draft-yegin-dmm-ondemand-mobility-02.txt, IETF Secretariat (2014).

[116] D. Liu, H. Deng, C. Perkins, Address selection for DMM, Internet-Draft draft-liu-dmm-address-selection-01, IETF Secretariat (2012).

[117] J. Korhonen, B. Patil, S. Gundaveli, P. Seite, D. Liu, IPv6 Prefix Properties, Internet-Draft draft-korhonen-6man-prfix-properties, IETF Secretariat (2013).

[118] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, X. Amatriain, Watching television over an ip network, in: Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement, IMC '08, ACM, New York, NY, USA, 2008, pp. 71–84. doi:10.1145/1452520.1452529.
URL http://doi.acm.org/10.1145/1452520.1452529

[119] D. von Hugo, H. Asaeda, P. Seite, Context Transfer for Multicast support in Distributed Mobility Management (DMM), Internet-Draft draft-vonhugo-multimob-dmm-context-00.txt, IETF Secretariat (2012).

[120] F. Giust, A. De La Oliva, C. J. Bernardos, R. Da Costa, A network-based localized mobility solution for distributed mobility management, in: Wireless Personal Multimedia Communications (WPMC), 2011 14th International Symposium on, 2011, pp. 1–5.

[121] H. Asaeda, IGMP/MLD-Based Explicit Membership Tracking Function for Multicast Routers, Internet-Draft draft-ietf-pim-explicit-tracking-05, IETF Secretariat (Feb. 2013).

[122] T. Schmidt, M. WÃ€hlisch, Analysis of handover frequencies for predictive, reactive and proxy schemes and their implications on ipv6 and multicast mobility, in: P. Lorenz, P. Dini (Eds.), Networking - ICN 2005, Vol. 3421 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2005, pp. 1039–1046. doi:10.1007/978-3-540-31957-3_118.
URL http://dx.doi.org/10.1007/978-3-540-31957-3_118

[123] S. Jeon, N. Kang, Y. Kim, Mobility management based on proxy mobile ipv6 for multicasting services in home networks, Consumer Electronics, IEEE Transactions on 55 (3) (2009) 1227–1232. doi:10.1109/TCE.2009.5277980.

[124] B. Quinn, K. Almeroth, IP Multicast Applications: Challenges and Solutions, RFC 3170 (Informational) (Sep. 2001).
URL http://www.ietf.org/rfc/rfc3170.txt

[125] L. M. Contreras, C. J. Bernardos, I. Soto, Rams: A protocol extension to pmipv6 for improving handover performance of multicast traffic, Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 2 (2) (2011) 67–82.

[126] A. Diab, A. Mitschele-Thiel, K. Getov, O. Blume, Analysis of proxy mipv6 performance compared to fast mipv6, in: Local Computer Networks, 2008. LCN 2008. 33rd IEEE Conference on, 2008, pp. 579–580. doi:10.1109/LCN.2008.4664245.

[127] G. A. Leoleis, G. N. Prezerakos, I. S. Venieris, Seamless multicast mobility support using fast mipv6 extensions, Comput. Commun. 29 (18) (2006) 3745–3765. doi:10.1016/j.comcom.2006.07.013.
URL http://dx.doi.org/10.1016/j.comcom.2006.07.013

[128] D.-H. Kwon, W.-J. Kim, Y.-J. Suh, An efficient mobile multicast mechanism for fast handovers: A study from design and implementation in experimental networks., Computer Communications 31 (10) (2008) 2162–2177.
URL http://dblp.uni-trier.de/db/journals/comcom/comcom31.html#KwonKS08

[129] K. deok Seo, J. soo Kim, S. heung Jung, J. Yoo, A practical rtp packetization scheme for svc video transport over ip networks, ETRI Journal 32 (2010) .281–291.

[130] S. Wenger, Y.-K. Wang, T. Schierl, A. Eleftheriadis, RTP Payload Format for Scalable Video Coding, RFC 6190 (Proposed Standard) (May 2011).
URL http://www.ietf.org/rfc/rfc6190.txt

[131] ?, Ieee standard for information technology-telecommunications and information exchange between systems-local and metropolitan area networks-specific requirements - part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications, IEEE Std 802.11-2007 (Revision of IEEE Std 802.11-1999) (2007) C1–1184doi:10.1109/IEEESTD.2007.373646.

[132] D. Corujo, C. Bernardos, T. Melia, M. Wetterwald, L. Badia, R. Aguiar, Key function interfacing for the medieval project video-enhancing architecture, in: K. Pentikousis, R. Aguiar, S. Sargento, R. AgÃŒero (Eds.), Mobile Networks and Management, Vol. 97 of Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Springer Berlin Heidelberg, 2012, pp. 230–243. doi:10.1007/978-3-642-30422-4_17.
URL http://dx.doi.org/10.1007/978-3-642-30422-4_17

[133] S. Gundavelli, M. Townsley, O. Troan, W. Dec, Address Mapping of IPv6 Multicast Packets on Ethernet, RFC 6085 (Proposed Standard) (Jan. 2011).
URL http://www.ietf.org/rfc/rfc6085.txt