Mobility management in a converged fixed-mobile operator's network

Souheir Eido

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

UNIVERSITE BRETAGNE LOIRE

THÈSE / IMT Atlantique

sous le sceau de l'Université Bretagne Loire

pour obtenir le grade de

DOCTEUR D'IMT Atlantique

Mention : Informatique

École Doctorale Matisse

Présentée par

Souheir Eido

Préparée dans le département Informatique
Laboratoire Irisa

Contrôle de la mobilité dans un réseau d'opérateur convergé fixe-mobile

Thèse soutenue le 12 juillet 2017

devant le jury composé de :

Xavier Lagrange
Professeur, IMT Atlantique / président

Hossam Afifi
Professeur, Télécom SudParis / rapporteur

Farid Naït Abdesselam
Professeur, Université Parsi Descartes / rapporteur

Houda Labiod
Professeur, Télécom ParisTech / examinateur

Yassine Hadjadj Aoul
Assistant Professor, Université de Rennes 1 / examinateur

Annie Gravey
Directeur d'études, IMT Atlantique / directeur de thèse

Philippe Bertin
Senior Research Engineer, Orange Labs - Cesson Sévigné / invité

Stéphane Gosselin
Responsable de projets de recherche, Orange Labs - Lannion / invité

"It always seems impossible until it's done...!!"

Nelson MANDELA

Résumé

Les travaux de la présente thèse de doctorat ont été réalisés dans le contexte du projet européen FP7 COMBO (COnvergence of fixed and Mobile BrOadband access/aggregation networks, Grant Agreement number: 317762). Ce projet a démarré en janvier 2013, dans le cadre d'une collaboration entre 16 partenaires, notamment: Orange, Telefonica, Deutsche Telekom, Télécom Bretagne, Université de technologie et économique de Budapest, Politecnico di Torino, Ericsson, FON, ADVA.

Historiquement, les réseaux fixes et mobiles ont été optimisés et ont évolué de manière indépendante. Les organes de normalisation des réseaux fixes et mobiles sont différents et traitent donc les problématiques de déploiement des deux familles de réseaux séparément. COMBO propose une optimisation conjointe des réseaux d'accès et d'agrégation fixe et mobile autour du concept innovant des Points de Présence de Nouvelle Génération (NG-POP).

Dans un avenir proche, les réseaux fixes et mobiles devront supporter des volumes de traffic considérables. Par exemple, selon les prédictions de Cisco dans [1], le traffic IP global annuel devrait atteindre 2,3 ZettaBytes d'ici 2020, ce qui représente près de 100 fois le volume du traffic IP global mesuré en 2005. Cette croissance provient du nombre croissant de périphériques connectés (smartphones, tablettes, PC, ordinateurs portables, etc.) mais aussi du fait que les services IP utilisent des contenus vidéo de haute définition tels que les jeux en ligne, Internet Protocol TéléVision (IPTV), la vidéo à haute définition (streaming, conférence, etc.) [2] [3]. Une telle demande risque d'accroître le risque de congestion, et donc de dégrader la qualité de service (QoS) délivrée en augmentant la charge du traffic dans les différentes parties du réseau (accès, agrégation et cœur).

Pour faire face à ces demandes, les opérateurs de réseaux fixes offrant des services Triple Play (accès Internet, téléphonie et IPTV) envisagent de déployer une architecture de distribution de contenus avec des centres de données (DC) et des serveurs vidéo à la demande (VoD) situés à proximité des utilisateurs, généralement dans toute grande ou moyenne ville (au niveau du réseau d'agrégation et sur les bordures du réseau d'accès). Le déploiement d'une telle architecture, en s'appuyant également sur des accords passés entre opérateurs réseaux et opérateurs de distribution de contenu (CDN), devrait aider à réduire les dépenses de capital et d'exploitation (CAPEX et OPEX), la charge des serveurs et la latence des services. Par contre, comme l'architecture actuelle du réseau mobile est centralisée, les utilisateurs mobiles ne peuvent pas bénéficier d'une telle implémentation distribuée. En effet, aujourd'hui le traffic d'utilisateurs mobiles est

acheminé à travers un tunnel de bout en bout entre l'équipement de l'utilisateur (UE) et une passerelle de réseau de données par paquets distante (PDN-GW ou PGW), qui donne accès au réseau cœur-IP et donc aux services Internet [4]. L'utilisation d'un tunnel de bout en bout sécurise la connexion et permet aux utilisateurs mobiles d'avoir une mobilité transparente avec une bonne qualité de service (QoS). En revanche, même lorsqu'un utilisateur demande un contenu disponible sur un serveur géographiquement proche de lui, le traffic demandé doit d'abord être envoyé à la PGW dans le réseau cœur mobile (EPC) pour ensuite être encapsulé dans un tunnel pour être acheminé vers l'UE. Ce routage sous-optimal conduit à gaspiller les ressources du réseau et à surcharger les éléments du réseau cœur mobile. Cette politique de routage est sous-optimale pour les opérateurs convergents proposant des services fixes (Triple-Play) et mobiles.

Afin de permettre un routage optimal de données mobiles, l'organe de normalision 3GPP a défini des nouvelles méthodes permettant aux utilisateurs mobiles d'avoir accès aux services IP soit au niveau du réseau d'agrégation ou même au du réseau d'accès. Les bénéfices à prévoir, en termes de gain en bande passante, dus au déploiement d'une telle architecture mobile distribuée ont été mis en évidence comme première contribution de cette thèse [5]. Cette publication montre que, bien qu'une architecture mobile centralisée ait été justifiée jusque dans un passé récent, les gains dus à la distribution augmenteraient rapidement au cours des prochaines années, et ce dès 2020.

L'une des méthodes les plus prometteuses pour implémenter cette architecture mobile distribuée est SIPTO (sélection du traffic IP à décharger). SIPTO permet à un opérateur mobile de servir les demandes des utilisateurs soit directement sur le réseau local en utilisant des femto-cellules, sinon au-delà du réseau d'accès radio (RAN) l'aide des macro-cellules. En particulier, grâce à SIPTO un opérateur mobile peut sélectivement décharger une partie du traffic IP de l'utilisateur, tout en continuant à acheminer l'autre partie du traffic vers le cœur du réseau mobile (EPC). L'un des principaux objectifs pour l'utilisation de SIPTO est d'assurer une meilleure connectivité de services, c'est-à-dire permettre à un UE d'utiliser le meilleur chemin de données disponible vers le réseau IP externe. Toutefois, 3GPP a identifié, dans le cadre de SIPTO [6, 7], certains cas d'usage de la mobilité dans lesquels la continuité de la session n'est pas supportée par les procédures 3GPP actuelles. Dans certains cas, typiquement lorsque SIPTO s'appuie sur l'utilisation des passerelles locales (Local Gateways, LGW) ou bien d'une PGW alternative localisée au-delà du réseau d'accès radio, mais au plus près de l'UE, la mobilité de cet UE nécessite la modification de l'adresse IP attribuée à l'UE, ce qui entraîne l'interruption des sessions actives.

La continuité de services et la sélection des chemins de données sont considérées comme des problèmes clés à résoudre au sein du 3GPP. Ceci est du au fait que les solutions

proposées pour le Mobile IP ne fonctionnent qu'au niveau IP et non au niveau de l'accès à la couche IP, contrairement aux mécanismes définis par le 3GPP.

La proposition cœur de cette thèse, "Smooth SIPTO", assure une mobilité transparente pour les utilisateurs de SIPTO. Smooth SIPTO combine l'utilisation de MPTCP avec les procédures classiques de handover spécifiant le transfert des données mobiles entre les différentes cellules durant la mobilité des utilisateurs. La contribution consiste à faire fonctionner MPTCP sur l'architecture LTE et à modifier (à la marge) les procédures définies par 3GPP pour supporter le handover afin de permettre la continuité de services dans une architecture mobile distribuée. Smooth SIPTO est basé sur l'idée de connecter systématiquement l'usager à une PGW centralisée au niveau de l'EPC afin de garantir la continuité de services durant la mobilité des utilisateurs, tout en le connectant également à une LGW ou à une PGW géographiquement plus proche afin de favoriser la distribution du trafic. MPTCP est utilisé pour permettre à l'UE de recevoir le trafic d'une unique session sur plusieurs chemins simultanément (en phase de handover) ou alternativement (une fois le handover terminé).

Smooth SIPTO s'inscrit dans le cadre général de la convergence fixe mobile (FMC). L'architecture FMC a d'abord été considérée dans le cadre du service téléphonique, mais ce concept est maintenant devenu plus large, en intégrant un support global de la convergence fixe et mobile au niveau du service, c'est-à-dire les services IP et IMS [8], ainsi que la convergence fixe et mobile au niveau du réseau [9, 10]. Les opérateurs de réseaux envisagent actuellement le déploiement de la FMC afin d'une part de mieux supporter les services 5G et d'autre part de pouvoir partager l'architecture des services entre les utilisateurs fixes et les utilisateurs mobiles.

La troisième contribution de cette thèse est l'analyse de la mise en œuvre des solutions proposées pour smooth SIPTO sur les architectures FMC proposés dans le cadre du projet européen COMBO. J'ai présenté tout d'abord un bref résumé des travaux réalisés dans le cadre du projet européen COMBO Project [11] concernant la convergence fonctionnelle des réseaux fixes et mobiles [12, 13]. Ensuite, l'applicabilité des procédures proposées pour smooth SIPTO sur les architectures FMC a été évaluée en alignant les fonctions mobiles 3GPP, ainsi que les différents éléments fonctionnels du smooth SIPTO, sur les blocs fonctionnels proposés dans COMBO qui formalisent la gestion universelle des chemins de données (uDPM) entre client et réseau. Plusieurs architectures permettant d'implémenter les procédures proposés pour smooth SIPTO sur les entités fonctionnelles convergés fixe-mobile du projet COMBO ont été considérées. Par la suite, sachant qu'aucune des architectures FMC proposées par COMBO n'a considéré l'implémentation d'une PGW centralisée, et que les solutions proposées pour un smooth SIPTO sont basés sur l'idée de connecter l'usager à une PGW centralisée au niveau

de l'EPC afin de garantir la continuité de services durant la mobilité des utilisateurs, j'ai introduit la notion d'un "Anchor PGW" qui permet à un opérateur de supporter la signalisation MPTCP dans une architecture FMC. J'ai enfin décrit comment smooth SIPTO fonctionnait lorsque l'utilisateur et/ou le serveur ne supporte pas la MPTCP.

La dernière contribution de cette thèse se concentre sur le support de la mobilité des utilisateurs lorsque les solutions smooth SIPTO sont implémentées dans les architectures COMBO centralisées et distribuées [13]. Pour chaque architecture FMC, j'ai identifié les différents scénarios de mobilité pour lesquels j'ai considéré différents emplacements potentiels des fonctions du plan de données et de plan de contrôle FMC. Ensuite, une étude quantitative a été réalisée afin d'estimer le temps d'interruption potentiel et le volume de signalisation lorsque les scénarios de mobilité de smooth SIPTO sont implémentés dans chacune des architectures proposées par COMBO. En utilisant cette évaluation, j'ai défini les scénarios de déploiement préférés pour chacune des propositions pour smooth SIPTO. Enfin, j'ai proposé une évaluation qualitative de la manière dont ces scénarios s'appliquent aux cas d'usage typiques de la 5G.

La conclusion du manuscrit ouvre des perspectives relatives au support de smooth SIPTO par les nouveaux paradigmes de réseaux tels que la virtualisation de fonctions des réseaux (NFV) ainsi que le découpage du réseau en morceaux (Network Slicing). Typiquement, dans une architecture NFV, un opérateur FMC peut considérer le déploiement des machines virtuelles spécifiques pour le plan de données et d'autres machines virtuelles pour le plan de contrôle. Par ailleurs, pour le Network Slicing, une des implémentations possibles est de dédier des éléments fonctionnels FMC qui servent uniquement les utilisateurs machines (tels que Internet of Things ou IoT) et d'autres spécifiques pour les utilisateurs humains.

Mots clés : Réseaux mobiles 5G, Réseau LTE, Convergence Fixe-Mobile, Distribution de données, Déchargement de données mobiles, SIPTO, MPTCP, IoT, NFV, Network Slicing.

Abstract

The present Ph.D thesis was performed in the context of the FP7 European project COMBO (COnvergence of fixed and Mobile BrOdband access/aggregation networks), started on January 1st, 2013, under a joint collaboration between 16 partners, most notably: Orange, Telefõnica, Deutsch Telecom, Telecom Bretagne, Budapest University of Technology and Economic, Politecnico di Torino, Ericsson, FON, ADVA.

Up to now, fixed and mobile networks have been optimized and evolved independently. Standardization work groups dealing with fixed and mobile networks separately address their respective deployments, e.g. 4th Generation (4G) mobile networks and Fiber To The Home (FTTH) fixed networks. COMBO proposes a joint optimization of fixed and mobile access / aggregation networks around the innovative concept of Next Generation Point of Presence (NG-POP).

In the future, as IP services are moving towards high definition quality, existing fixed and mobile networks shall have to support steeply increasing demands in terms of data traffic volume. To cope with these demands, fixed network operators are deploying content services and data centres close to the users' locations, within the boundaries of aggregation and access networks. However, as the current mobile network architecture is centralized, mobile users cannot benefit from such a distributed implementation.

Mobile data offloading approaches such as those defined by 3GPP represent some of the most promising solutions that allow mobile users to access IP services at the aggregation or even the access segments of the network. The benefits, in terms of bandwidth gain, to be expected were mobile traffic offload made possible are highlighted as a first contribution of this Thesis. It is shown that, while a centralized mobile architecture was justified till now, expected gains would quickly increase in the next few years, as soon as 2020.

A distributed mobile architecture relying on SIPTO does not support session continuity during users' mobility. Session continuity and data path selection have indeed been considered as key issues to be solved within 3GPP, as none of the solutions proposed for Mobile IP directly apply in the context of LTE. In some cases, typically when SIPTO relies on using Local Gateways or alternate Packet Data Network Gateway, user's mobility makes it necessary to change the IP address allocated to a User Equipment, yielding the interruption of active sessions.

The core proposal of this thesis is a novel approach supporting seamless mobility for users relying on SIPTO-based mobile access. We propose to combine the use of MPTCP

together with the handover procedures specified for LTE in order to provide a “smooth Handover” in those cases. We describe how to make MPTCP operate over the LTE architecture and how the procedures defined by 3GPP to support handover should be (slightly) modified in order to enable session continuity.

The applicability of the above procedures on candidate FMC architectures proposed by COMBO is next assessed by mapping the 3GPP mobile functions, as well as the different smooth SIPTO functional elements, on the proposed COMBO functional blocks for universal Data Path Management (uDPM). We then propose several architectures for implementing smooth SIPTO procedures on COMBO FMC functional entities. Thereafter, we map the different mobility scenarios of smooth SIPTO proposals on both Centralized and Distributed COMBO architectures.

Finally, a quantitative study is carried out to estimate the potential interruption time and signalling volume when smooth SIPTO mobility scenarios are implemented within both centralized and distributed COMBO architectures. Using these evaluations, we define preferred deployment scenarios for each of the smooth SIPTO proposals, and qualitatively assess how these scenarios apply to typical 5G use cases.

A conclusion opens perspectives relative to the support of smooth SIPTO by the new networking paradigms Network Function Virtualization and Network Slicing.

Key words : 5G Network, LTE Network, FMC, Content distribution, Mobile data offloading, SIPTO, MPTCP, IoT, NFV, Network Slicing.

أُهدِي هَذِهِ الأُطرُوحَة إِلَى أَغلَى إِنسَانَة فِي حَيَاتِي، إِلَى
مَن فَرِحَت لِفَرَحِي وَحَزِنَت لِحُزنِي وَسَهِرَت اللَّيَالِي تَرعَانِي فِي تَعَبِي.
إِلَى التِي هِيَ أَغلَى عِندِي مِن رُوحِي، إِلَى...

حَبِيبَتِي أُمِّي

I dedicate this thesis to the most precious human being in my life,
to whom that rejoiced to my happiness and sorrowed to my sorrow,
to whom that woke-up during nights taking care of me in my
sickness... to the person the most precious to my soul...
My dear Mother

Acknowledgements

Many acknowledgements and thanks go to all the persons who have contributed in the achievement of this work, by their advice, encouragements and constructive remarks. This thesis is the harvest of four years of research work whereby I had the opportunity to benefit of real on-job training and learning, combined with a considerable and valuable support from many persons, to whom I am glad to show hereby all expressions of gratitude.

Lots of expressions of gratitude go first to my direct thesis supervisor Mme. Annie GRAVEY, for integrating me within her motivated and dynamic research team and accepting me among her PhD students, for directing my work and for the mark of trust that she has manifested towards me. Many of the professional qualities that I have developed during these years are owing to the valuable experience that I had accumulated while working in her team. She devoted a special attention to the progress of my research and supplied me with valuable technical advices whenever I needed them.

I would like to thank all the professors who were responsible for the PhD courses that I have attended during these years. I have acquired considerable knowledge through their presentations and interactive scientific discussions.

Special thanks and signs of acknowledgement go particularly to the members of my PhD committee for the interest that they have shown in my thesis and for their efforts in reading and providing valuable comments on the dissertation of my PhD thesis: Mr. Hossam AFIFI, Mr. Farid NAÏT ABDESSELAM, Mr. Xavier LAGRANGE, Mme. Houda LABIOD, Mr. Yassine HADJADJ AOUL, Mr. Philippe BERTIN and Mr. Stéphane GOSSELIN.

This work is involved in the COMBO FP7 European project, under a joint collaboration between 16 partners, including Orange, Telefonica, Deutsch Telecom, IMT-Telecom Bretagne, Budapest University of Technology and Economic, Politecnico di Torino, Ericsson, FON, ADVA. I thank all the persons with whom I had the opportunity to work on this project, it was a pleasure to work with them.

I would also express my deep sense of gratitude particularly to my parents, my sisters and to my brothers, especially Oussama EIDO, for being always there for me and for giving me the best support and the best working conditions. All signs of gratitude go also to my love Mahdi, for his affection, encouragements and precious support during these years.

I would like specially to present all my gratitude to Mr. El Hadj Amadou TOURÉ, the director of the student residence of Telecom Bretagne, for helping me during the long period of illness that I have experienced in the last year of my PhD.

I express my warm thanks to my brother Thierry EIDO (Product Manager at SFR) and to my colleague Moufida FEKNOUS from Orange Labs, for their support and for the many discussions about fixed and mobile networks which helped me to improve my work. I also present my sincere gratitude to all my friends and colleagues, especially: Julie SAUVAGE-VINCENT, Ahmed TRIKI, Souhaila FKI, Marisnel OLIVARES, Iyas ALLOUSH, Serge-Romaric TEMBO, Pratibha MITHARWAL, Ion POPESCU, Bogdan USCUMLIC and Simona ANTIN for their friendliness, conviviality and the sympathetic moments that we shared together.

Specially, countless thanks and acknowledgements go to my beloved mother Mme. Amal KANAWATI who supported me and took care of me during the hardest time I went through during this PhD.

Contents

List of Figures

List of Tables

Abbreviations

3GPP	3rd Generation Partnership Project
AAA	Authorization and Accounting
ADSL	Asymmetric Digital Subscriber Line
APN	Access Point Name
AS	Autonomous System
ATM	Asynchronous Transfer Mode
BAS	Broadband Access Servers
BNG	Broadband Network Gateway
CAGR	Compounded Annual Growth Rate
CAPEX	Capital Expenditure
CDN	Content Distribution Network
cMTC	critical MTC
CMT-SCTP	Concurrent Multipath Transfer for Stream Control Transmission Protocol
CN	Concentration Node
CO	Central Office
CtP	Content Provider
CP	Control Plane
C-SIPTO	Change for SIPTO
DC	Data Centers
DHCP	Dynamic Host Configuration Protocol
DL	Down Link
DP	Data Plane
DRM	Digital Rights Management
DSL	Digital Subscriber Line
DSLAM	Digital Subscriber Line Access Multiplexer

EB	ExaBytes
eMBB	enhanced Mobile Broadband
eNB	evolved Node B
EPC	Evolved Packet Core
E-RAB	E-UTRAN Radio Access Bearer
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
FMC	Fixed and Mobile Convergence
FTTB	Fiber To The Building
FTTC	Fiber To The Curb
FTTH	Fiber To The Home
GMDT	Global Mobile Data Traffic
GSM	Global System for Mobile communications
GTP	General-Packet-Radio-Service Tunnelling Protocol
GTP-C	GTP for Control plane
GTP-U	GTP for User plane
HD	Hight Definition
HeNB	Home evolved NodeB
HGW	Home Gateway
HO	Handover
HSS	Home Subscriber Server
IMS	IP Multimedia Subsystem
IoT	Internet of Things
IoT-GW	IoT service Gateway
IP	Internet Protocol
IPTV	Internet Protocol Tele Vision
ISP	Internet Service Provider
ITU	International Telecommunication Union
ITU-T	ITU - Telecommunications Standard Sector
KB	KiloBytes
KM	KiloMetres
LGW	Local Gateway
LHN-id	Local HeNB Network identifier
LIPA	Local IP Access

LN	Local Network
LTE	Long Term Evolution
LTE-Uu	LTE-User universal-mobile-telecommunications-system
MC	Mobility use-Cases
MEC	Mobile Edge Computing
MIMO	Multiple-Input Multiple-Output
mIoT	massive Internet of Things
MME	Mobility Management Entity
mMTC	massive Machine Type Communications
mmWave	millimetre Wave
MN	Multi-service Nodes
MOCA	Mobile Cloud Offloading Architecture
MPLS	Multi-Protocol Label Switching
MPTCP	Multi-Path Transmission Control Protocol
ms	millisecond
MTC	Machine Type Communications
NFV	Network Funcion Virtualization
NGMN	Next Generation Mobile Networks
NG-PoP	Next Generation Point of Presence
NR	New Radio
OLT	Optical Line Termination
ONT	Optical Network Termination
ONU	Optical Network Unit
OPEX	operational expense
OTN	Optical Transport Network
OTT	Over The Top
PCEF	Policy Control Enforcement Function
PCRF	Policy Charging Rules Function
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
P-EGW	PDN Edge Gateway
PGW	Packet Data Network Gateway
PLMN	Public Land Mobile Network

PON	Passive Optical Network
PoP	Point of Presence
PPP	Point-to-Point Protocol
PSTN	Public Switched Telephone Network
PtP	Point to Point
QoS	Quality of Service
QUIC	Quick UDP Internet Connections
RAN	Radio Access Network
RFI	Radio Frequency Interference
RGW	Residential Gateway
RRC	Radio Resource Control
S1-AP	S1 Application Protocol
SAE	System Architecture Evolution
SDH	Synchronous Digital Hierarchy
SDN	Software-Defined Networking
SGW	Serving Gateway
SIPTO	Selected IP Traffic Offload
TAI	Tracking Area Identifier
TAU	Tracking Area Update
TCP	Transmission Control Protocol
TEID	Tunnel Endpoint Identifier
UAG	Universal Access Gateway
uAUT	universal Subscriber and User AUThentication
UDP	User Datagram Protocol
uDPM	universal Data Path Management
UE	User Equipment
UHD	Ultra-High-Definition
UL	Up Link
URLLC	Ultra-Reliable Low latency Communications
VDSL	Very-high-bit-rate Digital Subscriber Line
ViLTE	Video over LTE
VoD	Video on Demand
VoIP	Voice over IP

VoLTE	Voice over LTE
WDM	Wavelength Division Multiplexing
WP	Work Package
ZB	ZettaBytes

Chapter 1

Introduction

Today, fixed and mobile networks are experiencing a dramatic growth in terms of data traffic. For instance, according to Cisco [1], the annual global IP traffic has surpassed the Zettabytes (zb; 1000 EB) threshold in 2016, and is expected to reach 2.3 ZB by 2020, which represents 95 times the volume of the global IP traffic in 2005. This growth is resulting from the increasing numbers of connected devices (smart-phones, tablets, PCs, laptops, etc.) and the demand for Internet-based services such as on-line games, Internet Protocol TeleVision (IPTV), high definition video (streaming, conferencing, etc.) [2][3]. Such a demand is expected to increase potential congestion risks by increasing the traffic load in the various network portions (access, aggregation and core) and thus to potentially degrade delivered QoS . In order to deal with this growth, network operators have to upgrade network capacity by making enormous investments in network infrastructures (e.g., multiply the number of nodes, deploy fiber based transport technologies, implement heterogeneous access architectures, etc.). Operators are thus seeking for new cost-effective solutions while ensuring a high performance of both fixed and mobile networks.

Currently, network operators offering Triple-Play services (Internet Access, Telephony and IPTV) consider deploying a content distribution architecture with Data Centers (DC) and Video-on-Demand (VoD) servers located close to the users, typically in every large or medium town. The deployment of such an architecture, possibly relying on agreements with Content Distribution Network (CDN) operators, is expected to help reducing both capital and operational expenditures (CAPEX and OPEX) [5], as well as server loads and latency delays for end-users. On the other hand, regarding the actual Long Term Evolution/System Architecture Evolution (LTE/SAE) (also called "4G") mobile network infrastructure, CDN architecture does not currently help optimizing mobile back-haul network. This is due to the fact that the LTE/SAE mobile architecture

is implemented nowadays by routing user's traffic through an end-to-end tunnel between the User Equipment (UE) and a distant Packet Data Network Gateway (PDN-GW or PGW), which allows the UE to access the IP backbone and the Internet [4]. The use of an end-to-end tunnel secures the connection and allows a seamless mobility with high Quality of Service (QoS) for mobile users. However, even when a user requests a content that is available in a geographically close server, the requested traffic must first be sent to the PGW in the Evolved Packet Core (EPC) network and then be tunneled and sent toward the UE. This sub-optimal routing leads to wasting network resources [5]. This routing policy is sub-optimal for converged operators offering both fixed (Triple-Play) and mobile services.

3GPP has proposed the Selected IP Traffic Offload (SIPTO) approach in [6] in order to selectively breakout some of the mobile IP traffic either directly at the local network using femto cells, or above the Radio Access Network (RAN) using macro cells. One of the main objectives for using SIPTO is to ensure a better connectivity service, i.e. to allow a UE using the best available data path towards the external IP network. SIPTO is used to re-assign a PGW, either co-located with the radio base station, or deployed as a separate entity; this alternate PGW would ideally be geographically closer to the current UEs locations than the PGW in the EPC. Consequently, non-offloaded traffic could be routed towards the EPC network while SIPTO traffic would be offloaded within the access/metro segment of the network.

Operators are currenly considering the deployment of a Converged Fixed and Mobile (FMC) network architecture, in order to better support 5G services and to mutualize service architecture between fixed and mobile users. FMC architecture was first considered in the framework of supporting telephone services but this concept is now broader, including a global support of fixed and mobile convergence at service level, i.e. IP services and IMS [8], and also fixed and mobile convergence at network level [9, 10] .

3GPP has identified, in the framework of SIPTO ([6], [7]), some mobility use cases in which session continuity is not supported due to the fact that the UE's mobility implies PGW relocation (and thus a modification of the UE's IP address), which is not supported by the current 3GPP procedures.

Through this thesis, in the context of future fixed mobile converged network architectures, we quantify the gain of bandwidth, in terms of offloaded traffic, brought by the generalized use of SIPTO approaches. We also propose a solution to provide seamless mobility for users relying on SIPTO. Lastly, we describe how this solution maps on different architecture options considered for fixed mobile convergence.

The present report is organized as follows:

Chapter 2 provides an overview of today's fixed, mobile and FMC network architectures. We first describe the potential fixed network implementation options over the different segments of the network: access, aggregation and core. Then, a global vision about the evolution of the mobile network architectures is depicted with a description of standardized mobile data offloading approaches that have been considered to enhance the currently deployed mobile architecture (4G) by offloading part of the mobile IP traffic within the access and aggregation segments of the network, and thus alleviating the load of the core network. Next, we present the different mobility scenarios currently deployed and/or standardized by 3GPP. The goal of this study is to identify the mobility use cases where session continuity is not ensured. We then compare these use cases with existing solutions and related works. Finally, we end this Chapter by providing an overview of CDN, MPTCP and FMC architectures.

In Chapter 3, we present a quantitative evaluation of the potential gain of bandwidth to be achieved assuming that SIPTO approaches are actually deployed. The study in this Chapter is based on [17] highlighting the traffic increase affecting metropolitan (or aggregation) network. In order to avoid certain bottlenecks at the core segment of the network, advanced scenarios based on content distribution are presented. The main idea of these new models is to deploy several content servers close to users' locations, typically at the edge of the backbone network and at the boundaries between access and metropolitan networks . Finally, this Chapter provides a detailed quantitative multi-criteria comparison between the different cases in terms of bandwidth gain and cost efficiency. The work reported in this Chapter has been published and presented in EUNICE 2014 [5].

Chapter 4 proposes "smooth SIPTO", which is a solution providing seamless mobility for users with ongoing SIPTO sessions when PGW relocation is performed. This solution relies on the Multi-Path Transmission Control Protocol (MPTCP) proposed recently by IETF [18]. We first explain how the initial SIPTO data path is established along with the MPTCP connection. During this establishment procedure, the MPTCP connection is established over the default (LTE) data path which shall be used for all MPTCP signalling messages. We then identify the key requirements allowing an operator to maintain the session continuity for users with ongoing SIPTO above RAN or SIPTO at Local Network (LN) sessions. In order to support smooth SIPTO at LN, we introduce a Proxy-SGW function, which is to be included within the LGW. This function is only seen by the co-located HeNB and LGW functions, which helps maintaining the global 3GPP network infrastructure unchanged. The sequence diagrams for proposed smooth SIPTO' handover procedures are then illustrated. These diagrams are based on the "inter eNB/inter SGW" S1-based handover procedure shown in A. Finally, we compare

our proposed smooth SIPTO mobility scenarios with the Classical LTE mobility scenario for S1-based handover previously described in Section 2.3.1. Compared with the currently proposed solutions for ensuring session continuity, the novelty of our solution is that it is compatible with the current 3GPP architectures. It also allows breaking out long-lived users data traffic directly at the local network, a solution which none of the existing works have yet proposed. This work was carried out in the frame of the European project COMBO [11] for fixed and mobile convergence.

In Chapter 5, we first present a brief summary of the work carried out within the European COMBO Project [11] regarding functional convergence of fixed and mobile networks [12], [13]. We then outline the possible mappings of the Classical LTE as well as the smooth SIPTO architectures on COMBO's Functional blocks. Finally, we present the possible mapping scenarios of smooth SIPTO architectures on each of the functional FMC network architectures proposed by COMBO.

In Chapter 6, we presented how mobility in smooth SIPTO can be supported within an operational FMC network. For each FMC architecture, we identify different mobility scenarios during which we consider different potential locations of FMC data plane and control plane functions. Next, we evaluate smooth SIPTO handover proposals on COMBO architectures, in terms of interruption time duration and signalling volume. These evaluations are then used to assess which implementation of the COMBO architectures allows a better applicability of each of the smooth SIPTO handover proposals. This in turn allows to identify which location would be the best for deploying data plane and control plane functions when smooth SIPTO handover solutions are implemented. Finally, we present how COMBO architectures can be applied to implement advanced 4G and 5G mobile architectures.

The work presented in Chapter 4, Chapter 5 and Chapter 6 was carried out within the European research project COMBO. Part of these results have been published in [19].

A final Chapter concludes the thesis report, recapping the main results that have been obtained and proposing some directions for further work.

Chapter 2

State of Art

This chapter first presents an overview of today's fixed network architecture. The fixed network is divided into different segments (Access, Aggregation and Core networks) and services are delivered to network customers over each of these segments. In the second part of this chapter, an overview of classical LTE and beyond LTE mobile architectures is described. The next section outlines the different mobility scenarios considered for both mobile architectures. Mobility with respect to session continuity support is considered as one of the key points to be ensured by operators when deploying any mobile architecture.

In Section 2.4, we present an overview of the existing proposals regarding the fixed and mobile convergence. We also outline the benefits that could be achieved when such integrated network is implemented.

In the next section, an overview of existing multiple IP addresses management approaches for mobility support is presented. Nowadays, mobile devices are disposed with multiple available interfaces that could be used over multiple access technologies. These approaches are particularly used to allow a dynamic control of the delivery of data originally targeted for cellular network regardless of the access network.

Finally, we present an overview of content distribution and its benefits for both the user and the operator. For instance, today, operators are considering content distribution as one of the most promising solutions used to improve the user's experience by serving users demands from servers that are located close to their locations.

2.1 Fixed Network Architecture

The fixed network is composed of 3 segments: the access network, the metropolitan/regional aggregation network, and the core network. The network segments are interconnected hierarchically by a set of equipment ensuring that data traffic is correctly distributed. Figure 2.1 illustrates the global architecture of the fixed network.

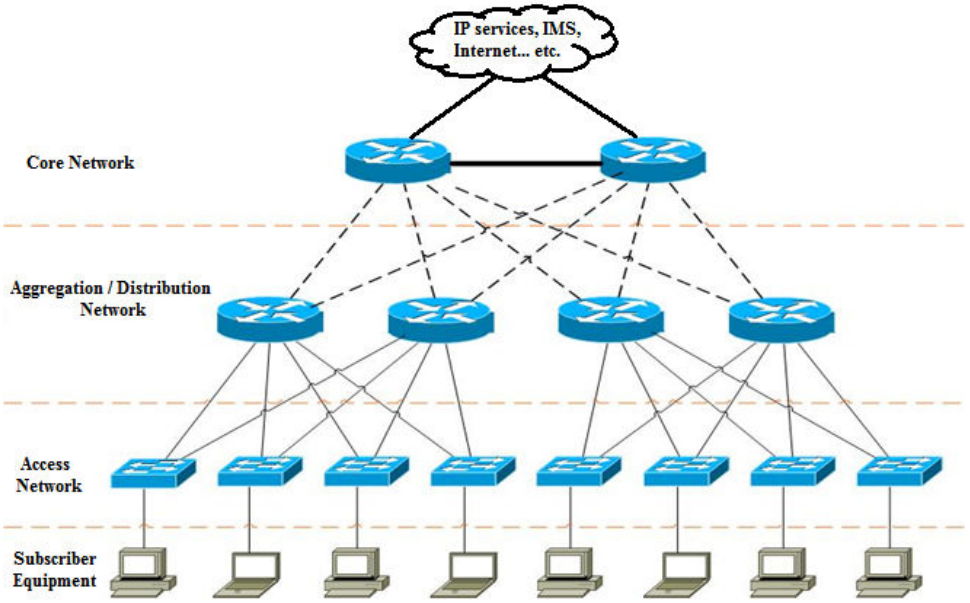


FIGURE 2.1: Fixed Network Architecture

In this section we present an overview of the fixed network architecture. We first describe how fixed traffic is distributed towards the end users. Then, we explain how packets are being aggregated between access and core networks. Finally, we conclude this section by showing how the core network gives access to the IP network services.

2.1.1 Fixed Access Network

The fixed access network consists in a set of equipment used to connect end users (subscriber equipment) to the aggregation network. Typical access equipment are Digital subscriber Line Access Multiplexer (DSLAM) and Optical Line Termination (OLT). In the access segment of fixed network, services are carried to customers either over copper access technologies such as xDSL [38], or fibre access technologies, e.g., FTTx.

The Digital Subscriber Line technology is deployed over the existing telephone lines starting from the central offices (CO). It provides a high speed data digital transfer using specific modulation and coding mechanisms. It is typically used for Voice over IP (VoIP) and broadband access: Internet, phone, IP television, etc. However, this

technique presents some limitations as the available bandwidth depends strongly on the copper quality and loop length (i.e. distance between the CO and the home). For instance, the Asymmetric DSL (ADSL) limits the loop length to a maximum of 7 km. Moreover, the upstream traffic is often limited to a few hundreds of Kbits per second. ADSL flows are concentrated by DSLAMs. According to France Telecom (Orange's network) in [20], a DSLAM connects an average of 900 users.

Fibre based access technologies are proposed by telco operators in order to deliver access services to end users while reducing the access bandwidth bottlenecks thanks to optical fibre. The use of optical fibre for traffic delivery allows increasing the distance between the CO and final customer. Typically, multiple users (up to 64 or 128 users) can share a single fibre that is used for implementing a Passive Optical Network (PON). For instance, the 10 Gigabit Passive Optical Networks (10GPON) solution standardised by ITU-T in [21] to replace the current G-PON systems supports a range of optical budgets between 33 decibel (dB) and 35 dB. A PON with such an optical budget supports a loop length as long as- 25 km.

In a fixed network architecture with optical fibre, three possible implementations for FTTx technologies have been considered [22]:

- Fiber To The Building (FTTB)
- Fiber To The Home (FTTH), which provides the fiber directly to the residential/home network enabling data rate capacities of several hundreds Mbit/s.
- Fiber To The Curb (FTTC)

The selection of an FTTx model depends on how close the fibre should get to the subscriber. Figure 2.2 illustrates today's fixed access network architecture with both potential access technologies (xDSL and FTTx). In this figure five access cases are identified relying on the same infrastructure for the aggregation network and the interconnection to the Internet Service Provider (ISP). The first two cases shows DSL access models using legacy copper access technology. The last three cases on the other hand, define the potential Fibre access scenarios using new fibre infrastructure to the users.

Case 1: DSL access architecture In this case the DSL modem is located in the Residential Gateway (RGW) while DSLAMs are located at the CO. Each user has a dedicated point-to-point DSL connection carried on the copper loop linking the user to the CO.

Case 2: VDSL access architecture The Very-high-bit-rate digital subscriber line (VDSL) access aims to improve the upstream and downstream rates offered to the user

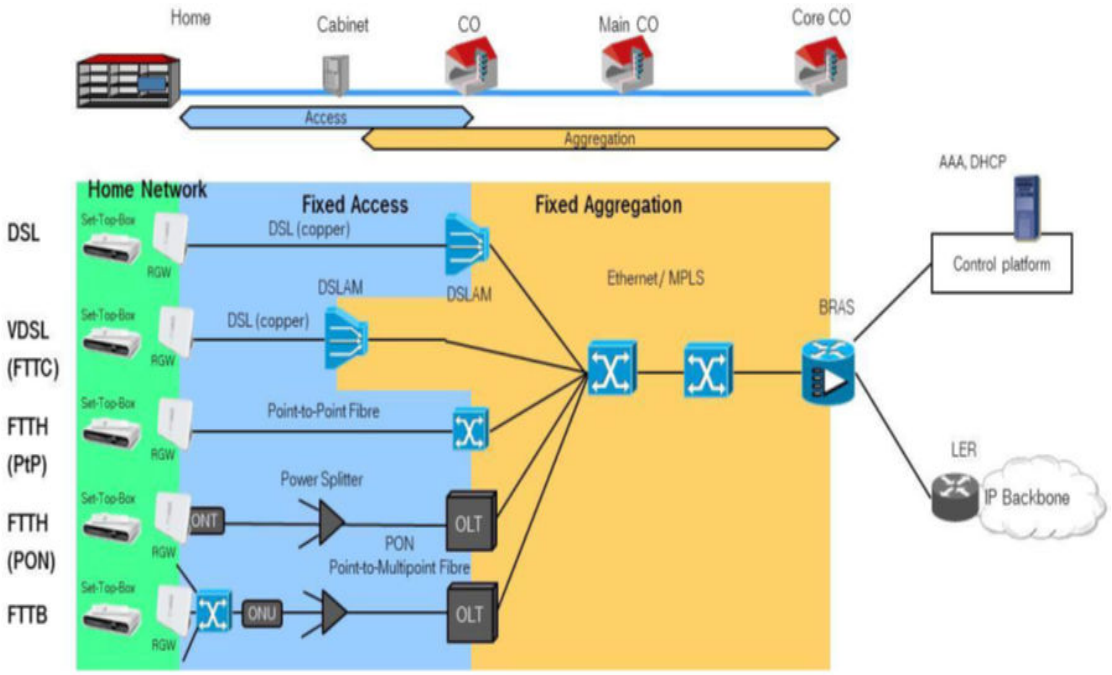


FIGURE 2.2: Fixed Access Network Architecture [14]

by deploying the DSLAM closer to the user's location, in the so-called "cabinet". The cabinet is the first operator network point in where a first level of aggregation is performed. In traditional architectures, the cabinet hosts only passive equipment, whereas in a VDSL architecture, a DSLAM is located within the cabinet (i.e. the cabinet supports active equipment).

Case 3: PtP-based FTTH access architecture This architecture relies on a Point-to-Point (PtP) fibre between the user and the CO. An FTTH modem with optical interfaces replaces the DSL modem in the RGW. Similar to the DSL based architectures represented in case 1 and 2, the PtP-based FTTH architecture requires an interface per user at the CO.

Case 4: PON-based FTTH access architecture unlike the PtP based FTTH architecture, the PON based architecture relies on a Point-to-Multipoint distribution tree. The FTTH modem includes an Optical Network Termination (ONT), dedicated to a single user. The Access Node (AN) in this architecture is presented by an active equipment located in the CO, called Optical Line Termination (OLT). One or several (passive) splitters are located between the OLT and the (multiple) ONTs. The OLT ensures the interconnection of PON with the aggregation network. It distributes the downstream traffic (e.g., data coming from aggregation network and service platform) toward the users, and arbitrates between the users sending upstream traffic. Downstream traffic is copied on all fibres at the splitter, and selected by each ONT. The splitters serve as

passive multiplexing equipment in the upstream direction between the ONTs and the OLT.

Case 5: FTTB access architecture In FTTB architecture, the optical modem is an Optical Network Unit (ONU). The ONU is an active element generally located in buildings or at the edge of a residential area. Ethernet multiplexing is used between the end users and the ONU modem, which implies that RGW now requires a simple Ethernet interface. A first multiplexing stage is provided by the ONU. Multiple ONUs are linked to an OLT, as in Case 4, thanks to a Multipoint-to-Point tree.

It is possible to co-locate DSLAMs and OLTs in a CO. However, fibre based and copper based access networks are separately managed.

2.1.2 Aggregation Network

The aggregation or "metropolitan" (metro) network is the intermediate network segment between access and core networks. The traffic from DSLAMs and OLTs is aggregated within "Edge Nodes" (ENs) which forward this traffic towards the core network. The aggregation network is linked to the core network at a "Point of Presence" (PoP). The PoP hosts several relevant devices, such as Multiservice Nodes (MN) and Broadband Access Servers (BAS):

- a MN is the node provides access to the service platforms of the operator (e.g., VoD and TV services).
- a BAS:
  - aggregates the upstream traffic coming from DSLAMs and/or OLTs, and forwards it to the core network,
  - distributes the downstream traffic towards DSLAMs and/or OLTs,
  - intermediates between the users and the operator's Authentication, Authorization and Accounting (AAA) mechanisms (e.g. setting up user's Point-to-Point Protocol (PPP) sessions, redirecting Dynamic Host Configuration Protocol (DHCP) traffic, etc.)

A metro network may be designed as a (logical) ring, or as a set of several interconnected rings, where a primary ring (Metro/Core) is connected to several secondary rings which represent the Metro/Access network. The basic idea of ring aggregation networks is derived from traditional Synchronous Digital Hierarchy (SDH) architecture. Figure 2.3 illustrates the aggregation segment of the fixed network architecture.

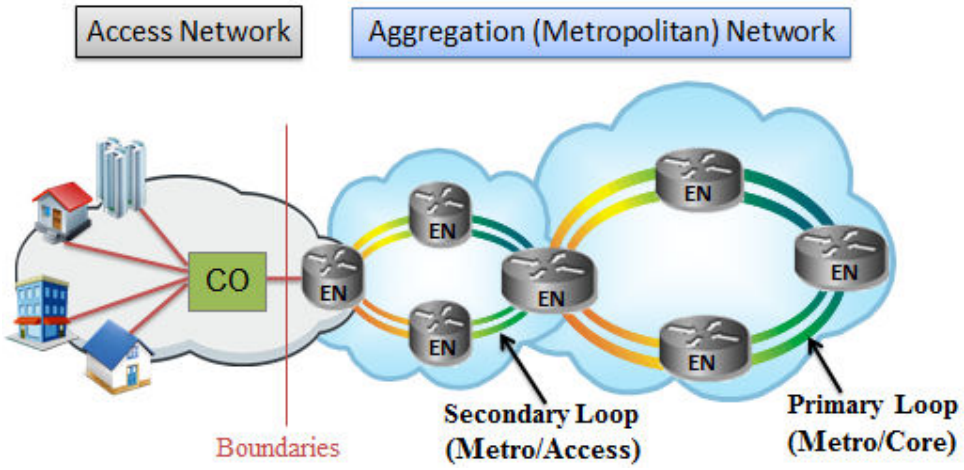


FIGURE 2.3: Metropolitan Network Architecture

The aggregation network is currently deployed on an Ethernet/Multi-Protocol Label Switching (MPLS) architecture relying on a Wavelength Division Multiplexing (WDM) architecture for transport between EN routers. Historically, aggregation networks were deployed on ATM over SDH architectures. Indeed, traffic forwarding within the aggregation network can be done either by using Asynchronous Transfer Mode (ATM) [23] or Ethernet. An ATM interconnection relies on interfaces with low bit rates (155Mbit/s, 622Mbit/s) [24], whereas an Ethernet interconnection, over a WDM transport network, may support higher bit rates (1Gbit/s or 1GE, 10Gbit/s or 10GE) [25]. The aggregation segment may support functions such as QoS management and traffic policies [26].

According to [20], which refers to the Orange network, there may be 10 to 30 ENs in an aggregation network. An EN located in a primary metropolitan ring aggregates at most 64000 users connected to an average of 70 DSLAM (i.e., 900 user per DSLAM).

Today, with the growth of both fixed and mobile networks data traffic predicted in [3], operators are looking for new solutions to cope with such a traffic increase with the lowest possible costs. Indeed, research to improve metropolitan network capacity is still active. Various proposed solutions include Conventional-Optical Burst Switching (C-OBS) [27], [28], [29], Labelled-OBS (L-OBS) [30], [31], Packet Optical Add/Drop Multiplexing (POADM) [32], [33], [34], [35], and Time-domain Wavelength Interleaved Networking (TWIN) [36], [37].

2.1.3 Core Network

The core network, also known as the backbone network, aggregates all customers' traffic.The first aggregation point in the core network is the Concentration Node (CN). It aggregates the traffic flows coming from the metropolitan network through several ENs

with higher bit rates (typically, from 1 to 10 GE) [38]. The core network connects the metropolitan network together and provides them with an access to the Internet and more generally to external IP networks. The backbone network is a national network as it typically covers an entire country. Backbone network topology is generally meshed on IP/MPLS and relies on Optical Transport Network (OTN) technologies.

Among the different telecom operators in France, the Orange operator for France Telecom represents the largest with around almost ten millions of Internet subscribers. Orange operator's core network is known as the RBCI (Réseau Backbone de Collecte IP). The RBCI is an IP Autonomous System (AS). It is recorded in global level as AS3215 [39].

Typical functions performed by core network are: user traffic routing to the Internet, users voice calls management, distribution of TV content, user management, policy and charging, lawful interception, network monitoring and management [40].

2.2 Mobile Network Architecture

This section presents an overview of the existing 3GPP mobile network architectures. We first present the classical LTE architecture. Then, we outline the different mobile data offloading mechanisms in a "beyond LTE" architecture model.

2.2.1 Classical LTE Architecture (4G)

Classical mobile architecture was first defined by 3GPP in Release 8 of technical specifications, where the following working groups were set up:

1. **Long Term Evolution (LTE) Group:** LTE was initiated by 3GPP in November 2004. The main objectives of this group were to improve the radio interface and optimise the architecture of the mobile access network. The results of the group's work were published in 2007 at which time the new radio access network, Evolved Universal Terrestrial Radio Access Network (E-UTRAN), was introduced for the first time [41]. At the end of 2009 , the first LTE deployments were announced by TeliaSonera and Verizon Wireless [42].

2. **System Architecture Evolution (SAE) Group:** In parallel to LTE, the SAE working group was defined by 3GPP in order to optimise the mobile core network and offer a complete IP-based architecture. As a successful outcome from SAE, the new Evolved Packet Core network (EPC) was specified by 3GPP in [43].

LTE and SAE working groups together defined the Evolved Packet System, which represents the "4G Mobile Network".

2.2.1.1 LTE Network Functions

E-UTRAN represents the cellular network between the mobile users and the core network. It consists of radio antennas and radio base station gathered in a single element called the evolved NodeB (eNB).

EPC represents the core network for the 4G mobile network. EPC was revised to separate signalling traffic (control plane) from user's data traffic (data plane). This functional split was motivated by 3GPP ([4], [44]) to allow an independent evolution and development of both software and hardware control. This would help operators to dimension and adapt their network easily by controlling their network behaviour using high level software programs.

The SAE group has introduced the following entities to the network [4]:

- Mobility Management Entity (MME):
  The MME handles all signalling traffic between the UE and the EPC network, using bearer and connection management functions such as user's attachment to the network, mobility, security, session activation/de-activation, etc.
- Packet data network GateWay (PGW):
  The PGW allows users to interconnect with an external IP network such as: the operator's IP service and IP Multimedia Subsystem (IMS) domains. The PGW filters each user's data plane packet into different EPS bearers, depending on the QoS information. Moreover, the PGW is responsible for UE's IP addresses allocation function.
- Serving Gateway (SGW):
  The SGW routes the IP packets of user's data plane between the eNBs and the PGW. It is considered as the mobility anchor for the inter-eNBs and the inter-operator's handovers. The SGW is responsible for buffering the downLink IP packets coming to users during their non-active or idle mode status.
- Home Subscriber Server (HSS):
  The HSS stores users' subscription data base. It holds the list of allowed Access Point Names (APNs) or PDNs connection information for each user in addition to any access restriction information for roaming.

- Policy Charging Rules Function (PCRF):
  the network element that manages the Policy Control Decision Making implements the PCRF. It also deals with the Flow-Based Charging functions in the Policy Control Enforcement Function (PCEF) which resides in the PGW.

The classical LTE architecture was designed to fully rely on IP, for both control plane and data plane traffics. This means that both voice and data services are carried over IP and are based on packet switching (contrary to 3G mobile networks that still relied on circuit switching for voice services).

Figure 2.4 shows the mobile signalling and data paths. The latter uses an end-to-end tunnel between the UE and the PGW. Considering the example when a user is using his phone to access an application server, an IP packet would first be created at the UE's level. This packet consists of application data, TCP or User Datagram Protocol (UDP) header for transport and then IP header carrying source address of UE and destination address of application server (e.g. Youtube).

- The IP packet is routed from UE to eNB over the LTE-User Universal-Mobile-Telecommunications-System (LTE-Uu) air interface with Packet Data Convergence Protocol (PDCP).
- Once received by the eNB, the IP packet is first encapsulated encapsulated in UDP and IP; the IP header of the packet now contains the eNB IP address as a source address and the SGW IP address as a destination address. The packet is then encapsulated in the General-Packet-Radio-Service Tunnelling Protocol (GTP-U) for User plane with a header carrying information related to Tunnel Endpoint IDentifiers (TEIDs). Lastly, it is switched in Ethernet towards the SGW over the (S1-U).
- The SGW encapsulates the packet in UDP and IP; the IP header of the packet now contains the SGW IP address as a source address and the PGW IP address as a destination address. The packet is then encapsulated in GTP-U, with a header carrying information related to TEIDs, and forwarded it to the PGW over the S5 interface.
- The PGW decapsulates the packet and forwards it towards the external IP network, thanks to the original IP header.

Furthermore, signalling paths rely on the following protocols:

- Radio Resource Control protocol (RRC) on the LTE-Uu air interface between the UE and the eNB;

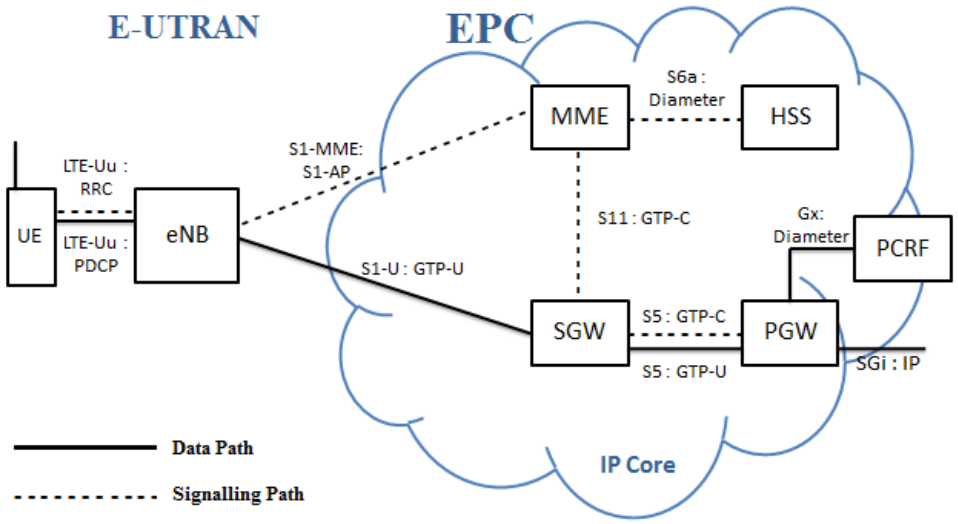


FIGURE 2.4: Today's Mobile Network Signaling and Data Paths

- S1 Application Protocol (S1-AP) on the S1-MME interface between the eNB and the MME;
- GTP-Control (GTP-C) protocol on the S11 interface between the MME and the SGW and the S5 interface between SGW and PGW;
- Diameter protocol on the S6a interface between the MME and the HSS;
- Diameter protocol on the Gx interface between the PGW and the PCRF.

2.2.1.2 Connectivity in Classical LTE Network

The mobile architecture is built on an "Always-On" connectivity concept. According to 3GPP in [4], two connectivity state models are identified:

1. **Idle Mode:** A network connection is set with Idle mode when a UE remain inactive for a specified time. The aim of such a connectivity is to reduce the load of the RAN network by performing a "Resource Release Procedure [4]" where all allocated signalling paths between the UE and the EPC network are released. At this level, the eNB covering the UE's location area would have no knowledge of the UE context. To stay reachable by the network during mobility, an idle mode UE has to choose a feasible radio network and a cell to camp on. Therefore, a "Public Land Mobile Network (PLMN) selection" and cell selection and re-selection" functions must be performed. The UE then has to report its location (Tracking Area Identifier (TAI)) to the HSS via a "Tracking Area Update Procedure" provided by 3GPP in [4] and illustrated in Appendix A. The TAI would then be used to reconnect the user to the EPC if a service is available.

2. **Connected Mode:** A network connection could change from idle to connected mode when a service is either requested by the UE or triggered by the network:

   - Service Requested by the UE: When a UE that is in idle mode has to send an uplink packet to the EPC, a "Service Request Procedure" (in Appendix A) would be performed by the UE in order to re-activate the previously released tunnels.
   - Service Triggered by the Network: When the EPC network receives a downlink packet intended for a UE in Idle mode, a "Paging Procedure" should be performed by the MME. The paging procedure, illustrated in Appendix A, allows the network to locate the UE and identify to which eNB it should be connected to. Once the user is paged, the UE perform a "Service Request Procedure" toward the network in order to re-establish the tunnels previously released.

   In the connected mode, a UE would have at least a single active signalling path with the EPC. Mobility of connected mode users is then handled by a handover process (see Section 2.3.1).

In the currently deployed mobile architecture, network operators typically deploy a small number of SGWs and PGWs. This makes sense as long as the amount of data traffic carried over LTE is small. However, this shall not probably be true in the near future as the mobile data traffic is expected to grow at a compound annual growth rate of 53% from 2015 to 2020, reaching 30.6 exabytes (EB) per month in 2020 [3]. Bandwidth demand will increase at every segment of the network (access, metro and backbone). This traffic growth includes in particular the growth in traffic for video applications such as video streaming, video conferencing, etc., which is expected to reach 75% of the overall mobile data traffic, by 2020. In order to accommodate this growth, 3GPP considers a distributed LTE architecture as key for reducing the bandwidth demand on network segments. The basic idea is to distribute small radio base stations within the local residential network and mobile IP edges (SGWs and PGWs) within access and metro segments in order to implement mobile data offloading approaches.

2.2.2 Beyond LTE Architectures: Mobile Data Offloading

The present subsection reviews popular existing 3GPP offloading approaches relying on femtocells and/or macrocells. We first point out the benefits of deploying femtocells for telecom operators and associated consequences. Next, we present the LIPA and SIPTO approaches that have been introduced by 3GPP ([4], [45]). Finally, we outline the benefits of mobile data offloading for end-users and operators.

2.2.2.1 Radio Access Network Offloading

Femtocells, also called Home eNodeBs (HeNBs), are low power cellular base stations. They were first defined by 3GPP in order to offload the mobile traffic from the standard base stations (eNBs or macrocells) and thus improve the indoor voice and data coverage of mobile networks [46]. HeNBs transmit the mobile traffic using the same spectrum than macrocells [47]. Femtocells are connected to the EPC by the means of an IP home router, which uses a direct broadband connection (FTTx or xDSL). This connection replaces the backhaul infrastructure currently used for the macrocells.

However, due the fact that femtocells use the same spectrum as macrocells, the following problems are still open:

- Radio interferences;
- Femtocell location detection.

Moreover, access network offloading does not help in solving the issue raised in Chapter 1, namely mobile traffic routing through the EPC.

2.2.2.2 Mobile Aggregation and Core Network Offloading

3GPP has extended the use of femtocells in order to limit the loads of both aggregation and core networks. Indeed, femtocells potentially allow a direct access to the public IP network via the fixed network. Hence, by adding Local PDN gateway (LGW) which is either co-located with the femtocells or represented within a separate (or standalone) entity connected to the femtocell via a different interface, mobile data traffic will bypass the EPC and thus reduce the load of the standard gateways (SGW and PGW) used in 4G networks.

3GPP in [4] has also proposed alleviating the load of the standatd SGW and PGW by selecting new SGW and/or PGW located above the RAN i.e., beyond the macrocells, closer to the user's location. These approaches are explained in more details below.

1. **Local IP Access (LIPA):**

   LIPA has been presented by 3GPP [4] in order to allow a mobile user to directly access the private IP network services using a HeNB with either co-located or separated (standalone) LGW. The LGW is a gateway towards the external IP network. It supports PGW functions such as UE's IP address allocation and DHCP (or DHCPv6 for IPv6) functions. Moreover, the LGW supports some of

the SGW's functions such as downlink packet buffering as well as direct tunnelling towards the HeNB. However, the LGW is not a full SGW since the UE is already linked to a different SGW for its non-offloaded traffic, and 3GPP states that a UE can only be attached to a single SGW. The user's uplink packets are first routed towards the HeNB which first filters them on the basis of the destination IP address and tunnel ID, and then send them either to the LGW or to the SGW.

The LIPA architecture introduces new signalling and data paths using the existing protocols defined in LTE/SAE architecture [4]. An S5 interface is introduced between SGW and LGW and built over user and control plane with GTP-U and GTP-C protocols for tunnel management. Also, a direct link is built between HeNB and LGW with GTP-U protocol. Fig. 2.4-b illustrates the new paths standardised in 3GPP in [4] for the LIPA architecture. The introduction of new interfaces for LIPA does not affect the regular LTE/SAE signalling and data traffic routing.

2. **Selected IP Traffic Offload (SIPTO):**

   SIPTO was first defined within 3GPP SA2 group in [6] in order to breakout selected IP traffic (e.g., Internet) at, or above the RAN i.e., beyond the macrocell (eNB). In particular, SIPTO allows alleviating the loads on the mobile aggregation and core networks by selecting a SGW and a PGW that are topologically/geographically close to the UE.

   According to 3GPP in [4], a UE can only be connected to one SGW at a time. Therefore, at the establishment of SIPTO PDN connection, the MME selects its preferred SGW. SIPTO above RAN architecture relies on the same architecture model and concepts of LTE/SAE described in [4]. Consequently, the selected SGW and PGW might either be co-located in a single gateway or separated from each other within different equipments. 3GPP use the term "standalone" for the separated equipments use cases.

   SIPTO has been extended by 3GPP in [7] in order to support breaking out selected IP traffic within the residential IP network or at local network (at LN). SIPTO at LN architecture relies on the same architecture model and concepts of LIPA presented above. Hence, SIPTO at LN allows a direct connection between the UE and the local IP network using a femtocell HeNB with either co-located or standalone LGW [4], [48].

In mobile aggregation/core network offloading, LIPA and SIPTO both offer a direct connection between mobile users and the residential Local IP Network. In particular, a UE shall be able to have simultaneous access to the local IP network (using LIPA

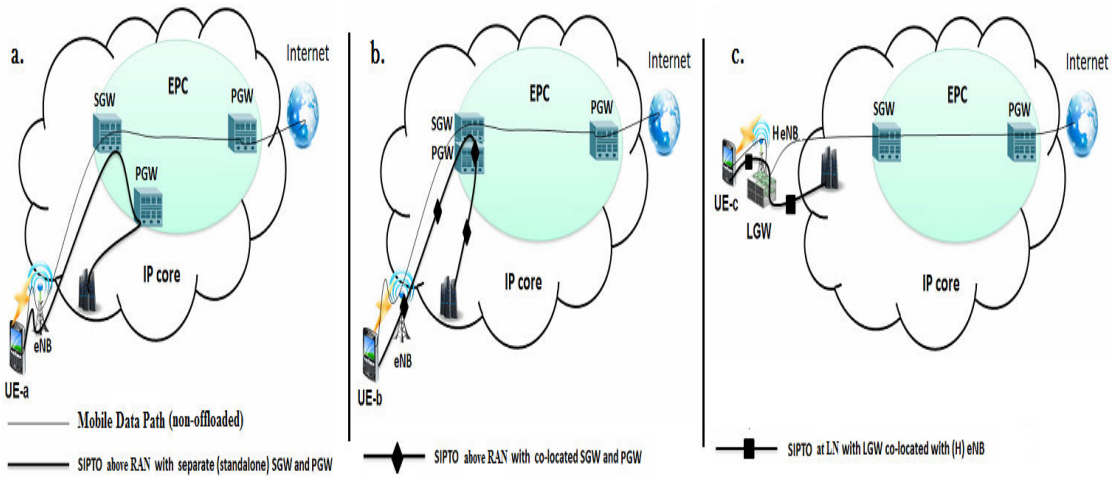


FIGURE 2.5: Mobile data Offloading Architecture

and/or SIPTO data paths) and to the operator's EPC network through the standard LTE path.

Figure 2.5 illustrates the different LIPA and SIPTO use cases where UEs (a, b and c) are having part of their IP traffic routed towards EPC, while another part is offloaded towards a VoD server, located within the IP network, and close to the user.

1. **SIPTO above RAN with standalone gateways:** UE-a (Figure 2.5-a) is traversing the SGW for all traffic, but relies on two different PGWs.
2. **SIPTO above RAN with co-located gateways:** UE-b (Figure 2.5-b) is traversing the SGW for all traffic. However, UE-b is accessing the server using co-located SGW and PGW.
3. **LIPA or SIPTO at local network:** UE-c (Figure 2.5-c) is traversing the SGW only for the non-offloaded traffic. Since the LGW includes functions of both PGW and SGW, this implies that UE-c is virtually connected to two SGWs simultaneously: the standard SGW, which is used by the non-offloaded traffic and the LGW, which is used by the traffic offloaded thanks to SIPTO.

Wi-Fi access points located in Home Gateways offer a method to offload traffic similar to LIPA or SIPTO at LN. The selection and filtering is not performed in the HeNB, but directly in the UE.

2.2.2.3 Benefits of Mobile Traffic Offloading

Besides the capacity, coverage and network performance's improvement, the deployment of femtocells helps operators reducing a significant part of capital expenditure (CAPEX)

as well as operational expense (OPEX). According to Cisco [49], "33% of global mobile data traffic (GMDT) was offloaded onto the fixed network using femtocells and Wi-Fi in 2012". By 2020, the offloaded mobile data traffic is expected to reach 55% of the GMDT [3], and thus, the use of femtocells and WiFi technologies to access the mobile core network will help saving 22% of CAPEX costs dedicated for macrocells deployment. Furthermore, considering the ongoing costs for running macro base stations, deploying and monitoring their backhaul as well as additional electricity, OPEX is reduced to only 200$ per year per femtocell down from 60.000$ per year per macrocell [50].

In Chapter 3, we investigate whether locating video servers close to the users, and relying on LIPA/SIPTO significantly limits bandwidth demand on the aggregation and core segments.

2.3 Mobility Support

The primary feature of all cellular networks is no doubt the possibility for mobility. In particular, mobile users shall be able to move anywhere within the coverage area of the network with no interruption of their ongoing data services. This section presents the basic 3GPP mobility procedures for users with offloaded and/or non-offloaded services. First, mobility in classical LTE architecture is defined. Next, mobility beyond in mobile data offloading architecture is discussed. We conclude this section with some related works.

2.3.1 Mobility Support in Classical LTE Architecture

Mobility management in classical LTE network mainly depends on the user's connectivity mode (Connected or Idle).

When a UE is in connected mode, at least one signalling path would be active between the UE and the EPC. Mobility of connected mode users can then be handled by the handover process defined by 3GPP in [4]. Typically, handovers are classified by the target system. In this thesis we focus on the "Intra-LTE handover process". This represents transitions to the same or different carrier frequency inside an LTE system.

2.3.1.1 Handover Types in LTE

3GPP defines two potential handover types within classical LTE architecture (see Figure 2.6).

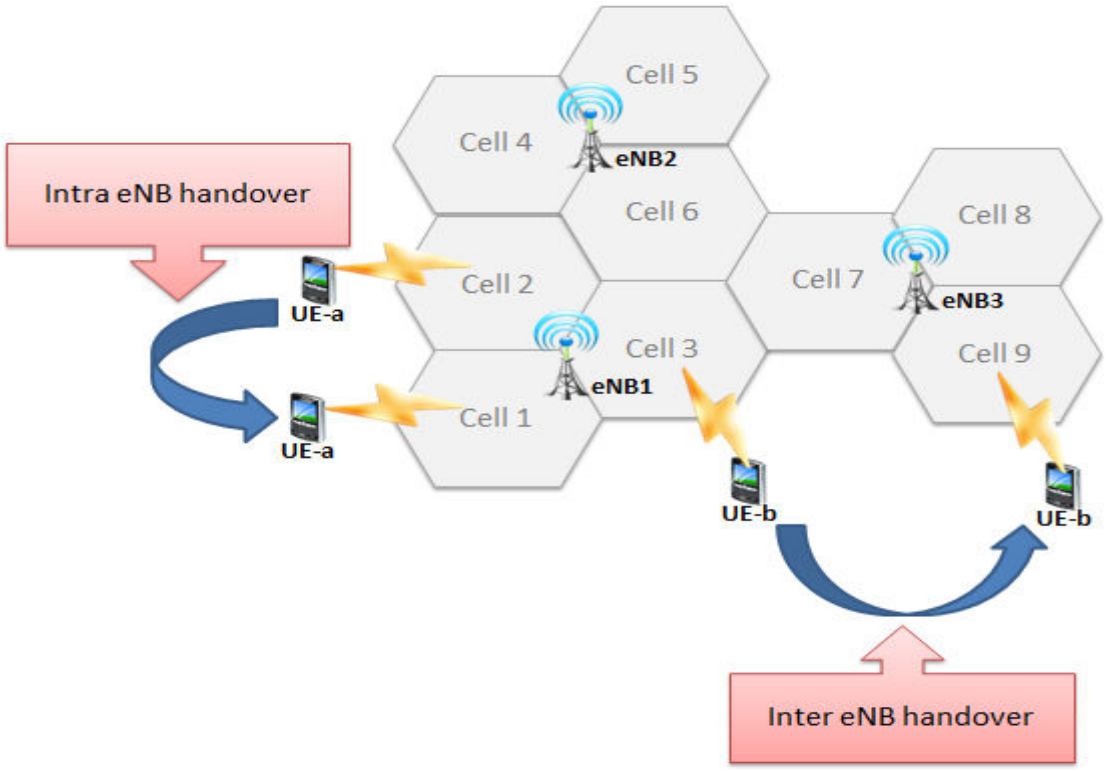


FIGURE 2.6: Inter eNB versus Intra eNB handover

- **Intra eNB handover:** in this case, the source and target cells reside in the same eNB. The handover performed inside the eNB and no traffic forwarding is performed.
- **Inter eNB handover:** This handover refers to a case where the source and target cells correspond to different eNBs. During an inter eNB handover, the MME may or may not be changed. Moreover, a SGW relocation could potentially be required. When the MME is not changed, either a "X2-based handover" or a "S1-based handover" procedures [4] is performed in order to transfer the user's data traffic from source to target eNB. However, if the MME is changed, only "S1-based handover" procedure can be applied.

As pointed out in Section 2.2.1, an idle mode UE has no active signalling connection to the EPC network. To stay reachable by the network during mobility, an idle mode UE has to choose a feasible radio network and a cell to "camp" on. This could be done by performing a "PLMN selection" and "cell selection and re-selection" functions. The UE then has to report its location or its TAI to the HSS via a "Tracking Area Update Procedure [4]" shown in Figure A.10 and Figure A.11 of Appendix A. The TAI is used to reconnect the user to the EPC if a service is available.

2.3.1.2 Handover Procedures in LTE

A LTE handover is a process that is reported by the UE (the UE detects that a better signal is delivered by a neighbouring cell), initiated by the eNB and controlled by the EPC. Handover consists in three phases: Preparation, Execution and Completion.

The **Preparation** phase establishes the forwarding data paths between the source eNB and the target eNB. First, the source eNB transfers the UE context to the target eNB in order to check whether the target eNB is capable of allocating required resources for this user or not. Then, depending on the target eNB's response, one of the following actions is performed:

- If no resources are available at the target eNB, the handover fails;
- If resources are available at the target eNB, the latter establishes the downlink forwarding tunnels with the source eNB.

During the **Execution** phase, data forwarding is performed. This phase starts when the source and target eNBs are both ready for the handover and the forwarding tunnels are established. The source eNB sends a Handover Command message the UE in order to initiate the handover. The UE then disconnects from its current cell within the source eNB and connects to the new cell in the target eNB. Downlink packets received by the source eNB shall now be forwarded to the target eNB, which buffers them until the UE is completely synchronised with the target eNB. The UE's uplink packets are not sent to the target eNB as soon as the UE is successfully connected to the target eNB.

Finally, during the **Completion** phase, the network switches the downlink data paths to use directly the new path that goes from the PGW, to the (target) SGW, then from the (target) SGW to the target eNB, to finally reach the UE. The forwarding tunnels as well as the user context at the source eNB are then released.

3GPP has introduced two handover procedures for an LTE inter eNB mobility use case:

- **X2-based Handover:** In LTE, eNBs are interconnected with an X2 interface. When a UE moves between two eNBs that are served by the same MME, handover from source to target eNB will be performed over the X2 interface. Let us assume the scenario in Figure 2.7-a where a UE is accessing the IP services through a PGW located within the EPC network. During mobility, the UE detects that a better signal is delivered by a neighbouring cell and forwards this information to the source eNB in a measurement report message (Figure 2.7-b). The source eNB identifies the target eNB using the Cell identifier information received from the

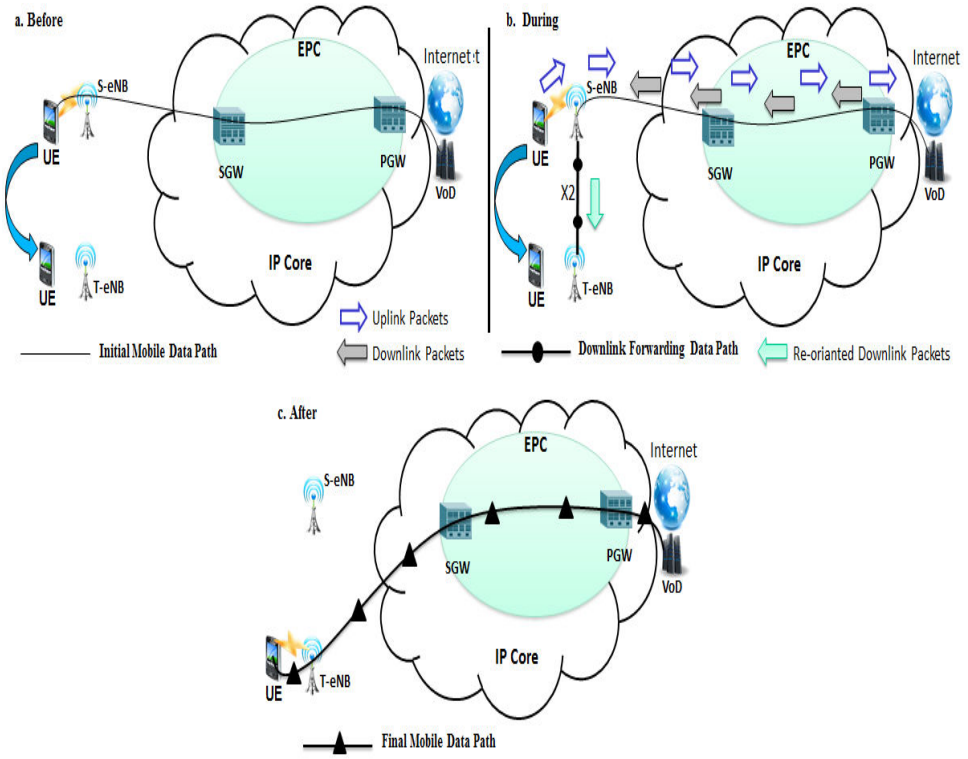


FIGURE 2.7: X2-based handover Procedure

UE. As both source and target eNBs are in the same MME pool, the handover preparation phase is performed, where X2 data path and X2 signalling connection are established over the X2 interface. At this point, the SGW is still not aware that a handover has occurred and sends downlink packets to the source eNB. The latter thus uses the X2 forwarding tunnel to forward the downlink data to target eNB. The handover execution phase then starts and the UE disconnects from the old cell and connects to the new cell. Once the UE is fully synchronised with the target cell, the user starts sending and receiving its data directly with the target eNB, the data path to the SGW is switched from source eNB to target eNB and finally the forwarding tunnel over the X2 interface is released. The final data path is shown in Figure 2.7-c.

- **S1-based Handover:** An S1-based handover procedure is performed either when the X2 interface between the source and target eNBs is unavailable or when the inter eNB handover requires an MME change. It relies on the S1-MME interfaces between the MME and both source and target eNBs. In particular, all signalling messages exchanged between source and target eNBs are first sent to the MME over the S1-MME interface and then forwarded to their ultimate destination.

  As previously, we assume that during mobility, the UE detects that a better signal is delivered by a neighbouring cell and forwards this information to the source eNB

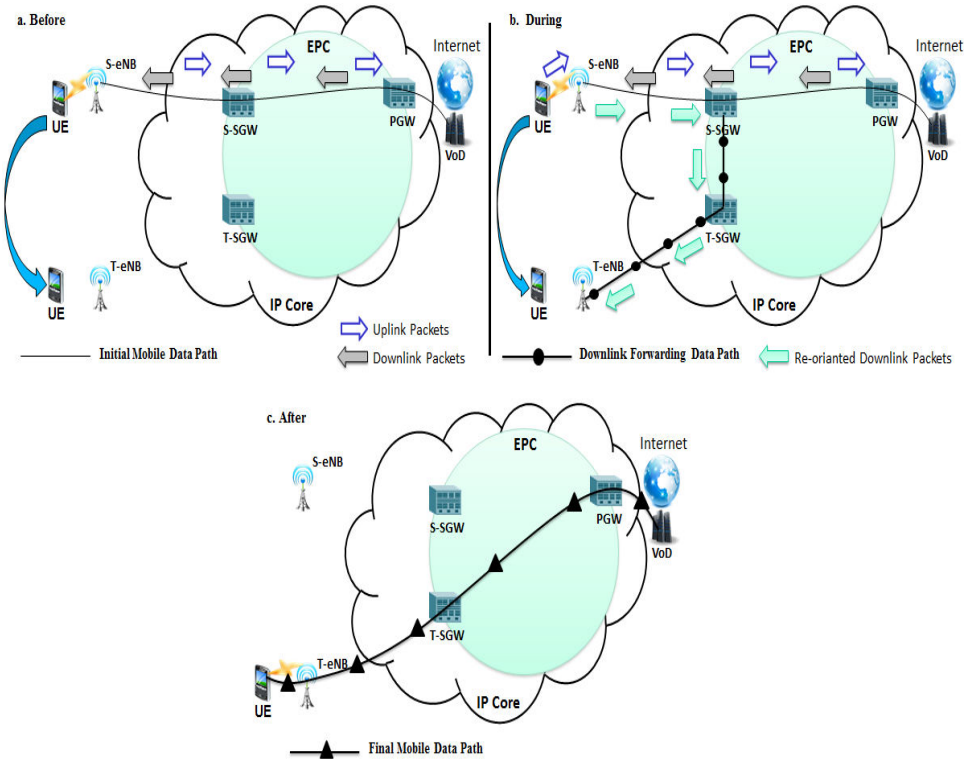


FIGURE 2.8: S1-based handover Procedure

in a measurement report message (Figure 2.8-a). Assuming that the source and target eNBs are connected to the same MME but belong to different SGW pools, the handover preparation takes place with a SGW relocation decision with no MME change. The target SGW is selected by the MME and an indirect forwarding tunnel is established between the source and target SGWs. Figure 2.8-b shows the indirect data forwarding tunnel connection used during the handover execution phase to forward downlink packets. During the execution phase, the PGW is unaware of the user's handover. The PGW then keep sending downlink traffic towards the source SGW. During handover completion, the radio access data path between the UE and the source eNB as well as the S1 data path between the source eNB and source SGW are switched to target SGW and target eNB. Finally, the resources previously established with the source eNB and source SGW as well as the resources dedicated for the indirect forwarding tunnel are released; the final situation is shown in Figure 2.8-c.

2.3.2 Mobility Support in Beyond LTE Architectures

In this section we present an overview of the different mobility scenarios in mobile data offloading architectures. We then identify the potential issues in each scenario, in which

most notably is session continuity support. Finally, we outline the different solutions proposed to maintain session continuity.

2.3.2.1 Mobility Support in LIPA

According to 3GPP [4], LIPA is only intended to allow UEs to access their own private Local Access Network via a femtocell. Thus, UEs may not apply LIPA when connected through a macrocell. Moreover, when a UE, which is having an ongoing LIPA session using a HeNB co-located with a LGW, moves towards a target HeNB, a LGW relocation would be required by the network and a deactivation procedure with type of reactivation would be performed by the MME (Appendix A). The relocation of the source LGW would result on loosing the IP address allocated to the UE by the latter. Consequently, no service continuity is supported in case of mobility of UEs with ongoing LIPA session (e.g. whenever a UE moves away from the femtocell, all his sessions offloaded with LIPA approach are terminated). Considering the initial scenario shown in Figure 2.5-c, we now illustrates in Figure 2.10 the mobility use case when UE-c moves from HeNB1 to HeNB2.

2.3.2.2 Mobility Support in SIPTO

Unlike LIPA, SIPTO allows alleviating the loads on the mobile aggregation and core networks regardless of the available radio access network (i.e., Macrocells, Picocells or Femtocells). Depending on the type and the location of gateways used to access the external IP network for SIPTO, 3GPP has distinguished three different Mobility use-Cases (MC).

1. **MC1: SIPTO above RAN with standalone SGW and PGW:**
   According to [7], service continuity for users with ongoing SIPTO above the RAN with standalone SGW and PGW session is supported using the existing mobility procedures defined in 3GPP specifications for LTE mobile architecture. This is due to the fact that even if a SGW relocation is required, the user's IP address allocated by the PGW remains the same during the whole handover procedure. This includes UE's mobility within Macro network, Femto network and between Macro and Femto cellular networks.

   However, if SGW and PGW are quite close to one another (e.g., located in the same building) there would be a potential relocation of PGW whenever SGW is relocated.

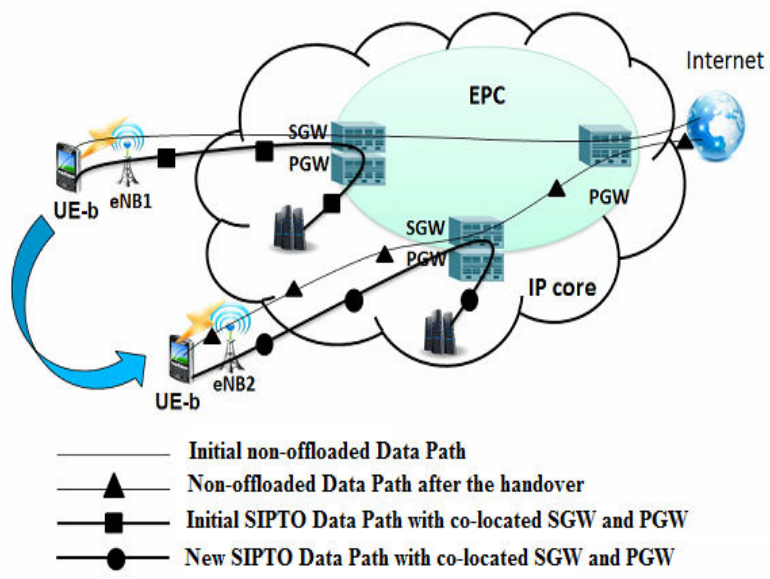


FIGURE 2.9: MC2: Mobility of UE having SIPTO above RAN Session with co-located SGW and PGW

2. **MC2: SIPTO above RAN with co-located SGW and PGW:**
   As result of UE mobility having an ongoing SIPTO above RAN session, a SGW relocation procedure might be provided by the MME. For SIPTO with co-located SGW and PGW, the gateway relocation decision would affect both SGW and PGW. Then, the relocation of SIPTO PGW will result in losing the IP address allocated for the UE by PGW. Consequently, the MME must disconnect the impacted SIPTO connection with reconnection cause required [4]. This procedure will probably be seamless to users having short-lived applications such as SMS-texting. However, the deactivation of a SIPTO connection will affect the sessions which require the conservation of the currently used IP address e.g., video streaming and online gaming. We illustrates this mobility use case in Figure 2.9 when UE-b shown previously in Figure 2.5-b moves from eNB1 to eNB2 and relocates from source co-located SGW and PGW towards new co-located SGW and PGW. As shown in this figure, local services are interrupted during all the UE's mobility procedure. Only, after the activation of the new SIPTO connection, the UE can request those services again.

3. **MC3: SIPTO with LGW co-located with HeNB:**
   Similarly to LIPA and SIPTO with co-located SGW and PGW, mobility of UEs with ongoing SIPTO sessions with HeNB co-located with LGW will affect the continuity of sessions requiring IP address maintain. Due to the user's mobility, the MME decides to perform LGW relocation. Consequently, the offloaded data traffic requiring IP address maintain is interrupted and a deactivation procedure with type of reactivation is performed by the MME on SIPTO at LN traffic during

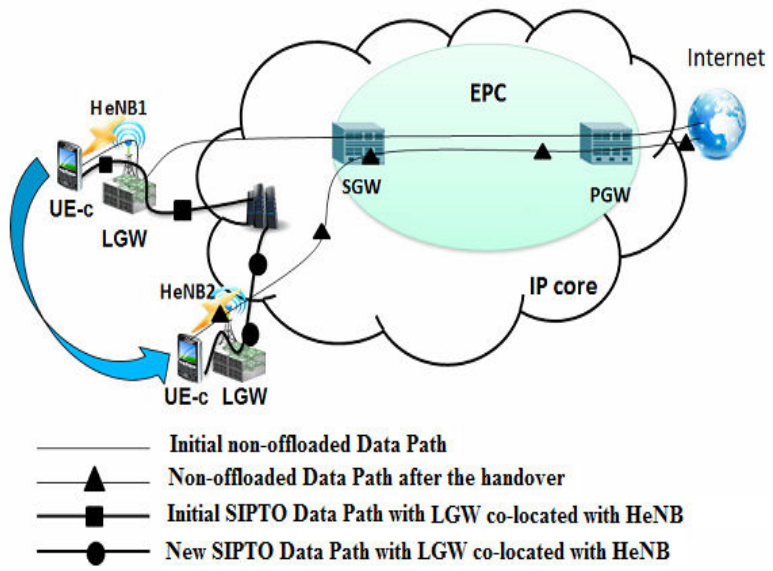


FIGURE 2.10: MC3: Mobility of a UE having SIPTO at LN /at RAN session with LGW co-located with (H)eNB

the change of LGW (Figure 2.10).

In the mobility use cases discussed above, we see that SIPTO mobility is not supported due to multiple data paths to a single UE which has multiple IP addresses. Multiple data paths issue can be solved with the help of multihoming protocols provided by IETF, which is discussed in the following subsection.

While LIPA approach is only applicable on femtocells, SIPTO approach on the other hand, provides its users with access to both local services using SIPTO at LN (similar to LIPA) as well as external IP services using SIPTO above RAN. This is why, in this thesis we focused our study on providing seamless mobility for mobile users that are having ongoing SIPTO sessions only.

2.3.2.3 Related Works on Mobility Support

As shown in the above Section, session continuity is not maintained during mobility with PGW/LGW relocation cases. Most of the current studies focusing on offloading mobile data traffic with seamless mobility aim at IP address conservation as the goal to reach for ensuring SIPTO session continuity in case of mobility.

3GPP has considered in [45], three different use cases under the "Change for SIPTO" (C-SIPTO) approach. C-SIPTO first use case relies on performing "deactivation procedure with reactivation" cause only for short-lived traffic, while continuing to forward the

ongoing long-lived traffic towards the initially chosen PGW. After the attachment to the target PGW is finalized, a new SIPTO connection is established with the target PGW to be used for future long-lived sessions. C-SIPTO second and third use cases have introduced the "always-on dual connection" and "on-demand dual connection" concepts. These concepts consist in redirecting all short-lived sessions towards a SIPTO data path using a selected LGW while continuing to forward all traffic flows that require a stable IP address towards the initially chosen PGW in the EPC network. As a result, for all three uses cases, C-SIPTO supports smooth handover in case of mobility only for traffic flows corresponding to ongoing short-lived sessions. However, none of C-SIPTO cases currently allow a smooth handover in case of mobility with gateway relocation for traffic flows corresponding to ongoing sessions that require a stable IP address.

A distributed mobility management solution was proposed in [51], where a PDN Edge Gateway (P-EGW) was introduced to support scalable LTE/SAE networks. In this solution, several P-EGWs are distributed close to the user's location, whereas each P-EGW supports both SGW and PGW functionalities defined in [4] and presented in Section 2.2.1 for classical LTE/SAE network architecture. In order to differentiate between both functionalities, two types of P-EGWs were considered: Anchor P-EGWs and Access P-EGWs. Indeed, the Anchor P-EGW is the default gateway to be used when a UE attaches to the network, whereas the Access P-EGW is selected whenever a UE that is having an ongoing session reaches an area that is not covered by the initial (Anchor) P-EGW. User's traffic is first routed from the target eNB to the Access P-EGW, and then from the Access P-EGW to the Anchor P-EGW in order to reach the external IP network. The selected Access P-EGW could be used as an Anchor PGW only for new requested PDN connections. This solution is very similar to the first C-SIPTO use case for long-lived sessions. However, in this solution in addition to the P-EGWs, new types of distributed MMEs have also been introduced to the network's architecture in order to dynamically manage the location binding information of ongoing sessions during user's mobility. Such new framework will increase the complexity of the network architecture and will not allow an optimized data path during the user's mobility as the user keep being attached to the initial Anchor P-EGW during the lifetime of its ongoing session.

Taleb T. et al. have introduced the Follow Me Cloud framework [52] for seamless users mobility using interworking federated clouds with distributed mobile networks. Follow Me Cloud enables mobile users to access cloud services using always the most optimal path by migrating services to the nearest available Data Center (DC) and/or data anchor gateway with no session interruption. However, to allow a smooth mobility of users in this solution, authors propose to replace data anchoring at the network layer by service anchoring and IP addressing by service/data identification. Moreover, a Follow

Me Cloud controller and a DC/GW mapping entity have also been introduced to the network architecture in order to allow an optimum session migration from one location to another during users mobility. Consequently, even though Follow Me Cloud has allowed a best data path selection while ensuring a smooth session migration when a user changes its network point of attachment, it has done so by significantly modifying and adding complexity to the current 3GPP architecture.

A lightweight Mobile Cloud Offloading Architecture (MOCA) has also been introduced in [53]; MOCA relies on cloud infrastructure and Software-Defined Networking capabilities to offload part of users' IP traffic. The basic idea of MOCA is to introduce a cloud platform, inside the mobile network, in which operators can instantiate a software instance of SGW, PGW and Content Server engine. Even though MOCA has enabled the adoption of Software-defined networking (SDN) and cloud technologies in a new mobile data offloading architecture, MOCA realization has included extensions to signalling protocols and modifications to the MME and the new cloud based SGW. Additionally, the use of an SDN middle-box for packet interception and the additional forwarding rules introduced into the eNBs and SGWs, potentially increase latency and in any case adds complexity to the standard mobile architecture. Lastly, session continuity during mobility between core EPCs was initially not considered in MOCA.

2.4 Fixed and Mobile Convergence

Fixed and Mobile Convergence (FMC) represent one of the most promising trends taking place in the telecommunications industry. The objective behind the FMC network is to allow a seamless switching between fixed and mobile devices/networks, e.g., one can consider handsets that switch between cellular and WiFi technologies or a user watching a movie on the big-screen TV and continue it on its smartphone or tablet and vice versa.

Today, fixed and mobile convergence is only implemented at service level with all IP services and IMS [8] [15]. Indeed, FMC was first considered by Telecom operators to unify streaming media, multimedia, legacy public switched telephone networks (PSTNs), IMS, and Web services onto a common network infrastructure. The resulting network environment has offered rapid and inexpensive application enrolment across a common IP infrastructure. However, FMC approaches can be used to explore new innovative solutions as it can present several benefits for both operators and customers. FMC network architecture allows reducing the cost for end users and operators by sharing fixed and mobile access/aggregation network infrastructures and hardware (equipment, links); e.g., as pointed out in Section 2.2.2.1, the use of femtocells and WiFi technologies (whenever is possible) to offload mobile data traffic would help saving up to 22%

of the CAPEX costs used for macrocells deployment. The main benefits behind the implementation of FMC network are :

- Minimization of the number of network elements and their locations,
- Optimization of the number of caches and the cache locations,
- Simplification of network control and route management,
- Improved service delivery due to the availability of different paths over the different access technologies,
- Improved network performance (higher throughput, reduced latency, higher services quality e.g., 3D video services or video streaming/conferencing with Hight Definition (HD) quality),
- Lower complexity by simplifying the network structure,
- Reduced cost and energy consumption by sharing control functions and improving network structure,
- Improved usage of available network resources,
- Unifying authentication for users regardless of the access network's type,
- Unifying IP edges of all networks (fixed and mobile gateways) within a single entity, and thus allowing flexible use of common functions,
- Enabling efficient network level load balancing schemes,
- Potential separation between control and data plane by using SDN technology.
- Enhancing mobile segment of FMC network architecture, which will be able to support enormous traffic growth generated by future HD services.

At a short-term basis, operators are focusing on the integration of fixed and mobile networks in the aim of making them converged in the perspective of 5G network [54]. In particular, the Next Generation Mobile Networks (NGMN) Alliance identified such integration as recommended option for 5G design [55].

At a long term transition, operators aim to allow the convergence of fixed and mobile networks themselves, combining both an optimal and seamless quality of experience for the end user together with an optimized network infrastructure ensuring increased performance, reduced cost and minimized energy consumption. Recent publications [9] [10] propose distributing multiple IP edge nodes closer to the users. In particular,

[9] proposes to implement an integrated IP edge for fixed, mobile, and Wi-Fi access networks within a functional entity called Universal Access Gateway (UAG) as a major building block of a converged fixed and mobile network. The co-location of such a UAG with application servers and data centers within a so-called Next Generation Point of Presence (NG-PoP) would allow a more efficient control of network resources [10]. NG-POP represents a location in the network where all essential functions, equipment and infrastructures of convergent networks would be distributed. It could typically be located either at the Core Central Office (CO) (the current IP edge for fixed and Wi-Fi traffic) or at the Main CO, thus moving the IP edge closer to the users, even for fixed and Wi-Fi traffic. This would improve the deployment flexibility of both fixed and mobile architectures and would also reduce the load of both data and control functions at the Core segment of the network. Moreover, this solution is expected to facilitate the implementation of Mobile Edge Computing (MEC) and fog architectures.

The overall NG-POP concept is illustrated in Figure 2.11 below.

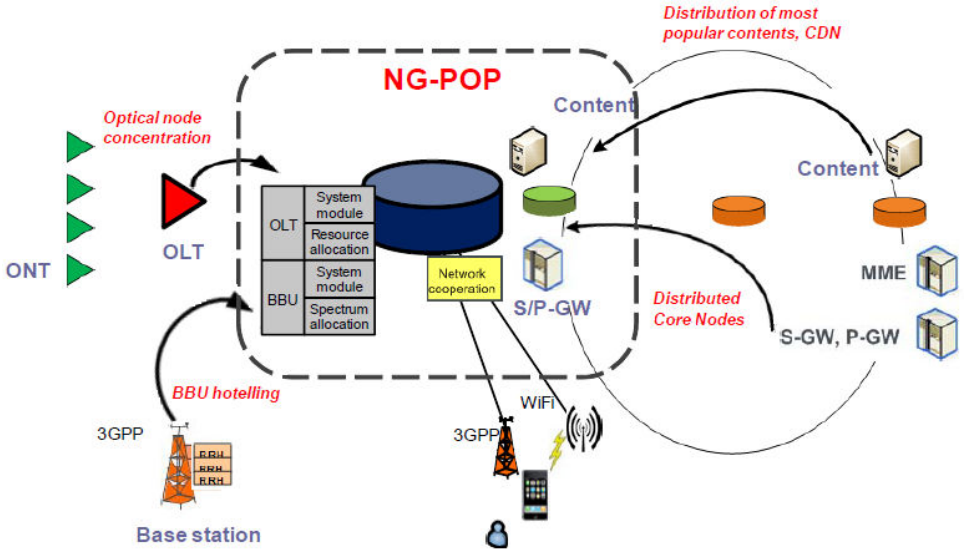


FIGURE 2.11: Overall Architecture for the Next Generation - Point of Presence [15]

The ultimate NG-POP-based COMBO architectures specifically targets two different aspects of carrier-based network convergence, which are both keys for a global FMC network [12, 40]:

1. **The Functional Convergence:** in which key functionalities of fixed and mobile networks should be implemented in a way to ensure a better use of network resources/interfaces by collaborating various access technologies and unifying control plane functions. It will also impact the data plane through an optimization of protocol stack and a better distribution of data flows in the converged network.

2. **The Structural Convergence:** defined as pooling/sharing of fixed and mobile access/aggregation network infrastructures and hardwares (cable plants, cabinets, sites, equipment, buildings) with expected CAPEX and OPEX cost saving for a converged network operator.

The work achieved in this thesis is part of the functional convergence of COMBO Project. COMBO in particular targets a better distribution and localization of EPC advanced functions such as SGWs and PGWs to offload the EPC network. A detailed overview of COMBO project is presented in Section 5.1 of Chapter 5.

2.5 Multiple IP Addresses Management Approaches for Mobility Support

Currently, mobile devices are emerging with multiple interfaces with diverse access technologies. This means that multiple IP addresses would be assigned to each user. As pointed out in Section 2.2 different solutions have been addressed in order to dynamically control the delivery of data originally targeted for cellular network bypassing the metropolitan and the backbone networks. Some of these solutions allow the delivery of mobile data traffic using the fixed access network i.e., through the home gateway, while other solutions were based on the use of wireless (Wi-Fi and cellular) accesses.

The use of multiple interfaces to a single user could probably improve the user's throughput by distributing data traffic over different available data paths. It also allows a failover from one interface to another in case the connection breaks on the first. However, current Transmission Control Protocol (TCP) [56] which is used for 95% of Internet communications does not support a simultaneous use of multiple interfaces for a single path transport. To overcome this limitation, IETF has proposed a Multi-Path Transmission Control Protocol (MPTCP) [18] [57] and a Concurrent Multipath Transfer-Stream Control Transmission Protocol (CMT-SCTP) [58] [59], dedicated for managing multiple IP addresses in a single host.

1. **Multi-Path Transmission Control Protocol (MPTCP):**
   Multi-Path TCP or MPTCP represents a set of extensions to the regular TCP enabling a single data connection to use several IP-addresses/interfaces simultaneously while in fact spreading data across several sub-flows. Indeed, MPTCP provides a multi-homing service for users with a single TCP session. It offers a better resource utilization, higher throughput by using several interfaces simultaneously for data delivery and smoother reaction to failures by allowing fail-over between

interfaces. Mainly, it is designed to be compatible with TCP (e.g., for non-MPTCP aware applications, MPTCP behaves exactly as a normal TCP). MPTCP is also compatible to existing application and network as it uses the standard socket API used by most Internet applications.

MPTCP connection establishment relies on the same signalling model "three ways handshake: SYN, SYN/ACK and ACK" used for a simple TCP connection establishment. However, IETF enhanced the TCP handshake with an MP_CAPABLE option carried over each signalling message and included in the TCP packet header as discussed in [18]. The MPTCP-CAPABLE option allows the host requesting the MPTCP connection establishment to identify whether the receiving host supports MPTCP or not. Each MPTCP connection is uniquely identified by the network thanks to the cryptographic tokens provided by both hosts and included in the MPTCP handshake.

MPTCP identifies multiple paths by the presence of multiple addresses at hosts. An MPTCP host can add/remove any of its IP addresses to/from the remote host at any time thanks to MPTCP's "ADD_ADDR" and "REMOVE_ADDR" options. When a host shares its list of addresses with a remote host, additional sub-flows would be added to this connection using MP_JOIN option. Similar to MP_CAPABLE option, MP_JOIN option is included within a three ways handshake TCP packet headers.

MPTCP provides the multiple TCP sub-flows with priority feature. Two priority options have been considered by IETF: Backup and Regular. Thanks to the MP_PRIO option, an MPTCP host (e.g., mobile user) can indicate to the remote host (e.g., content server) whether a new established TCP sub-flow should be used as "Regular" or "Backup" path. Regular sub-flows are the active data paths on which data could be transmitted. Whereas the sub-flow which is defined as backup path will be used only when all the regular sub-flows are not working.

Based on the two priority options presented above, Paaschy C. et al. have introduced three different handover modes for an MPTCP implementation: full MPTCP mode, backup mode and single-path mode [60]. The full MPTCP mode refers to the regular MPTCP operations, in which all TCP sub-flows are used simultaneously between the client and server hosts. This mode is mostly used for users targeting a best data transfer rates. In backup mode, MPTCP creates sub-flows over all the existing interfaces but data packets only flows on a subset of them. Finally, in the single-path mode only one TCP sub-flow is used at any time. Moreover, sub-flows are not pre-established over the backup interface (i.e., another TCP sub-flow can only be created if the ongoing TCP sub-flow is broke).

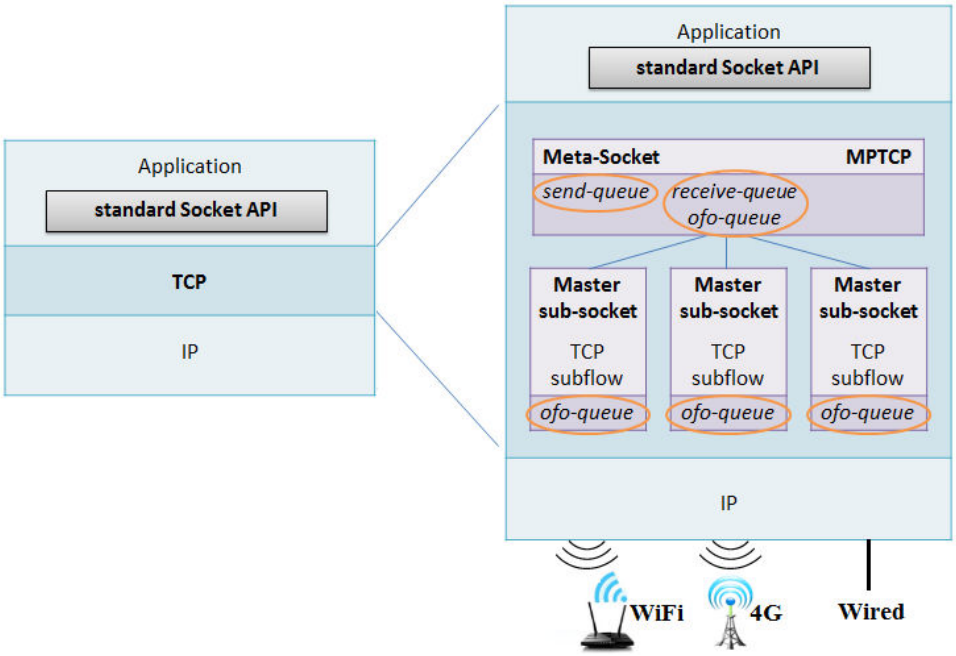


FIGURE 2.12: Comparison of Standard TCP and MPTCP Protocol Stacks

On the Linux implementation [61], when an application creates a socket and the system requests the establishment of new connection with another machine, the system initiates the MPTCP connection and creates the first TCP sub-flow called "Master Sub-socket" is created. The application thus begins by operating as a normal TCP socket. The Master Sub-socket then verifies if the remote host supports MPTCP and whether extra paths are available. In that case, additional TCP sub-flows called "Slave Sub-sockets" will be created. In order to handle the MPTCP connection, Linux Kernal has enhanced the Transport Layer with a "Meta Socket MPTCP" interface. The meta socket is the only data structure that is directly linked to the socket that is visible by the application as presented on Figure 2.12. Each meta-socket is provided with a send, receive and Out of Order (OfO) queues used to manage the traffic distribution of the different TCP sub-flows using an MPTCP scheduler. The OfO queue ensures the retransmission of lost packets and allows re-ordering the received data. The OfO queue is also implemented at each TCP sub-flows for packets re-ordering.

MPTCP provides a full mesh of possible network paths among the available interfaces. However, it serves only MPTCP compliant clients. For non-MPTCP compliant clients, IETF proposed the use of proxy MPTCP [62]. Proxy MPTCP represents the intermediate element between an MPTCP and a non-MPTCP compliant hosts. Its main role is to allow the MPTCP compliant host (e.g., mobile user) to initiate an MPTCP connection with the non-MPTCP compliant host. It is transparent to the MPTCP compliant host and all the TCP applications on

both hosts. For instant, when a mobile user that is MPTCP-CAPABLE initiates an MPTCP connection using a SYN packet, the proxy MPTCP will intercept the packet and create a temporary entry consisting of IP addresses and port numbers of both hosts (e.g., a user and a server) for the required MPTCP connection. Then, it will forward this SYN packet to the server. If the latter replies with MP-CAPABLE option in SYN+ACK packet (i.e., the server is also MPTCP compliant host), then the proxy MPTCP removes the temporary entry for this connection. Otherwise, the proxy MPTCP initiates an MPTCP connection with the user, as if it is the server. In that case, the temporary entry is maintained by the proxy MPTCP to record all the sub-flows.

2. **Concurrent Multipath Transfer-Stream Control Transmission Protocol (CMT-SCTP):**

CMT-SCTP is based on SCTP defined in [63]. Unlike TCP, SCTP already provides multi-homing capabilities which could be directly used for CMT-SCTP. An SCTP packet consists of an SCTP header and a set of "Chunks". The latter represents multiple information elements like control signalling (Control Chunks) or even user data (DATA Chunk). SCTP connection is initiated by a four-way handshake starting with an INITIATION (INIT) chunk. This message includes the list of all IP addresses of the host initiating the SCTP connection. when a remote host receives the INIT chunk it replies with with an INIT-ACK chunk. The INIT-ACK chunk acknowledges the reception of the INIT chunk and includes a list of all the IP addresses available on the remote host. Now each host is aware of the list of IP addressed of the other host.

During an SCTP connection establishment, each of the communicating hosts chooses one of its available IP addresses to be used as its "Primary" IP address, denoted as "IP1@MH" and "IP1@RH" in Figure 2.13. Using the primary IP addresses, a first path between the two hosts will be created. This first path is considered as "Primary Path". In standard SCTP, even when multiple paths are available on different interfaces, the Primary Path is the only path used for user data transmission. The other data paths would only be used to provide robustness (redundancy) in case of network failures. SCTP then does not allow load sharing parity on multiple available data paths. To overcome this limitation, IETF proposed the CMT-SCTP approach in [58]. Mainly, CMT-SCTP uses all additional IP addresses to create additional paths. CMT-SCTP provides multiple "disjoint" paths, as each secondary IP address can only be used for a single additional path creation.

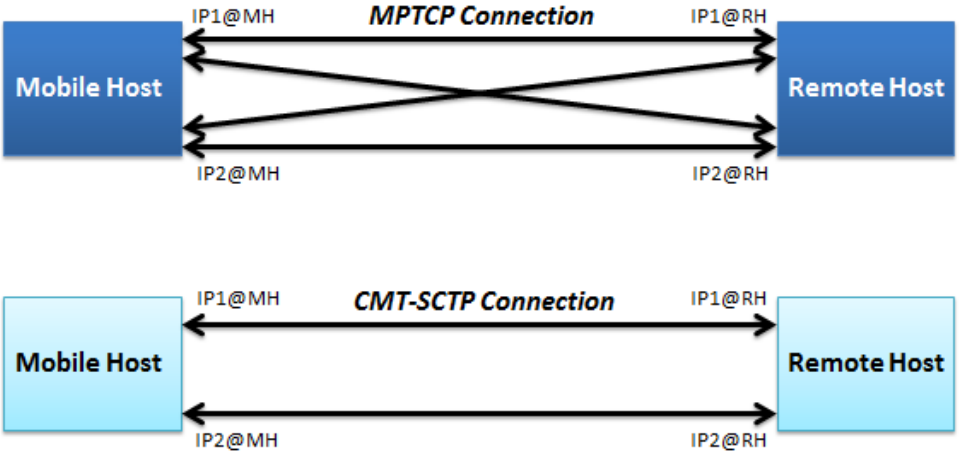


FIGURE 2.13: Data Paths over MPTCP and CMT-SCTP Architectures

Figure 2.13 shows a comparative architecture between MPTCP and CMT-SCTP paths management. The mobile Host (MH) can represent user equipment, while the Remote Host (RH) represents any peer node (e.g., content server). As shown in this figure MPTCP capable nodes are having a full meshed architecture using all available IP addresses, whereas CMT-SCTP nodes are having a point to point data traffic transmission.

As a result, both MPTCP and CMT-SCTP approaches support load sharing for end-to-end transport. However, while the MPTCP creates a full mesh of possible network paths among the available addresses, CMT-SCTP only uses pairs of addresses to set up communication paths [64]. Moreover, unlike MPTCP, CMT-SCTP is not transparent to the applications. Therefore, the solutions proposed in this thesis focus on the application of MPTCP to maintain session continuity.

2.6 Content Distribution Network Architecture

With an exponential growth, IP traffic is expected to become, within the next few years, dominant in both fixed and mobile networks. For instance, according to Cisco in [3], global IP traffic will reach 194.4 ExaBytes (EB) per month in 2020 up from 72,5 EB per month in 2015. As a result of this tremendous pace of expansion, network operators have started increasing their network capacity by deploying a distributed content architecture or the so-called Content Delivery Network (CDN).

Content distribution is a service of copying data strategically to a set of servers that are located closer to the end-users, typically in every large or medium town. These servers are known by the name of "Edge Servers". The deployment of such services would help

improving user's experience and network performance on the one hand and reducing both CAPEX and OPEX costs [5] on the other.

Content replication aims to alleviate the load of the origin content servers by offering an optimum content delivery. Indeed, for a best content delivery, a special "request-routing mechanism" has to be used to redirect the users requests to the appropriate edge server. A request-routing mechanism selects a target edge server based on a set of metrics such as distance, edge server load, response latency, network proximity and packet loss. The main routing mechanisms for content delivery are:

- Global Server Load Balancing (GSLB): In a distributed content architecture, a set of Web servers are attached to a Web switch. The latter should be aware of the health and performance of the Web servers attached to it. Using these information, a Web switch responds to the user's DNS request that is content-delivery enabled, with the IP address of the edge server most likely to give the best performance [65].
- DNS-based routing: In this approach, it is up to te DNS server to reply to the user's DNS request that is handled by a CDN. The DNS server chooses an IP address for the edge server that is geographically closest to the user. The selection of the edge server is based on the DNS resolver's IP address to which the user is routed to during its DNS request [66].
- HTTP Redirection: Typically, the information about replica server sets for CDNs are carried in the HTTP headers. Following an Hyper Text Transfer Protocol (HTTP) request, a Web server could use the HTTP protocols to reply to the user's request with an order to re-submit its request to another edge server [67].
- URL rewriting: In this approach, users are redirected dynamically by the origin server to surrogate servers that are close to the client by rewriting the generated pages URL links. [68].

The selection of content to be replicated could be based either on sensitivity to QoS [69] or on content popularity [70].

Generally, servers can be located either at the edge of the operator's network (close to the Internet) or within the operator's network itself [71–73]. In the latter case, two potential scenarios can be considered:

- The CDN belongs to the operator (e.g., VoD services),
- An interconnection agreement exists between the operator's CDN and the ISP's CDN (e.g. agreement for business services between Akamai and Orange).

2.7 Conclusion of Chapter 2

In this Chapter, I reviewed current fixed and mobile network architectures and identified that fixed and mobile convergence is the next step, especially in the context of content distribution. I also reviewed mobility support in mobile networks and showed that in some cases session continuity is not supported. Lastly, I reviewed several existing solutions to support multiple interfaces.

Chapter 3

Quantification of Bandwidth Gain via Mobile Data Offloading

In the present chapter, we assume that SIPTO "above the RAN" is implemented in order to allow content distribution to be performed non only from centrally located servers within the core network, but also from servers that are distributed closer to the users either in the metro network or at its edge. This could be the case for a network operated CDN, as described in Section 2.6. As shown in Section 2.2.2.2, SIPTO allows to selectively offload part of the mobile traffic, e.g. video traffic demands.

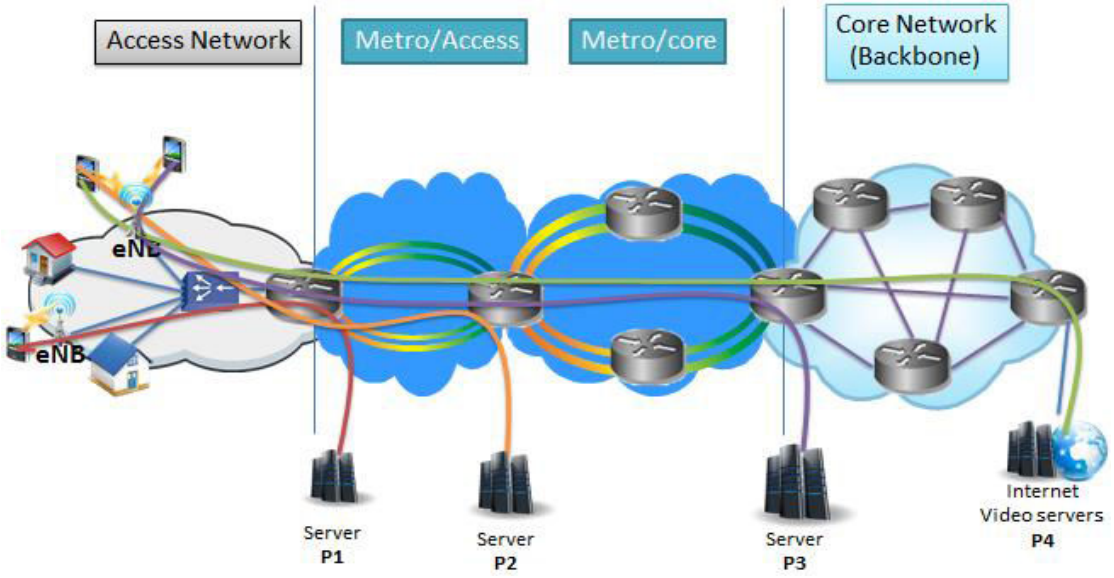


FIGURE 3.1: Locating servers in the metro and core networks

Figure 3.1 illustrates potential locations where servers for traffic offloading could be deployed.

We do not address here potential enhancements to LIPA and/or SIPTO mechanisms, which could solve the issue of session continuity in case of mobility (this is done in the next chapter). Assuming that session continuity is not an issue, we evaluate the potential gain in terms of bandwidth, of implementing SIPTO "above the RAN" and thus serving part or all content demands from servers closer to users (i.e. located in P1, P2, P3) instead of serving them from P4 as is currently done in actual mobile network implementations.

3.1 Considered Traffic Evolution

In future years, mobile data traffic is expected to grow faster than fixed data traffic, but to still remain significantly smaller in terms of volume. Indeed, mobile data traffic is expected to double each year during the next five years whereas during the same period, fixed data traffic growth should be only 20%. Table 3.1 shows the Cisco's forecast for the global IP traffic growth (for both consumer/residential and enterprise markets) between 2015 and 2020. In particular, the predicted Compounded Annual Growth Rate (CAGR) is introduced for the different traffic categories.

TABLE 3.1: Global IP Traffic, 2015 –2020. Source Cisco VNI 2016 [3]

	2015	**2020**	**CAGR 2015-2020**
By Type (EB per Month)			
Total IP Traffic	72.521	194.374	22%
Total Fixed Traffic	68.836	163.810	21%
Total Mobile Traffic	3.685	30.564	53%

Table 3.2, also from CISCO, gives more details on consumer Internet traffic and how it is distributed.

Considering the values in Table 3.1 and Table 3.2, we note that by 2020 Consumer Internet traffic will represent more than 80% of Total IP traffic with over 162 EB per month.

Also according to Table 3.2, Internet video traffic will reach 109,9 EB per month to become the most prevalent traffic of all consumer Internet traffic. Considering the Internet video traffic growth network by network, we see that mobile Internet video traffic will grow also faster than fixed Internet video traffic. Indeed, mobile Internet video traffic will grow at a CAGR of 62% between 2015 and 2020 compared to 27% for fixed Internet video traffic. However, during this period, mobile Internet video traffic will still remain significantly lower than fixed Internet video traffic in absolute terms, growing from 6% in 2015 to approximately 22% in 2020.

TABLE 3.2: Global Consumer Internet traffic, 2015 –2020. Source Cisco VNI 2016 [3]

	2015	**2020**	**CAGR 2015-2020**
Total (EB per Month)			
Consumer Internet Traffic	42.372	133.455	26%
Fixed	39.345	107.375	22%
Mobile	3.027	26.080	53%
By Sub-Segment (EB per Month)			
Internet video	28.768	109.907	31%
Fixed	27.011	90,239	27%
Mobile	1.756	19.668	62%
Web, email, and data	7.558	17.006	18%
Fixed	6.310	10.629	11%
Mobile	1.248	6.377	39%
File sharing	5.965	5.974	0%
Fixed	5.942	5.939	0%
Mobile	0.22	0.35	9%
Online gaming	0.82	0.568	47%

3.2 Distributing Video Services

Considering the traffic growth values shown in Table 3.1, it is essential to avoid tromboning and congestion, more particularly at the core network level. If services were delivered from servers close to the user's location this would reduce the load of the core network as presented in Section 2.6 of Chapter 2. Although, video is expected to be dominant in future traffic mixes, many other services can also benefit from distributed servers. This is true in particular for all cloud-based services such as e.g. on-line gaming and business services. However, in this Chapter, we focus on Video service distribution such as VoD and catch-up TV.

Considering the continuous growth of Internet video traffic, CDNs have prevailed as a dominant method to deliver such content. Actual deployments are based on distributing DCs and VoD servers close to the end-users locations. Globally, it is expected that 73% of all Internet video traffic will cross CDNs by 2020, up from 61% in 2015 [3]. CDNs can be operated either by ISPs, within their own network, or by "over-the-top" (OTT) providers, distributing Video through the Internet. Indeed, there are currently two different video distribution architectures: "IPTV" and "Internet Video".

1. IPTV:
   In the first one, the ISP also distributes video; this service is known as "IPTV". As the content provider controls the Digital Rights Management(DRM), and the access of the users to the various content, it is aware of all demands and can

dynamically select the server, which shall serve each demand. In particular, IPTV operators distribute servers within the metro network in order to improve user's experience and network performance. There are a few main video servers (large servers) that contain the complete set of offered video content, and several smaller, secondary servers, operating often as caches, which only store the more popular content. A recent publication [71] shows that the file hit ratio in such secondary servers can be relatively small (less than 35%) while the byte hit ratio in those servers is quite large (75%), as popular content are requested very often compared to the mean request rate.

Cisco VNI states in [3] that IPTV has increased 50% in 2015 and will continue to grow at a rapid pace, reaching 3.6 fold by 2020. Furthermore, IPTV traffic will represent 26% of consumer Internet video traffic by 2020, up from 24% in 2015.

2. Internet Video:
   In "Internet Video", the Content Provider (CtP) distributes the content over the Internet. Content distribution over the Internet relies on an overlay network linking multiple servers located at peering points. Popular content are present at multiple locations, while less popular are pulled on demand from central video servers. The Content Provider may directly distribute its own content (e.g. Google directly distributes YouTube content) or delegate content distribution to CDNs such as Akamai and Limelight. CDN servers are located close to the peering points with the ISP and are controlled by the CtP or the CDN. The ISP has a role neither in controlling the DRMs, nor in controlling the access to the content. As the ISP usually does not accept another provider to control servers within its own domain, there are no video caches in the metro network and the traffic cannot be offloaded. However, the video distribution ecosystem is constantly evolving, and strategic partnerships such as the one signed in 2012 between Akamai and Orange [72] may lead in the future to locating video caches used by Internet video services within ISPs' domains.

   Consumer Internet VoF traffic will grow at a slower pace than IPTV, to nearly double by 2020. In particular, Ultra-High-Definition (UHD) traffic will represent 20.7% of total IP VoD traffic in 2020, up from 1.6% in 2015.

3.3 Evaluation of the Amount of Offloaded Mobile Traffic

In this section, we present an evaluation of the amount of mobile data traffic that could be offloaded from the core and metro networks using distributed services within beyond LTE "Advanced-4G and 5G" mobile network architectures. The evaluation presented in

the following subsections is based on the study provided by Bell-Labs in [17] for analysing the impact of IP traffic growth on metro and backbone networks between 2012 and 2017. In this study, it was argued that by 2017, 75% of total metro traffic (totally fixed traffic) would be terminated within the metro network up from 43% in 2012.

We rely here on similar values to evaluate the amount of mobile traffic that could also be terminated within the metro network. We thus assume that the amount of total metro traffic that can be terminated within the metro network could reach 80% by 2020.

Two distribution models are considered in [17]:

1. Metro Centralized: In this model, servers are located at the edge of the backbone network (Figure 3.1, servers P3). This corresponds to the "Centralised COMBO" architecture considered in Section 5.1.3.1 of Chapter 5.
2. Metro Distributed: In this model, servers are also distributed within the metro network at the edges of the Metro/Access segment of the network (Figure 3.1, servers P1) and at the edge of the metro/core segment of the network (Figure 3.1, servers P2). Considering only P2 servers corresponds to the "distributed COMBO" architecture presented in Section 5.1.3.2 of Chapter 5. Considering also P1 servers corresponds to a (partial) implementation of Mobile Edge Computing (MEC).

Operators may choose to implement one of the above models, or a combination of models depending on the type of the area (e.g. densely versus sparsely populated).

3.3.1 Considered Scenarios

In order to estimate the gain of bandwidth in the metro network for 2020, we propose different scenarios. In all of them, we assume the same amount of fixed and mobile offloaded traffic (i.e. 57% in 2012 and 80% in 2020). The considered scenarios only differ in how offloaded traffic is served within the metro network. The rest of the traffic is served by central P4 servers shown in Figure 3.1.

Scenario 1: All offloaded traffic is served from P1 servers shown in Figure 3.1, that are located at the edge of the Access segment of the network; this corresponds to a MEC enabled architecture.

Scenario 2: All offloaded traffic is served from P2 servers shown in Figure 3.1, that are located within the current Metro network; this corresponds to the COMBO distributed architecture described in Section 5.1.3.2.

Scenario 3: All offloaded traffic is served from P3 servers shown in Figure 3.1, that are located within the current Metro network; this corresponds to the COMBO centralized architecture described in Section 5.1.3.1. This is also a typical location used by ISPs to distributed IPTV services to residential users relying on fixed and WiFi access networks.

We also consider hierarchical server architectures (Figure 3.1), where the most popular content are replicated very close to the users, while less popular content are distributed from more centralized servers. We (arbitrarily) selected some popularity distribution to design our last two scenarios.

Scenario 4 (No MEC implementation): 60% of offloaded traffic is served by P2 servers while the rest (40%) is served by P3 servers.

Scenario 5 (partial MEC implementation): 30% of offloaded traffic is served by P1 servers, the same quantity (30%) by P2 servers while the rest (40%) is served by P3 servers.

3.3.2 Gain in Terms of Traffic Demands Due to Distributing the Mobile Network Architecture

We now compare the respective amounts of traffic generated by mobile services for the years 2012 and 2020, over specific portions of the network and under the assumptions for our various scenarios.

Let X, Y, Z respectively represent the volumes in Total IP traffic, Total traffic generated by fixed services and Total traffic generated by the mobile services. Let also X_i, Y_i and Z_i respectively represent the volume in Total IP traffic that crosses the i network, the volume of Total fixed network's traffic that crosses the i network and the volume of Total mobile network's traffic that crosses network i; "i" represents one of the following network segments: Core, Metro/Core, Metro/Access.

Let G_i quantify the potential gain in terms of percentage of bandwidth demands made to the i network due to mobile traffic offloading. For instance, G_{Core} is the percentage of traffic crossing the Core network in a centralized mobile architecture that could be diverted from the Core network if mobile offloading were applied. The demands corresponding to this traffic would thus be served by servers located in the metro network (P1, P2 and P3 servers).

Obviously:

$$X_i = Y_i + Z_i \tag{3.1}$$

Let $\alpha_i =$ be the proportion of IP traffic that crosses the "i" network.

$$X_i = \alpha_i * X \tag{3.2}$$

In today's mobile architecture (Centralized LTE), none of the mobile traffic is offloaded, which implies that for all network segments i, $Z_i = Z$ and $Y_i = X_i - Z$.

In a distributed mobile architecture, mobile traffic can be offloaded. The assumption that the proportions of traffic generated respectively by fixed services and mobile services and are served in a given network portion are identical translates in:

$$\alpha_i = \frac{Z_i}{Z} = \frac{Y_i}{Y} \tag{3.3}$$

The volume of traffic demands generated by mobile services and offloaded from network segment i is $(Z - Z_i)$, which finally yields:

$$G_i = \frac{Z - Z_i}{Y_i + Z} \tag{3.4}$$

3.3.3 Gain in Terms of Bandwidth Demands at Core Network

The values for α_{Core} in 2012 and 2020 are respectively 0.43 and 0.2. Considering the distribution scenarios given in Section 3.3.1, we now apply in Table 3.3 the above derivation to the data given in Cisco [49] (2012) and Cisco [3] (2020) for X, Y and Z, in order to assess the impact on the Core network of distributing the mobile architecture.

TABLE 3.3: Quantifying gain for core network in terms of bandwidth demands due to distributing the mobile architecture in 2012 and 2020

	2012		**2020**	
X	43.570		194.374	
Y	42.685		163.810	
Z	0.885		30.564	
X_{Core}	18.735		38.874	
	Centralized LTE	Distributed LTE	Centralized LTE	Distributed LTE
Y_{Core}	17.850	18.354	8.310	32.762
Z_{Core}	0.885	0.380	30.564	6.112
G_{Core}	0	**2.62%**	0	**38.61%**

- In 2012, mobile traffic offloading allowed to reduce the volume of traffic supported by the core network by less than 3%. This gain could indeed be considered as

negligible compared to the investment (in terms of distributed PGWs deployment) required to achieve it, which fully justified a centralized mobile architecture.

- In 2020, almost 40% of the core network bandwidth can be offloaded if a distributed mobile architecture is implemented. This significant gain can probably justify the CAPEX increase due a distributed mobile architecture (which may be implemented for other reasons, not considered in the present Chapter).

3.3.4 Gain in Terms of Bandwidth Demands in the Different Network Portions

As the gain in distributing the LTE architecture is not significant in 2012, we focus the following study on predictions relative to 2020.

Considering the distribution scenarios given in Section 3.3.1, we can now apply the derivation given in Section 3.3.2 as in the previous subsection to derive the gain on various portions of the metro network due to using different distributions of servers in the metro network. The amount of gain in terms of bandwidth demands for each of these scenarios is shown in Figure 3.2.

Figure 3.2 of course shows that the gain in volume of traffic demands on the network segments depends on the considered scenario: the closer to the users the servers are located, the larger is the gain in terms of traffic demands, in each network portion.

It is important to notice that the gain in traffic demands translates differently, depending on the network portion, in various gains in terms of provisioned resources in these portions. Indeed, the closer the network portion is to the core network, the larger is the multiplexing gain to expect, and thus the smaller is the amount of resources to deploy for a given volume of traffic demands. Therefore an even better gain in terms of provisioned resources is to be expected by locating servers very close to the users.

For example, compared with Scenario 2, the deployment of the Scenario 1 would allow a significant gain of bandwidth at both segments of the metropolitan network. However, the amount of investments required for deploying scenario 1 would also be significantly higher than the investments required for Scenario 2. Indeed, to satisfy the large number of users/devices that are expected to connect to the network in the near future (billions), the network operators would have to multiply the number of PGWs and servers distributed within the Metro/Access segment of the network. Also, the procedures to ensure session continuity would have to be used at large scale, which would possibly increase the cost for supporting control and signalling traffic.

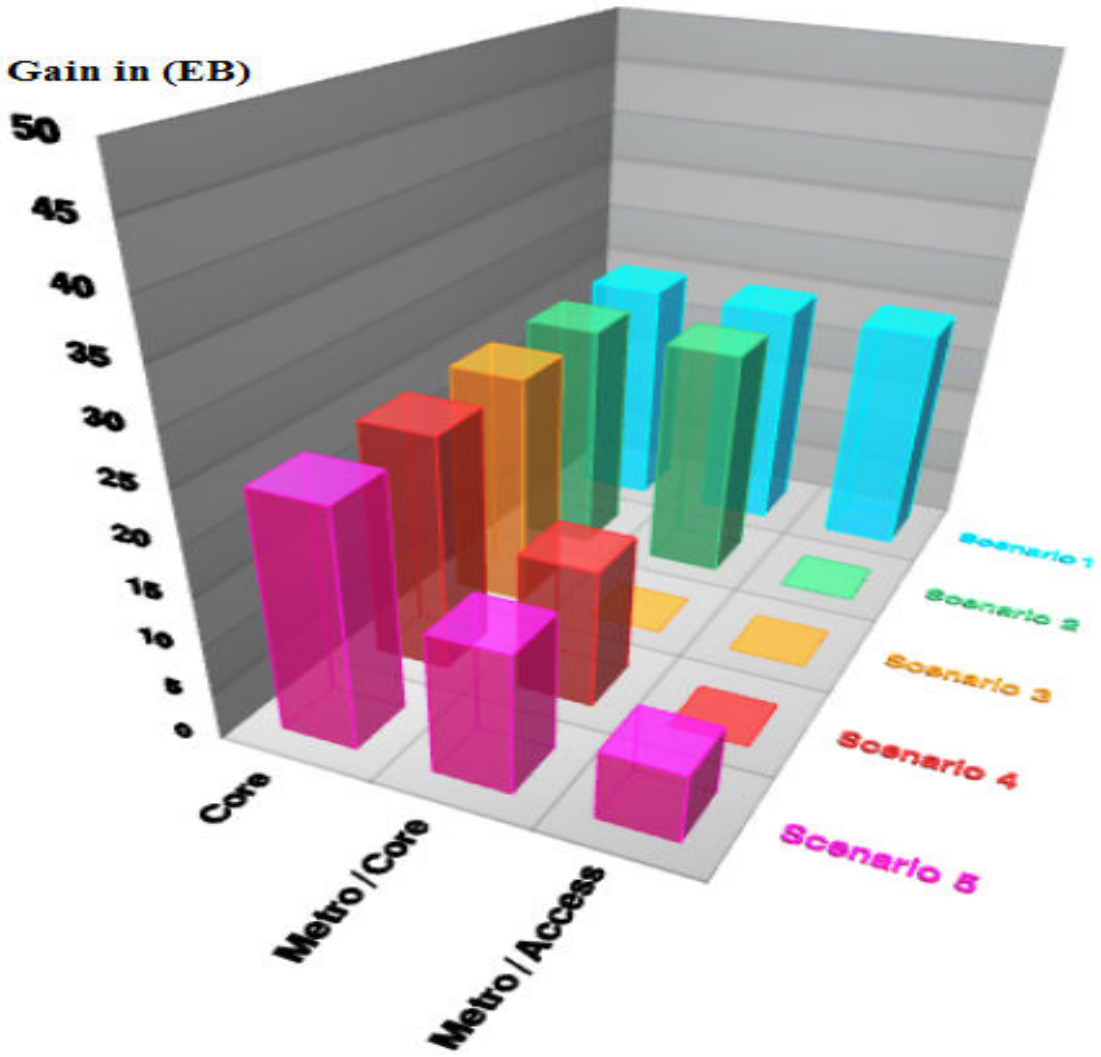


FIGURE 3.2: Gain (in volume) of bandwidth demands in the different network portions

As the amount of traffic generated by mobile services shall still be small in the near future compared to the amount of traffic generated by fixed services, the cost of distributing servers for mobile services may however be marginal as the same servers can potentially be used for fixed and mobile services. This type of consolidation is expected in the framework of fixed-mobile convergence for COMBO Project. In addition, the deployment of several 5G use cases also relies on moving the PGW data plane much closer to the user, as discussed later in Chapter 6. This in itself can be considered as a very significant gain for Telecom operators.

3.4 Conclusion of Chapter 3

In this Chapter, we have addressed the impact of distributing the mobile architecture on the amount of traffic generated by mobile services to be supported by different portions of

the network. First, we assumed that the various tools specified in the mobile architecture to offload mobile generated traffic were actually implemented: Classical LTE architecture and Beyond LTE architecture (described in Section 2.2.1 and Section 2.2.2 of Chapter 2). Then, using the Cisco documentations, we presented the potential forecasts on which content should or should not be distributed (offloaded). Next, using well-accepted traffic prediction assumptions, we assessed the traffic reduction on the core network and on various portions of the metro network, due to mobile traffic offloading, under different server distribution assumptions.

We have shown that the potential gain from server distribution for mobile traffic was indeed small in the past few years, but that in the near future, around 40% of backbone traffic could thus be offloaded. We also showed that, regardless of the investments required, distributing services within a full MEC architecture (scenario 1) could allow the same amount of gain at both segments of the metropolitan network (Metro/Core and Metro/Access).

In the next Chapter, we propose enhanced SIPTO implementations in order to support session continuity in case of mobility.

Chapter 4

Smooth Handover with SIPTO-Based Mobile Access

In a standard 3GPP architecture, a UE looses its previous connection whenever a PGW relocation is required. This corresponds to mobility use cases MC2 and MC3 presented in Section 2.3.2 of Chapter 2 for users with either SIPTO above RAN with co-located SGW/PGW or SIPTO at LN with LGW co-located with (H)eNB ongoing sessions. In these cases, the UE re-establishes a new SIPTO Connection, during which it gets a new IP address to re-access a host on the external IP network e.g. a content server. The content server may not be aware that the previous IP address and the new IP address reach the same user. This implies traffic interruption of the previous user session.

The "Smooth SIPTO" solution proposed in the present chapter is mainly used to maintain continuity of ongoing sessions in case of user mobility between different radio base stations.

This solution is driven by the idea of enhancing the 3GPP SIPTO architecture with MPTCP, in order to provide a smooth handover in the two mobility use cases MC2 and MC3 presented in 3GPP [4]. Part of the work described in this Chapter was published and presented in the HPSR Conference in 2015 [19].

In order to realize a smooth handover for SIPTO connections we propose to enhance the 3GPP SIPTO procedures by setting up an MPTCP connection between the UE and the server, assuming that both of them are MPTCP-capable. The enhancement of SIPTO with MPTCP will provide at least two data paths between the UE and the server: a first one towards a centralized PGW ("default PGW") within the EPC and a second one, relying on SIPTO to reach another PGW ("SIPTO PGW"), which is closer to the end-user. In order to provide a resilient MPTCP signalling channel, we also propose

to exchange all MPTCP signalling messages over the data path accessing the IP edge through the default PGW. The "always available" feature of this data path will ensure that MPTCP connection will not be broken even during user's mobility.

This Chapter describes the "Smooth SIPTO" proposal, which was designed by myself and in collaboration with my colleague Pratibha Mitharwal who mainly focused on MPTCP connection signalling.

Section 4.1 focuses on how "Smooth SIPTO" operates in MC2 and Section 4.2 respectively addresses MC3.

4.1 Smooth SIPTO Solution for MC2 (SIPTO above RAN with co-located SGW/PGW)

In this section, we first present how the MPTCP connection establishment is coordinated with the SIPTO data path establishment for MC2. Then, an overview of how the proposed solution provides handover in MC2. The successive steps of the solution are further detailed in the following subsections.

4.1.1 Coordinating MPTCP Connection Establishment with SIPTO Data Path Connection Setup

"Smooth SIPTO" relies on taking advantage of the multi-homing features provided by MPTCP. It is therefore necessary to initiate MPTCP whenever a user activates a SIPTO connection. Figure 4.1 illustrates how the MPTCP connection establishment is coordinated with the initiation of the SIPTO data path.

The successive steps to be carried out when the SIPTO connection is activated are the following (see Figure 4.1):

- When the UE is attached to the network, it receives a first IP address by the default PGW. Then, using this IP address, the UE connects to a server. The data path built between the UE and the server goes through the default PGW, and is thus called "default" data path.
- As mentioned previously, we assume that both UE and server are MPTCP-capable. Next, the UE establishes an MPTCP connection over the default data path. The default data path shall subsequently be used for all MPTCP signalling messages.

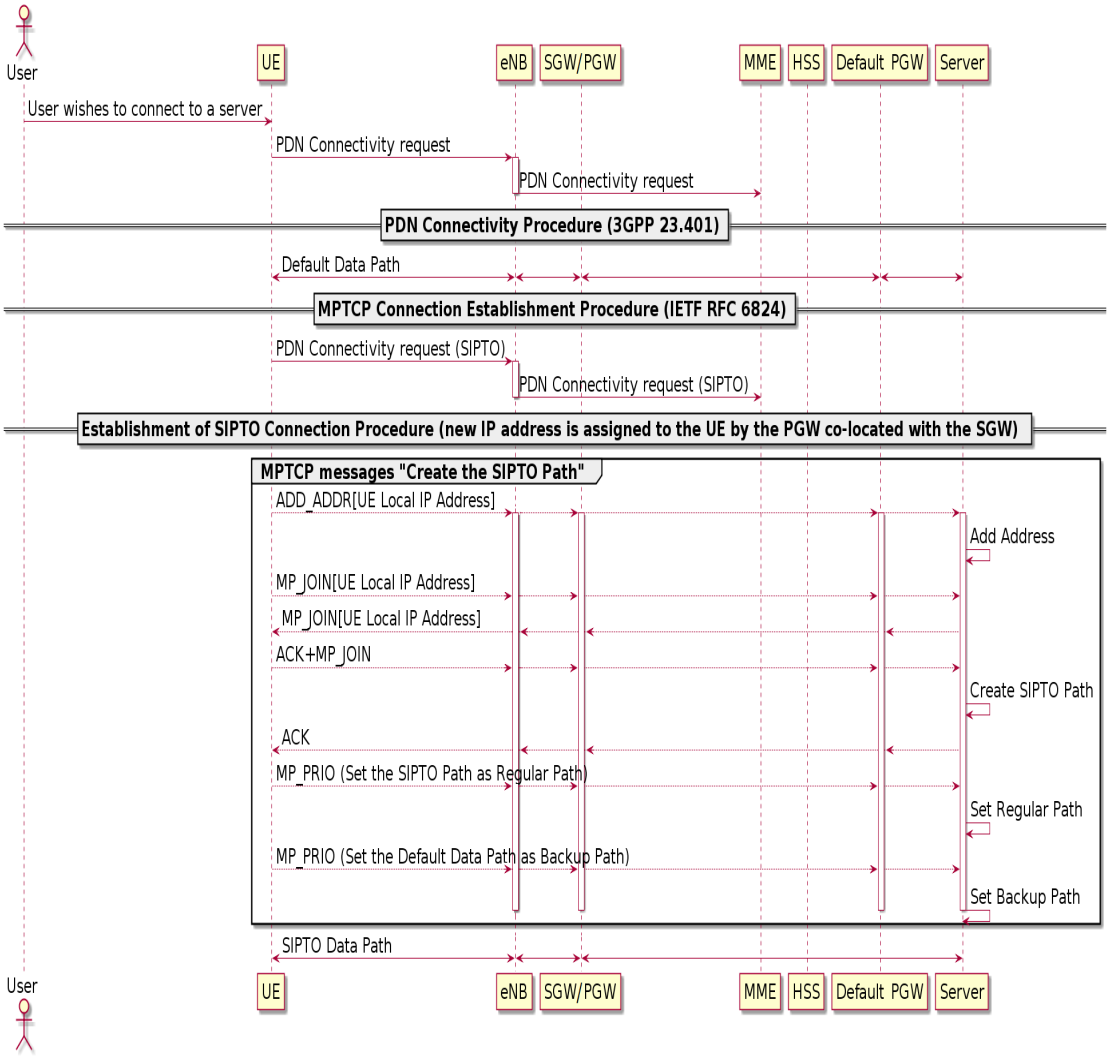


FIGURE 4.1: SIPTO Data Path Setup using MPTCP Connection

- Then, the UE requests the establishment of a SIPTO data path to the server. A SIPTO Connection Establishment Procedure is performed, and the UE thus obtains another IP address from another PGW, which is co-located with the SGW currently used. This address is called "local IP address". This procedure, which is part of SIPTO specification [4], is further detailed in Figure A.1.

- The local IP address is communicated to the server by the UE, in order to update the server's list of addresses used with MPTCP to communicate with the UE.

- Using the MP_JOIN option, the UE requests the creation of an MPTCP sub-flow between the server and the "local IP address".

- Finally, with the MP_PRIO option of MPTCP, the UE declares the default data path as "backup path" and the SIPTO data path as "regular path". This allows all downstream traffic from the server to be transmitted over the SIPTO path.

The SIPTO data path shown in Figure 4.1 corresponds to a SIPTO above RAN session using co-located SGW/PGW as illustrated in Figure 2.5 of Chapter 2.

4.1.2 Overview of Handover for Smooth SIPTO in MC2

Let us then assume that a UE initially attached to a first set of collocated SGW/PGW (see Figure 4.2-a) moves, and that it becomes necessary to attach it to a new co-located SGW/PGW. This could e.g., correspond to a user, travelling by train from one town to another. Handover consists in building new data paths for both Uplink (UL) and Downlink (DL) traffics; smooth SIPTO aims at maintaining session continuity while traffic moves from the initial paths to the new paths.

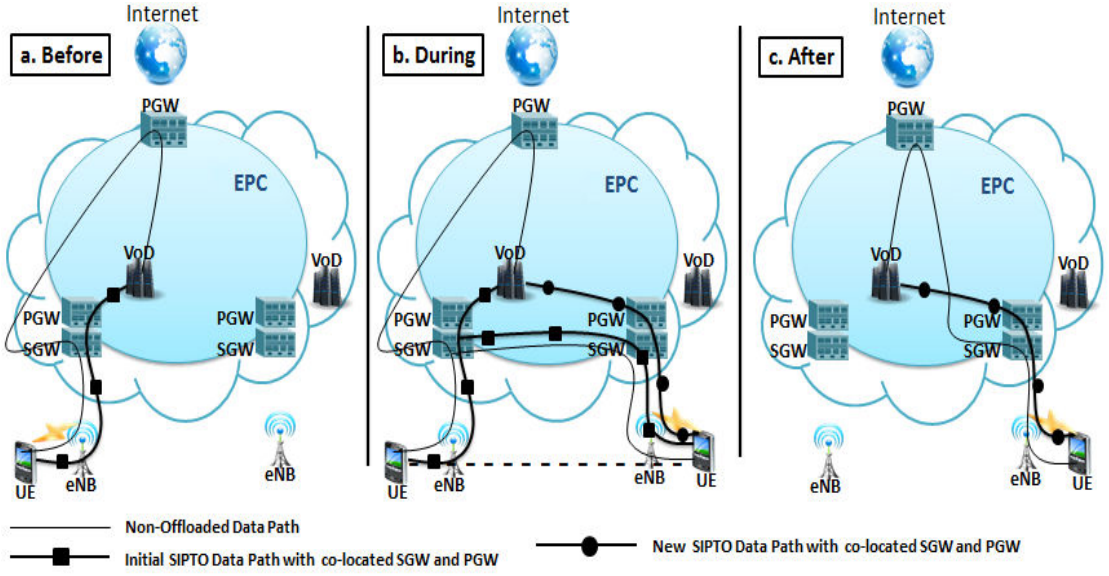


FIGURE 4.2: Mobility Scenario for a UE having a SIPTO connection with co-located SGW/PGW

In order to maintain session continuity for the offloaded traffic, our smooth SIPTO solution consists in handing over the current SIPTO data to the selected target eNB and target SGW while establishing a new SIPTO data path towards the target co-located SGW/PGW. As shown in Figure 4.2-b, thanks to MPTCP features, both SIPTO paths would be used in parallel to maintain the user's session continuity during user's mobility.

As shown in Figure 4.2-c, at the end of the smooth SIPTO enabled handover, the only path that would be used by downstream traffic is the new established SIPTO path.

In order to implement smooth SIPTO, new behaviours have to be introduced for the MME while other network elements' behaviours remain unchanged. The following requirements can be listed:

- Requirement 1: The MME must start an "inter eNB/inter SGW" S1-based handover procedure whenever a PGW relocation decision is taken; this differs from the standard 3GPP procedure in which the MME directly deactivates the initial SIPTO connection.
- Requirement 2: During the handover, the MME must request the establishment of new SIPTO connection towards the target co-located SGW/PGW, although the initial SIPTO connection is still considered as active.
- Requirement 3: The IP address provided by the new PGW is communicated to the server, and MPTCP procedures are used to allow traffic to be sent by the server to the UE over both initial and new SIPTO paths.
- Requirement 4: At the end of the Completion phase of the handover procedure, the MME must now deactivate the initial SIPTO connection.

4.1.3 Handover Procedure of Smooth SIPTO for MC2

Now that we have cited the requirements, let us see how they can be applied during user mobility. A global view of the handover procedure of smooth SIPTO for MC2 is illustrated in Figure 4.3.

As stated in [4] and mentioned in Chapter 2, the handover process is subdivided in to three phases: Preparation, Execution and Completion. The following sub-sections details each of the phases. All the steps carried out before and during Handover Preparation are illustrated in Figure 4.4. Handover Execution and Completion are respectively illustrated in Figure 4.6 and Figure 4.8. These figures and the successive steps identified to carry out handover are similar to the ones described in [4] and illustrated in Figure A.14; several modifications are proposed here in order to fulfill the requirements listed in Section 4.1.2.

4.1.3.1 Before Handover Preparation

Step 0. [S-eNB → UE] RRC Measurement Control:
According to 3GPP in [4], in mobile networks, eNBs are setup to send Radio Resource Control or RRC (as identified in Chapter 2) requests to their connected UEs for measurement control.

Step 1. [UE → S-eNB] RRC Measurement Report:
Upon the reception of a RRC Measurement Control message, a UE measures the signal

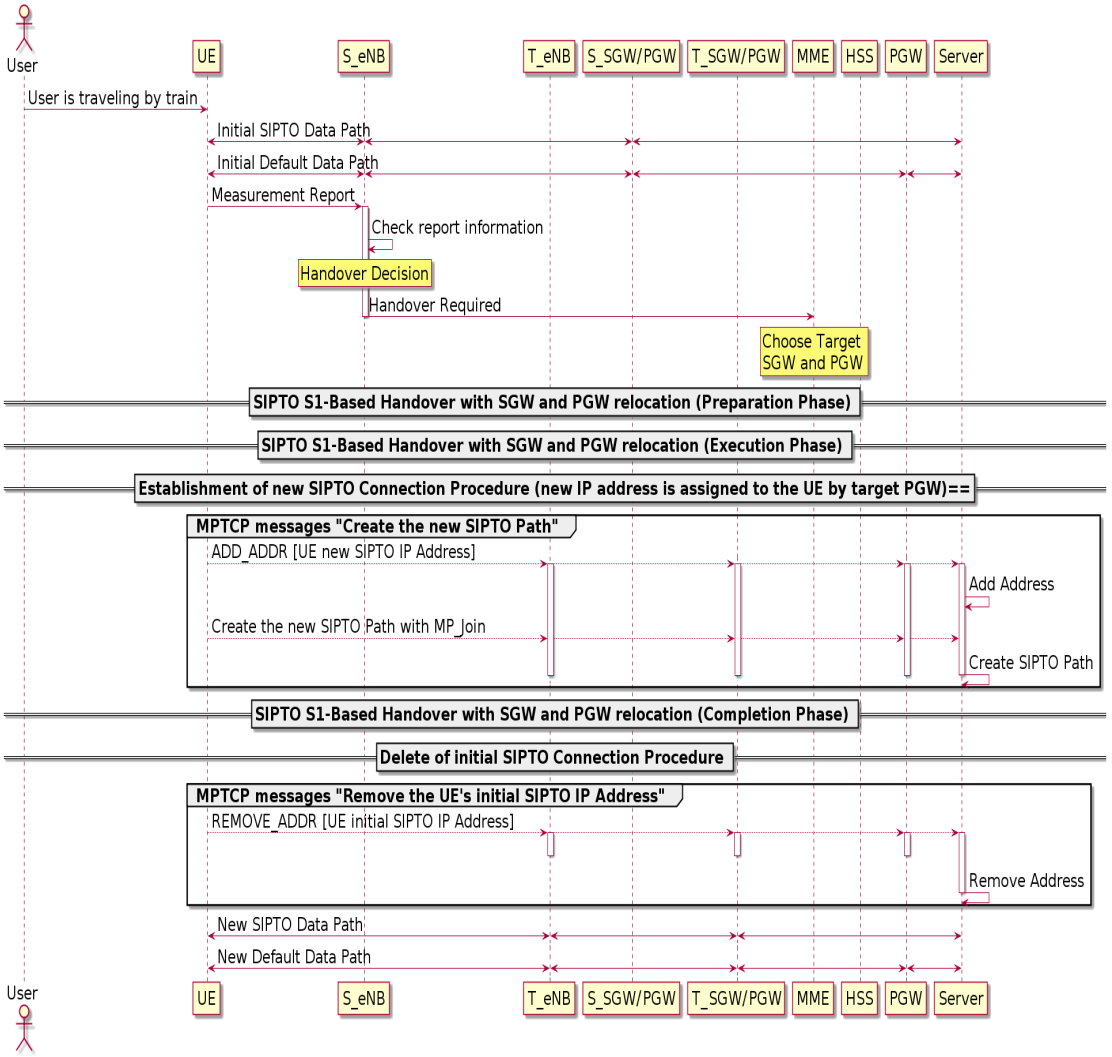


FIGURE 4.3: Handover Procedure of Smooth SIPTO for MC2

strength of neighbour cells and replies to its associated (source) eNB with an RRC Measurement Report message including the measurement result. Based on the information received in this message, the source eNB may select a target eNB which in MC2 implies an S1-based handover.

4.1.3.2 Handover Preparation

Step 2. [S-eNB → MME] Handover Required:
The source eNB sends a Handover Required message to the MME requesting a handover to the selected target eNB. The information included in this message is as follows:

- Target eNB ID: used to identify the target eNB.

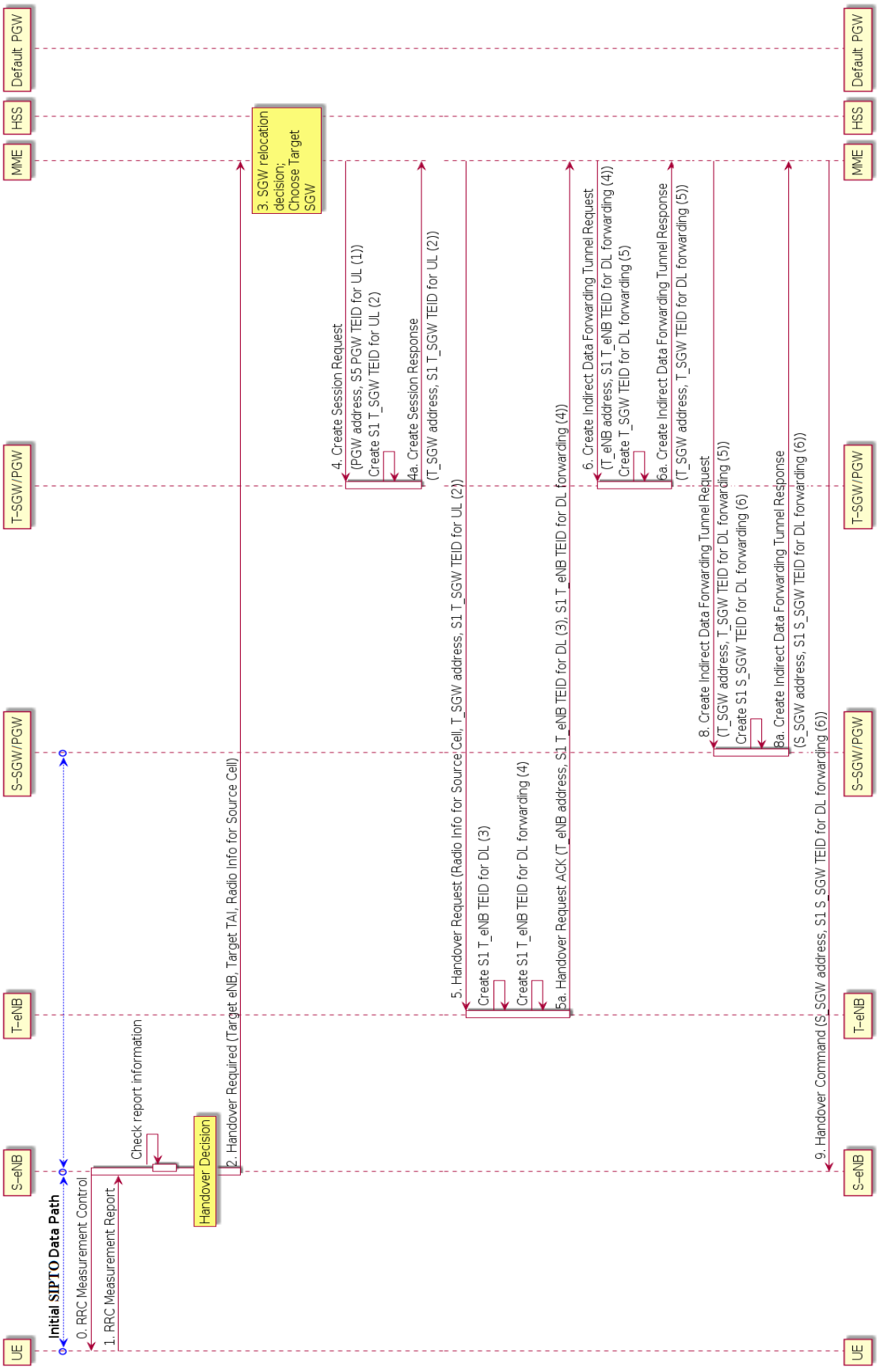


FIGURE 4.4: Handover preparation phase of smooth SIPTO for users with ongoing SIPTO above RAN sessions with co-located SGW/PGW

- Target TAI: used by the MME to identify whether the SGW, the PGW and/or the MME should be relocated or not.
- Radio Information for Source cell: used for re-ordering the packets during the forwarding of radio-related information of the source cell to target cell through the MME.

Step 3. [source SGW/PGW relocation decision and target SGW/PGW selection]:
The MME uses the TAI information received in Step 2 to decide whether an MME relocation is required or not. For the sake of simplicity, we assume here that no MME relocation is required, although the following can easily by adapted to the case of MME relocation. As we consider MC2, the TAI information indicates that the MME has to select a new co-located SGW/PGW and to initiate the establishment of an indirect forwarding tunnel between source eNB and target eNB.

Step 4. [MME → T-SGW/PGW] Create Session Request:
The MME sends a Create Session Request message to the selected target SGW. This message includes the source PGW's address and the TEID for the uplink traffic on the S5 interface (this is identified as S5 PGW TEID for UL (1) in Figure 4.4). The target SGW then allocates the address and TEID to be used by the target eNB for the uplink traffic on the S1-U interface (S1 T-SGW TEID for UL (2)).

Step 4a. [T-SGW/PGW → MME] Create Session Response:
The target SGW forwards the created information to the MME in a Create Session Response message.

Step 5. [MME → T-eNB] Handover Request:
The MME sends a Handover Request message to the target eNB on behalf of the source eNB. This message includes the information received from the target SGW (T-SGW address, S1 T-SGW TEID for UL (2)), in addition to the Radio information for Source cell received from source eNB in Step 2. The UE context, including security and tunnel information, is now created at the target eNB's level. The target eNB then allocates the downlink T-eNB address and TEIDs to be used by the network to reach the UE on the S1-U interface with the target SGW, for both the new path (S1 T-eNB TEID for DL (3)) and the forwarding path (S1 T-eNB TEID for DL forwarding (4)).

Step 5a. [T-eNB → MME] Handover Request ACK:
The target eNB forwards the downlink tunnel information to the MME in a Handover Request Acknowledgement message.

Step 6. [MME → T-SGW/PGW] Create Indirect Data Forwarding Tunnel Request:
Upon receiving the Handover Request ACK message, the MME creates an indirect

forwarding tunnel with the target SGW to ensure the seamless service provision for the UE. This message includes the S1 T-eNB address and TEID for DL forwarding (4) that have been allocated by the target eNB in Step 5. Following this message, the target SGW allocates the T-SGW address and TEID to be used by the source SGW on the downlink direction for the indirect forwarding with the target SGW (T-SGW TEID for DL forwarding (5)).

Step 6a. [T-SGW/PGW → MME] Create Indirect Data Forwarding Tunnel Response: The target SGW now sends the downlink forwarding information created on Step 6 (T-SGW TEID for DL forwarding (5)) to the MME.

Step 8. [MME → S-SGW/PGW] Create Indirect Data Forwarding Tunnel Request: The MME uses the forwarding information received in from the target SGW to created the indirect tunnel with the source SGW using a Create Indirect Data Forwarding Tunnel Request message. Becoming aware of the handover, the source SGW now allocates the address and TEID to be used by the source eNB to forward the user's downlink traffic on the uplink direction on the S1-U interface (S1 S-SGW TEID for DL forwarding (6)).

Step 8a. [S-SGW/PGW → MME] Create Indirect Data Forwarding Tunnel Response: The source SGW confirms the creation of the indirect forwarding tunnel and forwards the allocated address and TEID (S1 S-SGW TEID for DL forwarding (6)) to the MME.

Step 9. [MME → S-eNB] Handover Command:
At the end of the Preparation phase, the MME sends a Handover command message to the source eNB to inform it that the handover is now prepared. This message includes the information received on Step 8a.

The network status (i.e. established tunnels) following the handover Preparation phase is illustrated in Figure 4.5. At this stage, we note that traffic exchanged between the UE and the server still use the initial SIPTO data path, whereas the downlink forwarding paths as well as the new uplink paths have already been set up to proceed for the handover. As a result, each network element is now aware of the adequate destination of both UL and DL paths that should be used during the handover execution.

Compared with Classical LTE architecture with either separate SGW and PGW and/or co-located SGW/PGW where data path optimization is not a priority, the handover preparation phase of smooth SIPTO is very similar to the preparation phase of Classical LTE's S1-based handover explained in Section 2.3.1.2 of Chapter 2 and illustrated in Figure A.14 of Appendix A. The only difference is within *Step 3* where the initial PGW has to be relocated in order to fulfill Requirement 1.

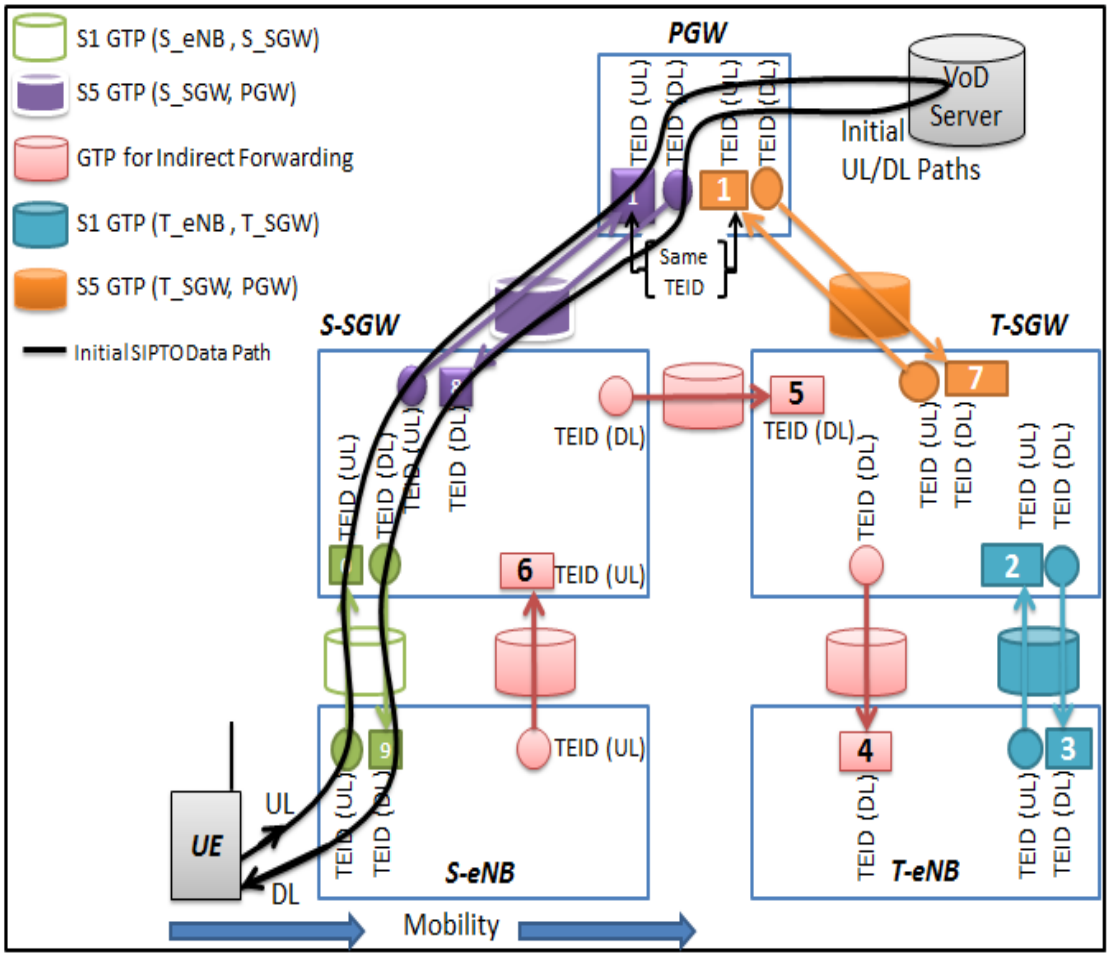


FIGURE 4.5: Network status at the end of the handover Preparation phase of smooth SIPTO for MC2

4.1.3.3 Handover Execution

Step 9a. [S-eNB → UE] Handover Command:
Now that the two eNBs are ready to perform a handover, the handover Execution phase starts by detaching the UE from the source eNB and by synchronising it to the target eNB. This is performed by the source eNB which first sends a Handover Command message including an RRC Connection Reconfiguration request to the UE in order to start the synchronization.

Steps 10/10c. [S-eNB → MME] then [MME → T-eNB] eNB status transfer:
The source eNB transfers its status to the target eNB over the S1-MME interface between both eNBs and the MME as shown in Figure 4.6.

Steps 11 to 11c. Forwarding DL Data Path:
At this stage, the downlink traffic forwarding between source and target eNBs over the indirect forwarding tunnel begins. The traffic received by source eNB is first sent to the

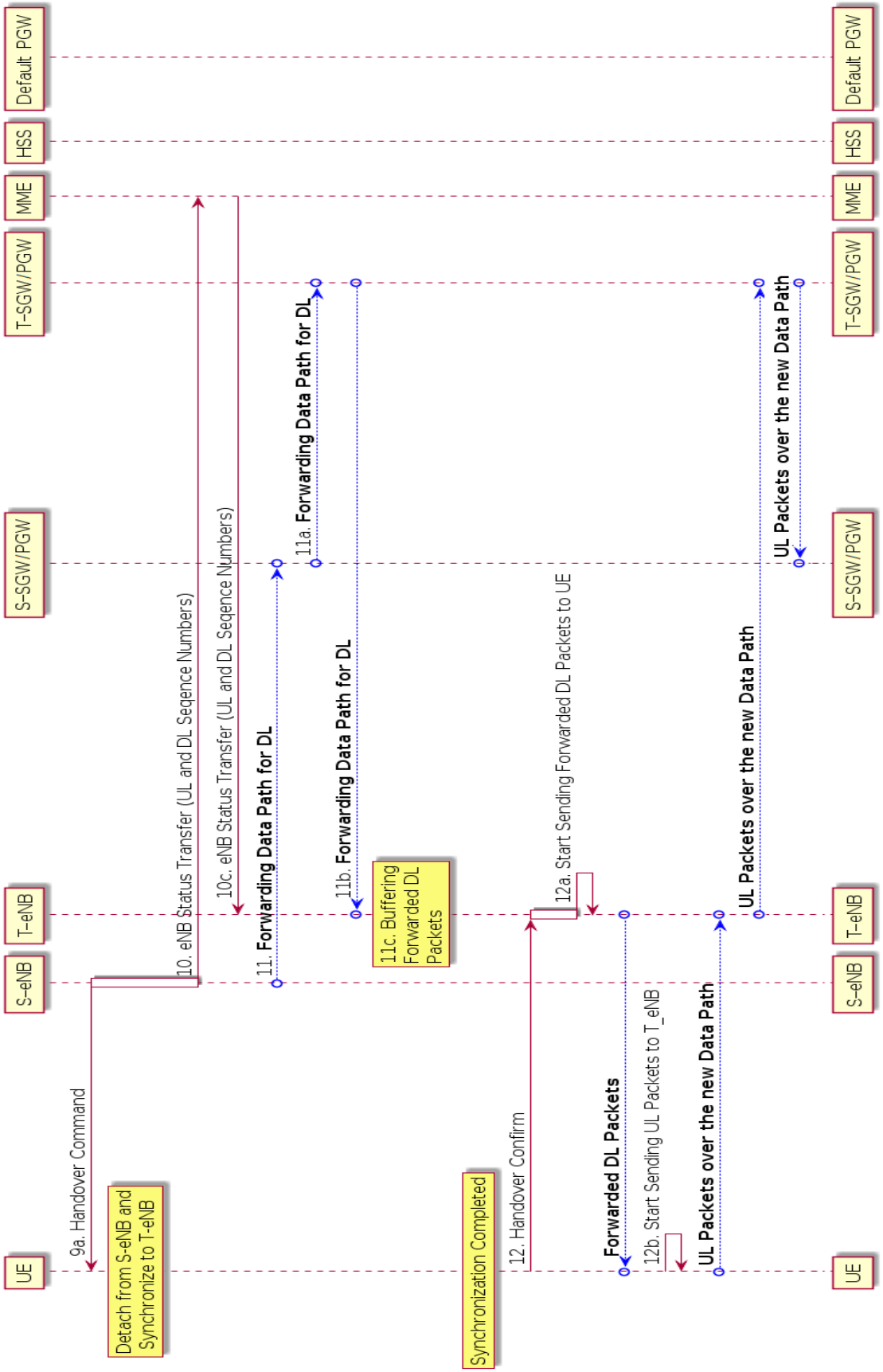


FIGURE 4.6: Handover execution phase of smooth SIPTO for users with ongoing SIPTO above RAN sessions with co-located SGW/PGW

source SGW, which then forwards it to the target SGW through the Indirect Forwarding Tunnel between source and target SGWs. The latter then sends this traffic to the target eNB, which start buffering the received packets until the UE's synchronization is completed. Finally, buffered packets are all sent to the UE and uplink packets are delivered to the source PGW directly through target eNB and target SGW. Handover is now executed, thanks to the Indirect Forwarding Tunnel.

Steps 12/12a. [UE → T-eNB] Handover Confirm:
Once the synchronization with the target eNB is completed, the UE confirms this information with the target eNB, which starts sending the forwarded downlink packets to the UE.

Step 12b. Data Transfer
The UE on the other side, starts sending the uplink packets to the target eNB. These packets are then sent to the target SGW on the S1-U interface and finally, forwarded to the source PGW on the S5 interface between the target SGW and the source PGW. Both uplink and downlink forwarding paths used during the handover Execution phase are illustrated in Figure 4.7.

Note here that the steps performed in the handover Execution phase of smooth SIPTO (9a to 12b) are identical to those performed in the handover Execution phase of Classical LTE mobility use case as illustrated in Figure A.14 of Appendix A.

4.1.3.4 Handover Completion

To realize a smooth SIPTO during users mobility, a new SIPTO data path must be established before releasing the initial SIPTO path resources. The basic idea here is to ensure that both initial and new SIPTO data paths will be simultaneously active to transfer packets between the UE and the server in order to avoid traffic interruption.

In Classical LTE, the traffic is interrupted within the execution phase of the handover process. In order to implement smooth SIPTO, the interruption time during the handover process of the initial SIPTO traffic should be minimized as much as possible and therefore, the new SIPTO path is established only when the handover Execution phase of the initial SIPTO traffic is completed. This is illustrated in Figure 4.8.

Step 13. [T-eNB → MME] Handover Notify:
The handover Completion phase of our smooth SIPTO solution for MC2 starts when the target eNB notifies the MME that the handover is executed and that the handover Completion phase can now begin.

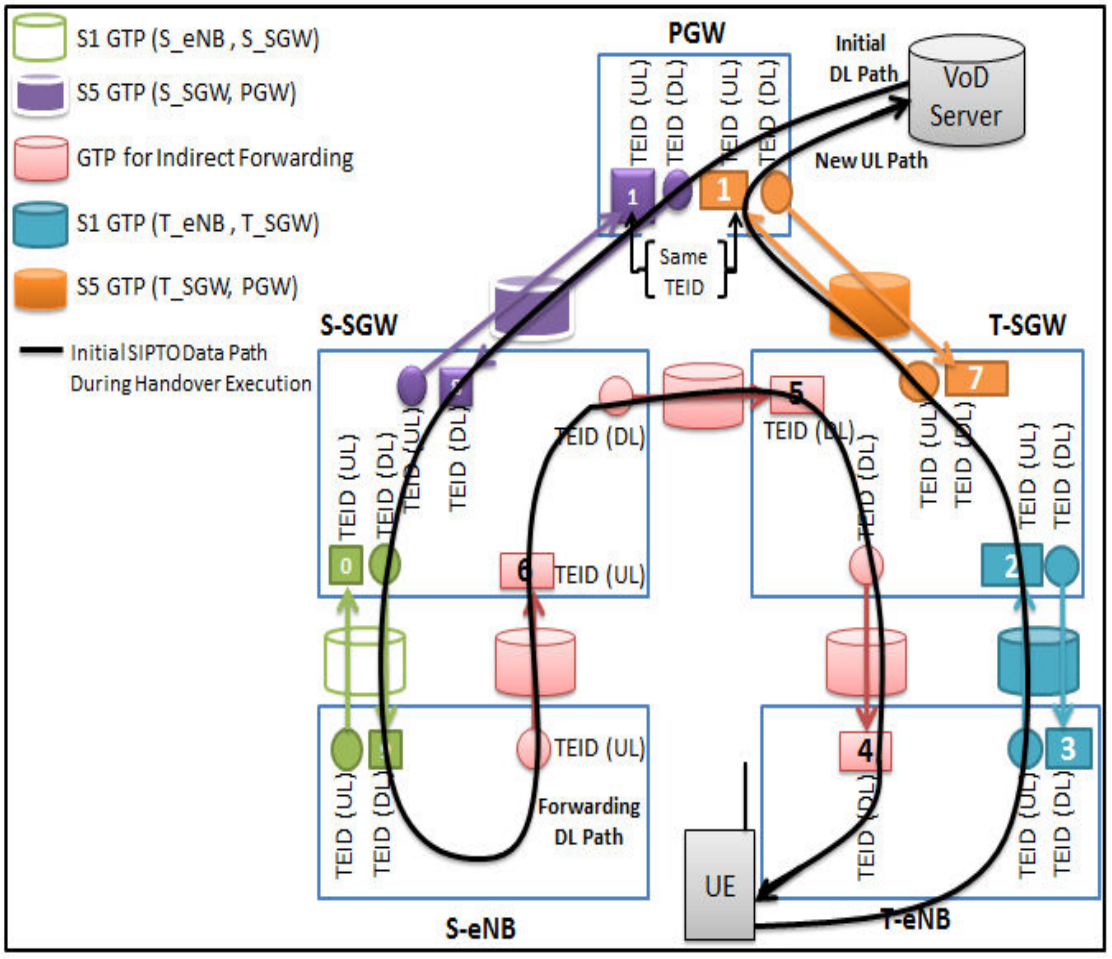


FIGURE 4.7: Network status during the handover execution phase of smooth SIPTO for MC2

Then a timer is launched. In order to ensure that the initial SIPTO data path will not be broken before the establishment of the new data path, the handover completion phase of the initial SIPTO data path must be delayed until the new SIPTO data path establishment procedure is completed. The standard handover process is thus modified by increasing the duration of the resource release timer at the MME level; the additional delay should be larger than the delay required for the establishment of the new SIPTO connection.

The new SIPTO data path establishment procedure towards the selected target co-located SGW/PGW can now be launched in order to fulfil Requirement 2. The UE communicates its new IP address to the distant server and creates new MPTCP sub-flow, using MP_JOIN option of MPTCP (see Figure 4.3), as "new SIPTO path"; this is done over the backup path which is always on. The UE has now three available data paths to the server: the default (backup) data path towards the EPC, the initial SIPTO data path towards the co-located source SGW/PGW using the Indirect Forwarding

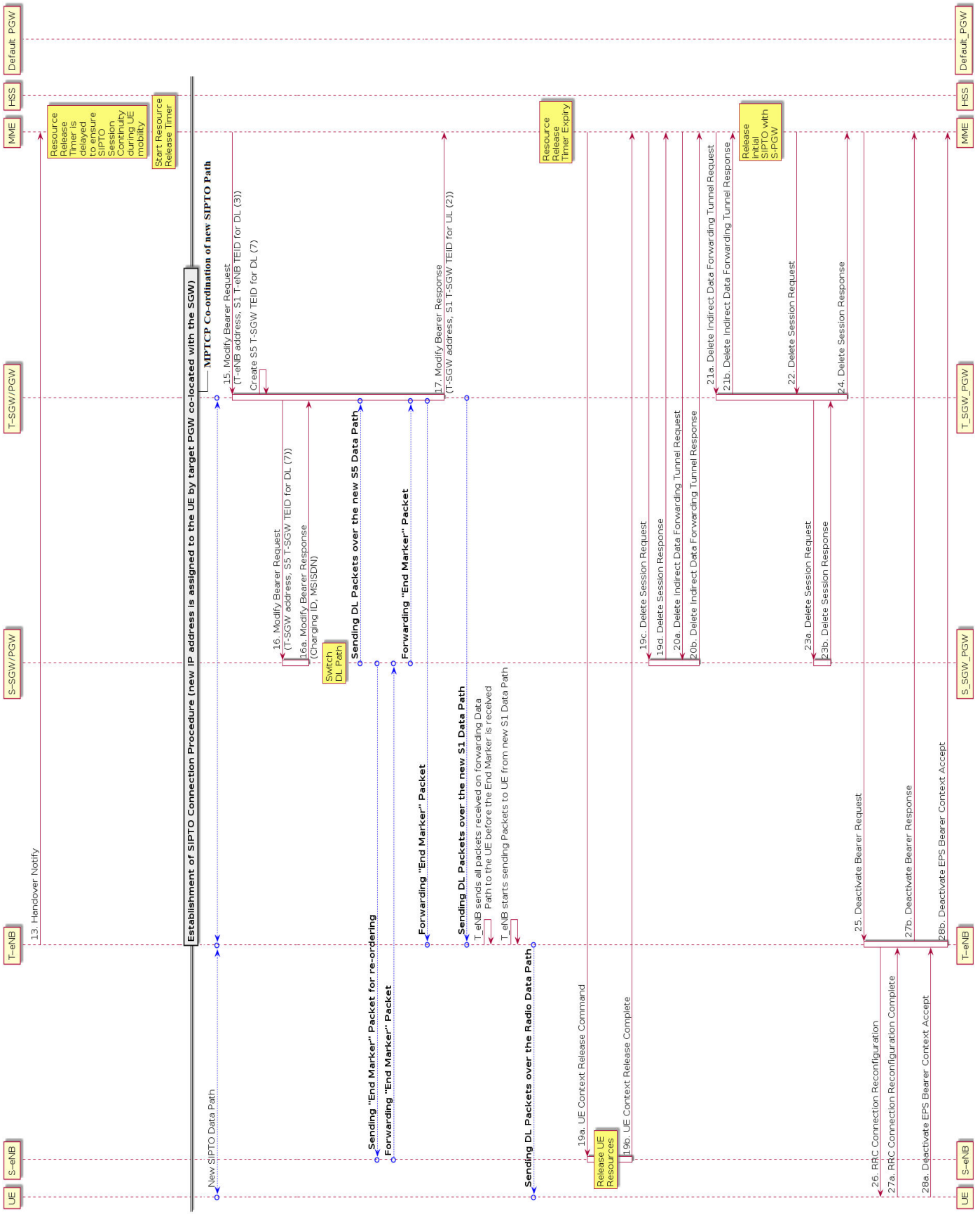


FIGURE 4.8: Handover completion phase of smooth SIPTO for users with ongoing SIPTO above RAN sessions with co-located SGW/PGW

Tunnel and the new SIPTO data path towards the co-located target SGW/PGW (see Figure 4.2-b). This fulfils Requirement 3.

Step 15. [MME → T-SGW/PGW] Modify Bearer Request:
The MME sends a Modify Bearer Request message to the target SGW. Since the source SGW is relocated, the target SGW assigns the IP address and TEID for downlink traffic coming from the source PGW (S5 T-SGW TEID for DL (7)).

Step 16. [T-SGW/PGW → S-SGW/PGW] Modify Bearer Request:
The target SGW forwards the Modify Bearer Request to the source PGW in order to switch the downlink packet delivery path from source PGW towards target SGW instead of source SGW.

Step 16a. [S-SGW/PGW → T-SGW/PGW] Modify Bearer Response:
The source PGW acknowledges the bearer modification with a Modify Bearer Response message and switches the Path on the S5 interface to go towards the target SGW. The source PGW then sends an "End Marker" packet on the initial SIPTO path to the source eNB, which forwards it to the target eNB over the indirect forwarding tunnel. This packet allows coordinating the data at the target eNB's level.

Step 17. [T-SGW/PGW → MME] Modify Bearer Response:
The target SGW confirms the switching by sending a Modify Bearer Response message to the MME including the target SGW address and TEID allocated in Step 4 for uplink traffic (S1 T-SGW TEID for UL (2)). The target SGW then starts sending downlink packets to the target eNB using the new downlink path on the S1 interface. At this stage, the target eNB sends all packets received on the indirect forwarding data path to the UE before the End Marker is received. Then, it starts sending packets to the UE from the new downlink path.

Note that, according to 3GPP in [4], in case the new (target) TAI that is sent to the user in Step 2 is not in the list of TAIs associated to the UE, the UE has to initiate a Tracking Area Update (TAU) procedure as shown in Appendix A.

Steps 19a to 19b. UE Context Release:
Upon completion of the TAU procedure, the MME asks the source eNB to release the UE resources via a UE Context Release Command message. Once this is done, the source eNB confirms the context release by replying to the MME with a UE Context Release Complete message.

Steps 19c to 19d. Delete initial SIPTO' resources:
Upon receiving the UE Context Release Complete message, the MME deletes the EPS bearer resources that were used by the source SGW to maintain the initial SIPTO path.

Steps 20a to 21b. Delete Indirect Data Forwarding Tunnels:
The MME deletes the indirect forwarding tunnels created in Steps 6 to 8a during the handover Preparation phase.

Steps 22 to 28b. Delete initial SIPTO Connection:
We now delete the initial SIPTO connection.First, MPTCP is used to delete the MPTCP sub-flow originally set up for the initial SIPTO path; this is done over the backup path. Then, the MME deactivates the initial SIPTO connection. This is done similarly to the PDN disconnection procedure illustrated in Figure A.16 of Appendix A. Besides, as per MPTCP procedures, the IP address initially allocated by the source PGW for this UE is deleted from the list of addresses stored within the server. Thus, Requirement 4 is now fulfilled.

The UE now only has two MPTCP sub-flows: the Backup data path towards the EPC and the SIPTO data path towards the new co-located SGW/PGW. The final user's offloaded and non-offloaded data paths are shown in Figure 4.2-c. The network status at the end of the handover procedure of smooth SIPTO for MC2 is illustrated in Figure 4.9 in which we note that all resources that were used for the initial SIPTO connection are now released.

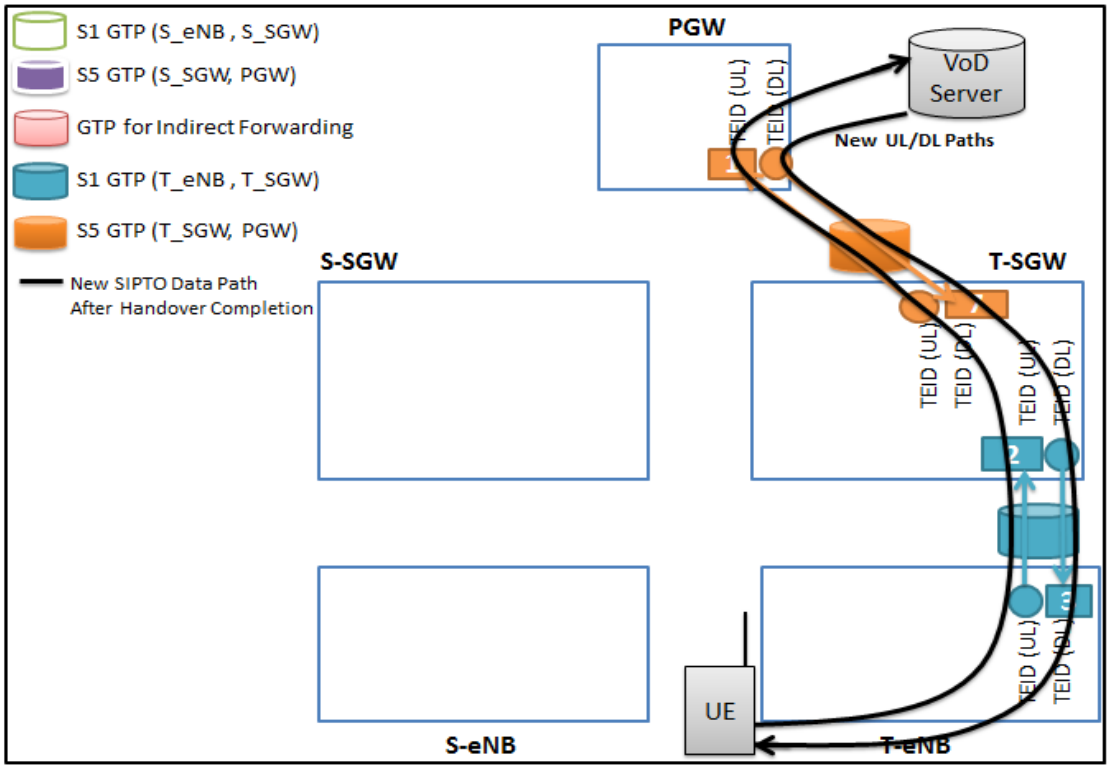


FIGURE 4.9: Network status at the end of the handover procedure of smooth SIPTO for MC2

Compared with the Completion phase of Classical LTE S1-based handover, our smooth SIPTO solution for MC2 proposes the use of MPTCP features to ensure the continuity of SIPTO above RAN sessions during users mobility. The main difference between the two Completion phases appears when we delay the resource release timer in order to take account of (1) the duration of a new SIPTO connection establishment and (2) the inclusion of this connection to the MPTCP connection using MP_JOIN. The rest of the handover completion phase is then performed similarly to Classical LTE.

Besides, steps 22 to 28b were introduced to the completion phase of our smooth SIPTO proposal for MC2 in order to delete the initial SIPTO path.

4.2 Smooth SIPTO Solution for MC3 (SIPTO at LN with LGW co-located with HeNB)

In this section, we first present an overview of how the proposed smooth SIPTO solution provides handover in local mobility (MC3). Then the successive steps of the handover procedure for local mobility support are further detailed in the following subsections.

4.2.1 Overview of the Smooth SIPTO Solution for MC3

To support user's local mobility, smooth SIPTO solution introduces the notion of "Proxy-SGW" to the 3GPP architecture with SIPTO at LN. A Proxy-SGW is a functional block which is included within the LGW. Proxy-SGW is a set of internal functions that is only seen by the co-located HeNB and LGW and is unseen by the rest of the network equipment. Proxy-SGW is seen as a HeNB by the LGW and as an LGW by the HeNB. In order to maintain the compatibility to 3GPP architecture, we propose to connect Proxy-SGW to the HeNB over an S1-U interface with GTP-U protocol and to the LGW over an S5 interface with GTP-U protocol for user plane and GTP-C protocol for control plane. This is illustrated in Figure 4.10.

Let us assume that a UE is having an offloaded session towards a server (e.g. a VoD server) relying on SIPTO at LN. The UE is then connected to a HeNB, which is co-located with a LGW including Proxy-SGW function (see Figure 4.11-a). Let us also assume that this UE moves towards another HeNB, that is also co-located with a LGW including Proxy-SGW functionality, and that due to this mobility it becomes necessary to attach this UE to the target LGW. This could e.g., correspond to a student walking around a large University Campus, and attaching successively its UE to different HeNBs.

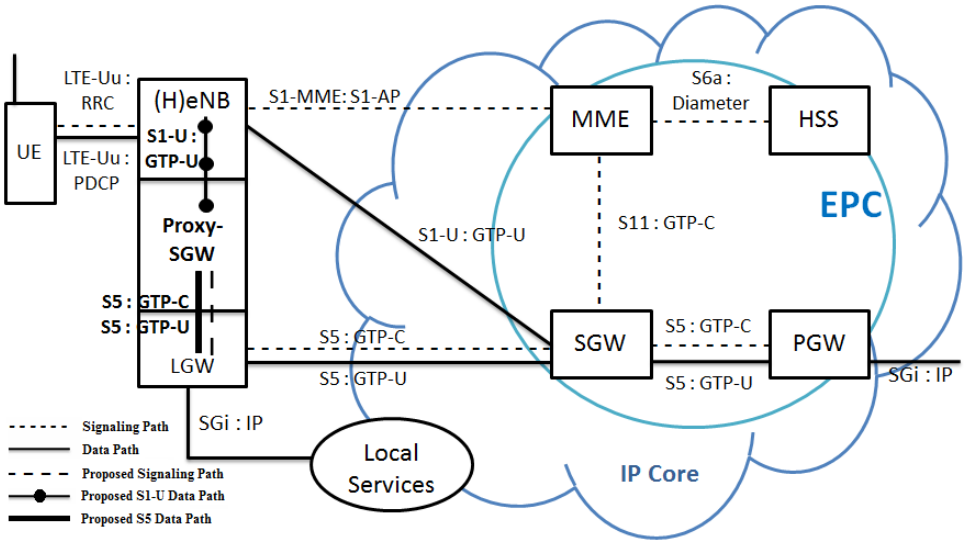


FIGURE 4.10: Location of a "proxy-SGW" in a LGW Co-located with HeNB

In order to maintain session continuity for the offloaded traffic, smooth SIPTO solution for MC3 consists in handing over the current SIPTO data to the selected target HeNB using an Indirect Forwarding Tunnel between the source and target HeNBs via the source and target Proxy-SGW/LGWs, while establishing a new SIPTO data path towards the target HeNB and its co-located LGW. This is shown in Figure 4.11-b. Thanks to MPTCP, both SIPTO paths would be used in parallel to deliver the user's data traffic of the current SIPTO at LN session.

The indirect tunnel between the source and target Proxy-SGWs is mainly used to ensure the forwarding of both uplink and downlink data traffic of the current SIPTO session.

As shown in Figure 4.11-c, at the end of the smooth SIPTO enabled handover for MC3, the only path that would be used for the offloaded local traffic is the new established SIPTO path.

4.2.1.1 Selecting the Target Proxy-SGW/LGW

In smooth SIPTO solution for MC3, the identification of the target co-located Proxy-SGW/LGW is indispensable to perform the handover.

Originally, during a Classical LTE handover procedure, the target SGW is selected thanks to the "SGW selection function" included within the MME, whereas no further PGW selection functionality is performed [4]. Since the LGW in SIPTO at LN architecture supports similar functions as PGW (e.g., UE' IP address allocation), it was considered in [4] that no further LGW selection could be performed in SIPTO for MC3 (see Section 2.3.2.2 of Chapter 2).

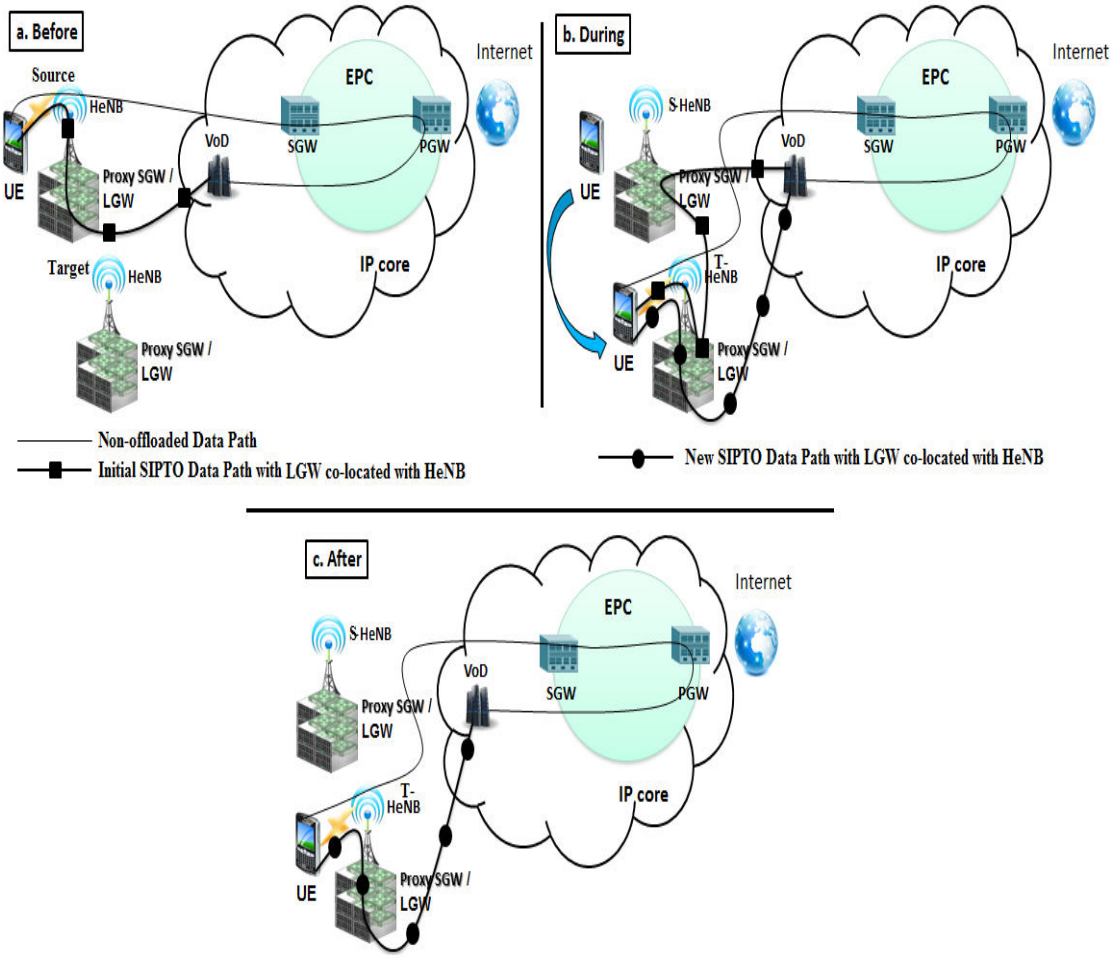


FIGURE 4.11: Mobility Scenario of a UE having a SIPTO at LN session with Proxy SGW function included in LGW

To achieve a seamless local mobility in our smooth SIPTO solution for MC3, the gateway selection function within the MME must then be enhanced to enable a target Proxy-SGW (and thus target LGW) selection, whenever a LGW relocation is required.

According to 3GPP in [4, 6], a HeNB includes the IP address of its connected LGW and sends it to the MME in every INITIAL UE MESSAGE and every uplink signalling message between the UE and the EPC network. However, in our smooth SIPTO solution for MC3 the target Proxy-SGW information address must be used during the handover before receiving any uplink signalling message from the target HeNB (i.e., before receiving the Handover Request ACK message).

Therefore, to select the target Proxy-SGW/LGW, I have been inspired, in smooth SIPTO solution for MC3, by the idea of identifying the set of HeNBs that are having IP connectivity to local PDNs via one or several LGWs with a "Local HeNB Network identifier (LHN-id)" [7]. Herein, as shown in Figure 4.12, I assume that each HeNB that

is having an IP connectivity towards one or different local PDNs using its co-located LGW, is identified with a LHN ID.

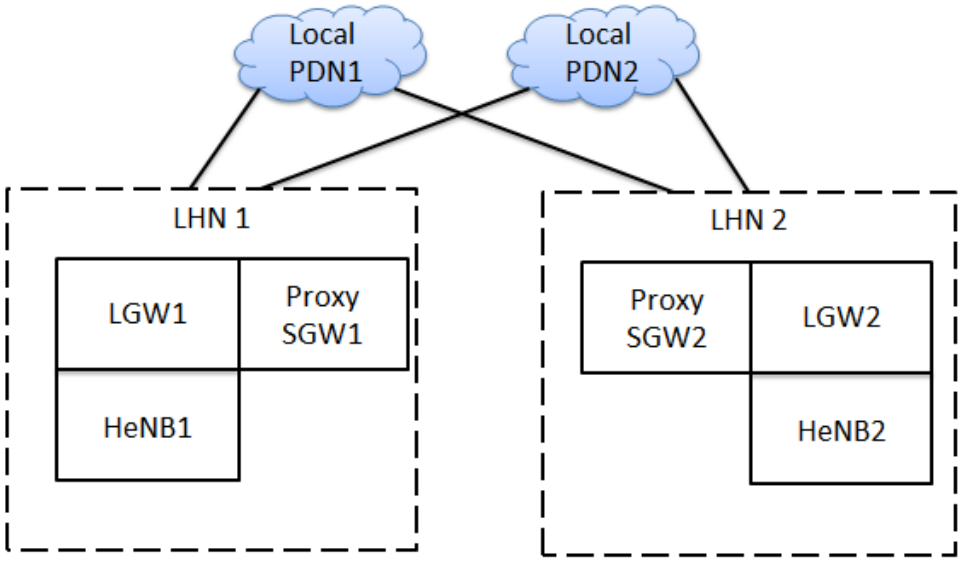


FIGURE 4.12: Identification of each HeNB that is co-located with a LGW with a "LHN id"

The LHN-id is then used by the MME to select the target Proxy-SGW/LGW using a DNS-based alternative (i.e., DNS Query/Response).

4.2.1.2 Requirements of Smooth SIPTO Solution for MC3

To allow a seamless local mobility, smooth SIPTO for MC3, differs from standards 3GPP by introducing new behaviours for MME, LGW and (H)eNB equipment while keeping other equipment's behaviours (SGW, PGW, etc.) unchanged. The MME new behaviours for SIPTO at LN mobility includes those defined in Section 4.1.2. HeNB's, LGW's and additional MME's new behaviours have to fulfil the following requirements:

- Requirement 5: The interface selection function within the source HeNBs, must be enhanced to enable "an always S1-based handover" for users with ongoing SIPTO at LN sessions.
- Requirement 6: Each HeNB must have a list of the set of LHN-identifiers of its neighbour HeNBs.
- Requirement 7: The source HeNB must include the target LHN-id in every "Handover Required" message.

- Requirement 8: Upon the reception of the target LHN-id, and if a LGW relocation is required, the MME must send a "DNS Query" message towards a DNS server in order to obtain the target Proxy-SGW/LGW IP address.
- Requirement 9: The MME must send the "Create Session Request message" to the source Proxy-SGW instead of the target ProxySGW, then it must send a "Create Tunnel Request" to the target Proxy-SGW in order to create a tunnel between the source and target ProxySGWs.
- Requirement 10: the LGW must forward all the signalling messages, received from the SGW (within the EPC) and related to the establishment of the indirect forwarding tunnel, to the Proxy-SGW.

4.2.2 Handover Procedure of Smooth SIPTO for MC3

Now that we have cited the requirements, let us see how they can be applied during the user's mobility. Figure 4.13 illustrates a global view of the handover procedure of smooth SIPTO for MC3.

Similar to Section 4.1.3, the different handover phases of smooth SIPTO for MC3 are detailed in the following sub-sections.

All the steps carried out during the Handover Preparation are illustrated in Figure 4.14. Handover Execution and Completion are respectively illustrated in Figure 4.16 and Figure 4.18. These figures and the successive steps identified to carry out handover are similar to the ones described in [4] and illustrated in Figure A.14; several modifications are proposed here in order to fulfill the requirements listed in Section 4.1.2 and Section 4.2.1.2.

4.2.2.1 Before Handover Preparation

As shown in Figure 4.13, before the handover process Step 0 and Step 1 presented in subsection 4.1.3.1 are performed to measure the signal strength of neighbour cells (HeNBs). Upon taking a handover decision, the HeNB selects S1 interface to be used for this handover, which fulfill Requirement 5.

4.2.2.2 Handover Preparation

Step 2. [S-HeNB → MME] Handover Required:
The source HeNB now sends a Handover Required message to the MME requesting a

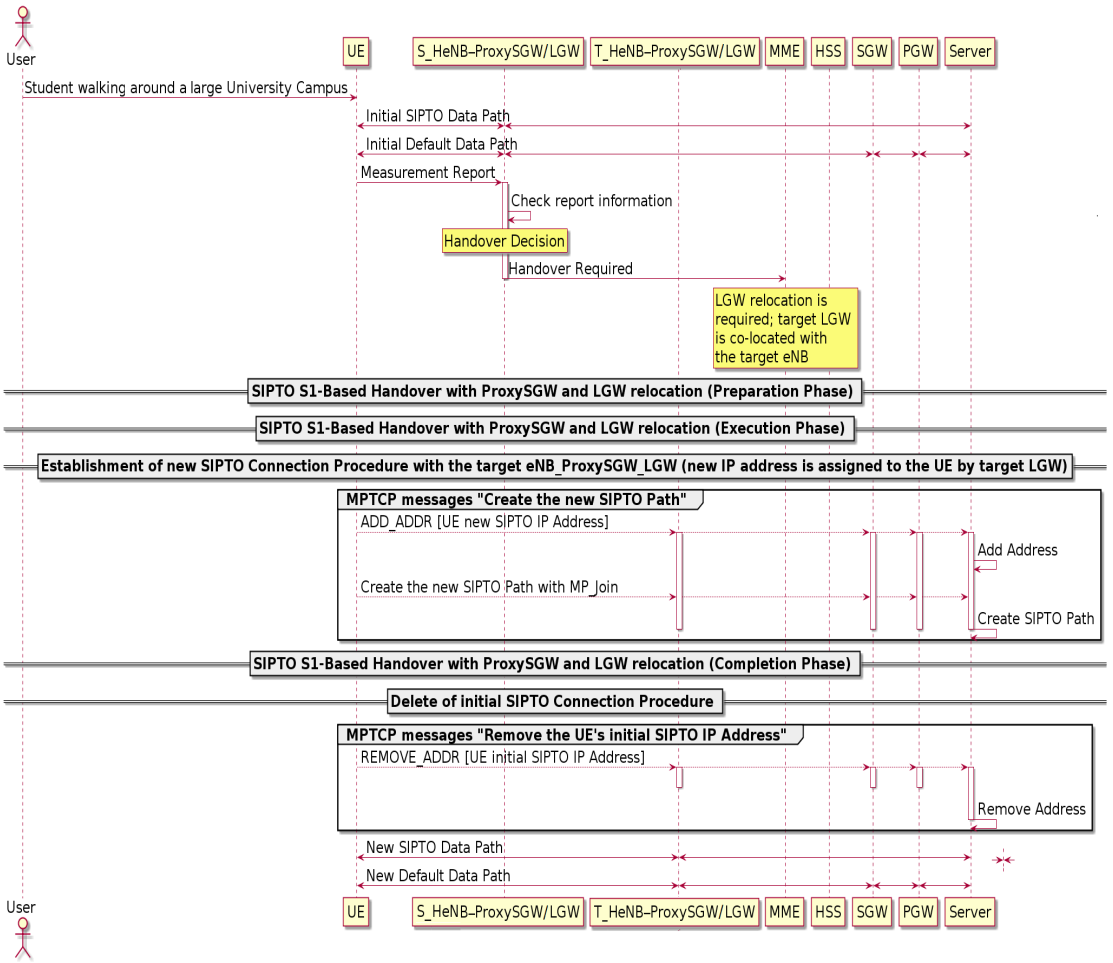


FIGURE 4.13: Handover Procedure of Smooth SIPTO for MC3

handover to the selected target HeNB. The information included in this message is as follows:

- Target eNB ID: used to identify the target eNB.
- Target Tracking Area ID: used by the MME to identify whether the SGW, the PGW, the LGW and/or the MME should be relocated or not.
- **Target LHN-id**: used to identify the target LGW. This fulfill Requirements 6 and 7.
- Radio Information for Source cell: used for re-ordering the packets during the forwarding of radio-related information of the source cell to target cell through the MME.

Step 3. [LGW relocation decision and selection of target Proxy-SGW/LGW]:
During the handover process of smooth SIPTO solution for MC3, since the users mobility

is limited by distance, then neither the MME nor the SGW are relocated. As we consider MC3, where source LGW is co-located with the source HeNB, an LGW relocation is then required.

A DNS Query including the target LHN-id received in (Step 2.) can now be launched by the MME in order to fulfil Requirement 8.

Upon receiving the Proxy-SGW IP address, the MME proceed for the establishment of the indirect forwarding tunnel between the source and target HeNBs.

Step 4. [MME → SGW → S-LGW → S-ProxySGW] Create Session Request:

We recall here that all messages addressed to a Proxy-SGW are first routed from the MME to the SGW over the S11 interface, then to the LGW (co-located with that Proxy-SGW) over the S5 interface, which finally forwards them to its co-located Proxy-SGW over the new direct S5 interface between them, and vice-versa. This fulfill Requirement 10.

The MME sends a Create Session Request message to the source Proxy-SGW in order to fulfill Requirement 9. This message includes the source LGW's address and TEID for the uplink forwarding traffic on the new direct S5 interface between the source co-located ProxySGW and LGW (this is identified as S5 S-LGW TEID for UL (1) in Figure 4.14).

The source Proxy-SGW, then allocates the address and TEID to be used by the selected target Proxy-SGW for the uplink forwarding traffic over the indirect tunnel to be established between the source and target Proxy-SGWs (S-ProxySGW TEID for UL (2)).

Step 4a. [S-ProxySGW → S-LGW → SGW → MME] Create Session Response:
The source Proxy-SGW replays to the MME with a Create Session Response message including its allocated address and TEID (S-ProxySGW TEID for UL (2)).

Step 4b. [MME → SGW → T-LGW → T-ProxySGW] Create Tunnel Request:
Since there is no identified (direct) interface between the source and target Proxy-SGWs, the MME sends a Create Tunnel Request message (instead of Create Session Request) to the target Proxy-SGW in order to create the data forwarding tunnel between both Proxy-SGW functions (and thus between source and target LGWs). This message includes the source Proxy-SGW's address and TEID (S-ProxySGW TEID for UL (2)) created in Step 4 to be used for uplink traffic forwarding path (from target Proxy-SGW to source Proxy-SGW).

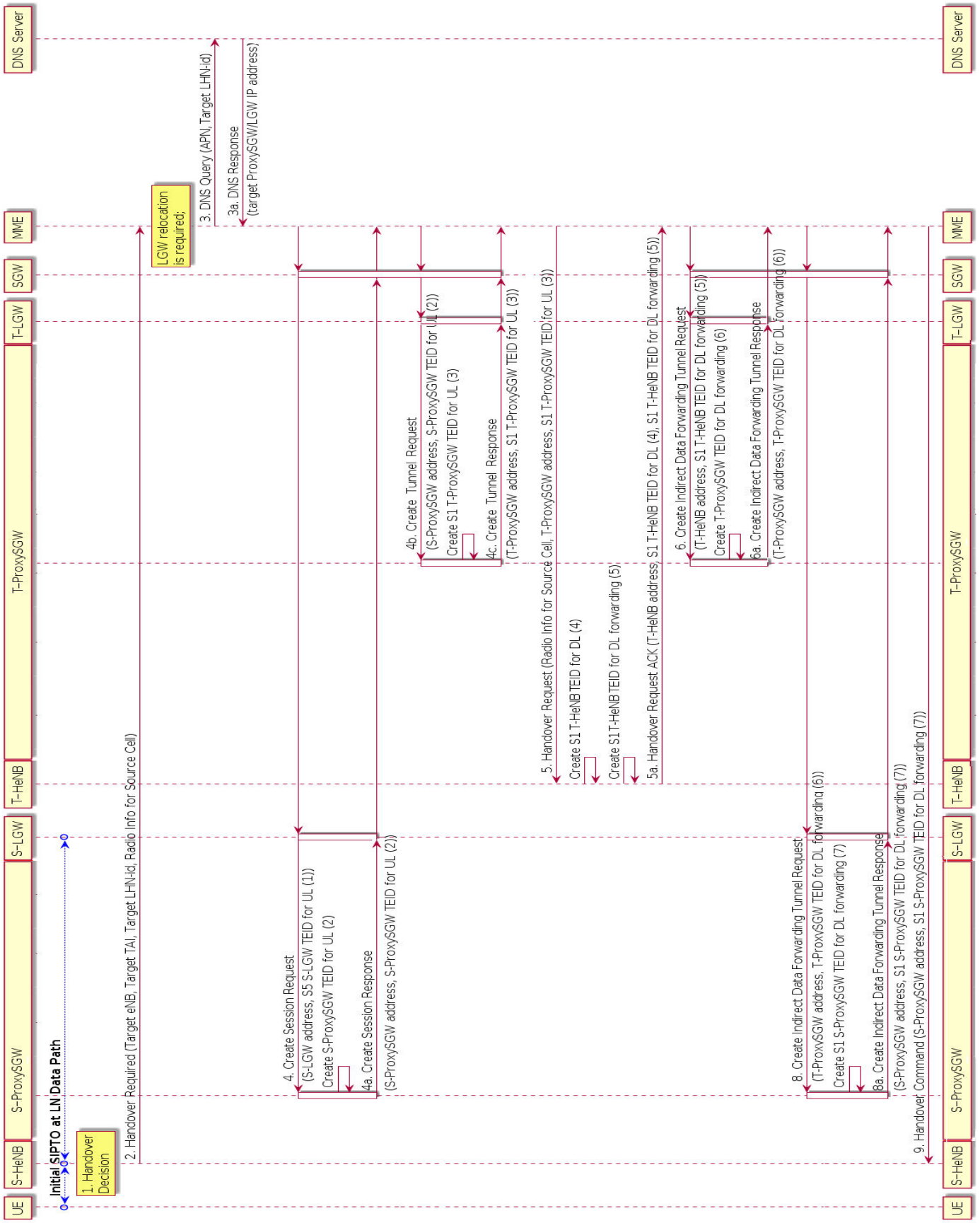


FIGURE 4.14: Handover preparation phase of smooth SIPTO for MC3

Step 4c. [T-ProxySGW → T-LGW → SGW → MME] Create Tunnel Response:
The target Proxy-SGW, upon receiving the Create Tunnel Request message, creates an indirect tunnel connecting to the source Proxy-SGW and then allocates the (S1 T-ProxySGW TEID for UL (3)). It then forwards this information to the MME through a Create Tunnel Response message, so that the target HeNB can create an indirect tunnel connecting to the target Proxy-SGW.

Step 5. [MME → T-HeNB] Handover Request:
The MME sends a Handover Request message to the target HeNB on behalf of the source HeNB. The information received in this message is as follows:

- Radio Information for Source cell received from source HeNB in Step 2: used for packets re-ordering.
- T-ProxySGW address and S1 T-ProxySGW TEID for UL (3) created in Step 4c: used to create an indirect tunnels between the target co-located HeNB and Proxy-SGW for uplink forwarding on the direct S1-U interface between them.
- E-RAB to be setup : information of E-UTRAN Radio Access Bearer (E-RAB) stored at source HeNB.
- Security Context: includes information used to derive the "Security base keys", e.g., K_{eNB} used to secure user's data over the radio link.

The target HeNB then creates the UE context including security and tunnel information. Based on the E-RAB to be setup information, the target HeNB checks if the same QoS provided at the source HeNB is available on the target HeNB as well. If available, it establishes an uplink S1 Tunnel connecting to its co-located Proxy-SGW, using the T-ProxySGW address and S1 T-ProxySGW TEID for UL (3) received from the MME. Then it allocates the downlink T-HeNB addresses and TEIDs to be used by the target Proxy-SGW on the S1-U interface for both the new path (S1 T-HeNB TEID for DL (4)) and the forwarding path (S1 T-HeNB TEID for DL forwarding (5)).

Step 5a. [T-HeNB → MME] Handover Request ACK:
The target HeNB sends the MME all information prepared in Step 5 regarding the tunnels admitted to be created in a Handover Request Acknowledgement message. This message includes also information about security algorithms supported by the target HeNB.

Step 6. [MME → SGW → T-LGW → T-ProxySGW] Create Indirect Data Forwarding Tunnel Request:
The MME sends a Create Indirect Data Forwarding Tunnel Request message to the

target Proxy-SGW requesting the creation of an indirect tunnel for delivering downlink packets during the handover. This message includes the S1 T-HeNB address and TEID for DL forwarding (5) that the target HeNB has allocated in Step 5.

Upon receiving this message, the target Proxy-SGW creates an indirect tunnel connecting to the target HeNB. It then allocates the address and TEID (T-ProxySGW address, T-ProxySGW TEID for DL forwarding (6)) to be used by the source Proxy-SGW on the downlink direction for the indirect forwarding .

Step 6a. [T-ProxySGW → T-LGW → SGW → MME] Create Indirect Data Forwarding Tunnel Response:
The target Proxy-SGW now sends the MME the downlink forwarding information created on Step 6 (T-ProxySGW address, T-ProxySGW TEID for DL forwarding (6)).

Step 8. [MME → SGW → S-LGW → S-ProxySGW] Create Indirect Data Forwarding Tunnel Request:
The MME creates the indirect tunnel with the source Proxy-SGW using a Create Indirect Data Forwarding Tunnel Request message including the T-ProxySGW's address and TEID (6) for DL forwarding received in Step 6a above.

Upon receiving this message, the source Proxy-SGW now allocates the address and TEID to be used by the source HeNB to forward the user's downlink traffic on the uplink direction over the direct S1-U interface between them (S1 S-ProxySGW TEID for DL forwarding (7)).

Step 8a. [S-ProxySGW → S-LGW → SGW → MME] Create Indirect Data Forwarding Tunnel Response:
The source Proxy-SGW confirms the creation of the indirect forwarding tunnel and sends the MME the allocated address and TEID (S1 S-ProxySGW TEID for DL forwarding (7)) within a Create Indirect Data Forwarding Tunnel Response message.

Step 9. [MME → S-HeNB] Handover Command:
The MME sends the source HeNB a Handover Command message that includes the information received in Step 8a as well as the security algorithms information provided by the target HeNB in Step 5a. At this stage, the handover Preparation phase is completed.

Figure 4.15 illustrates the network status following the handover Preparation. The tunnels prepared to forward the current SIPTO data traffic are numbered by their order of creation. As shown in this Figure, the data traffic exchanged between the UE and the server keep being routed over the initial SIPTO data path, whereas the new UL and DL forwarding paths are now ready to proceed for the handover.

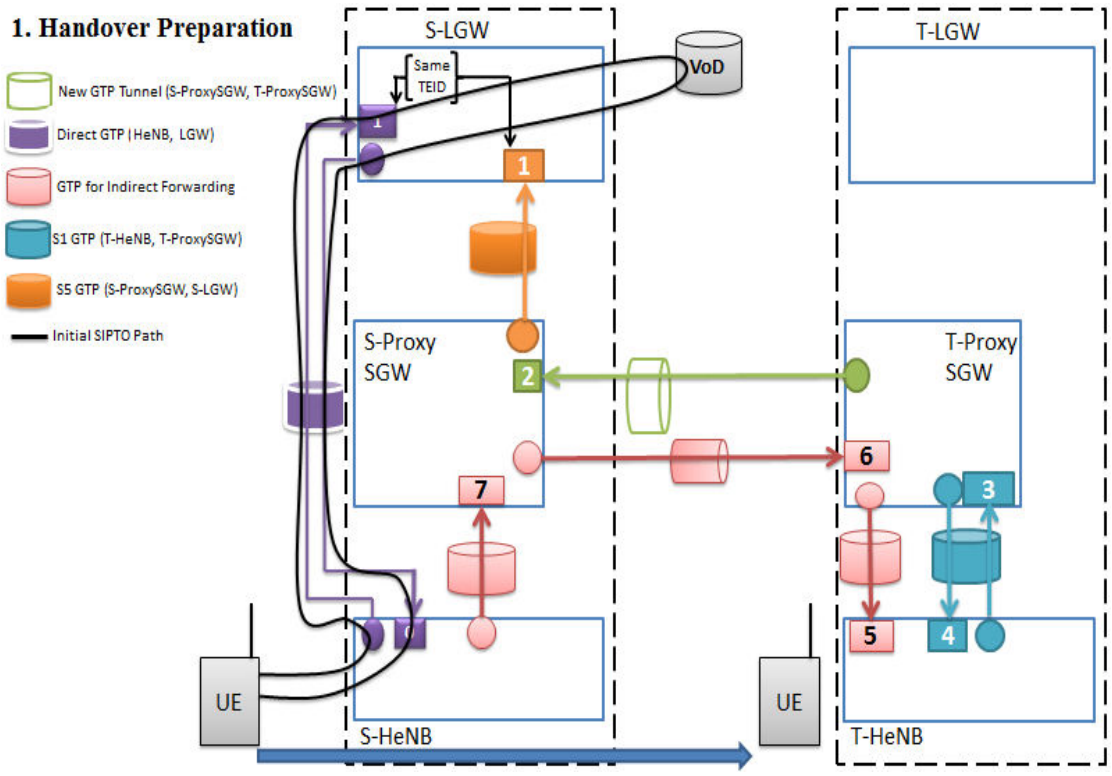


FIGURE 4.15: SIPTO Data Paths during the handover Preparation phase of smooth SIPTO for MC3

The handover Preparation phase of smooth SIPTO for MC3 differs from the S1-based handover Preparation phase of Classical LTE in:

- Step 3 where the initial LGW has to be relocated to fulfill Requirement 1 and a DNS Query has to be performed to fulfill Requirement 8.
- Step 4 where an indirect tunnel is established between the source Proxy-SGW and the target Proxy-SGW in order to fulfill Requirement 9.

4.2.2.3 Handover Execution

Step 9a. [S-HeNB → UE] Handover Command:
The handover Execution phase of smooth SIPTO for MC3 starts when the source HeNB sends a Handover Command message to the UE requesting it to detach itself from its associated cell (source HeNB) and synchronise to the selected target cell (tatget HeNB).

Steps 10 to 10c. [S-HeNB → MME] then [MME → T-HeNB] HeNB status transfer:
The source HeNB transfers its status to the target HeNB over the S1-MME interface between both HeNBs and the MME using a Status Transfer message as shown in Figure

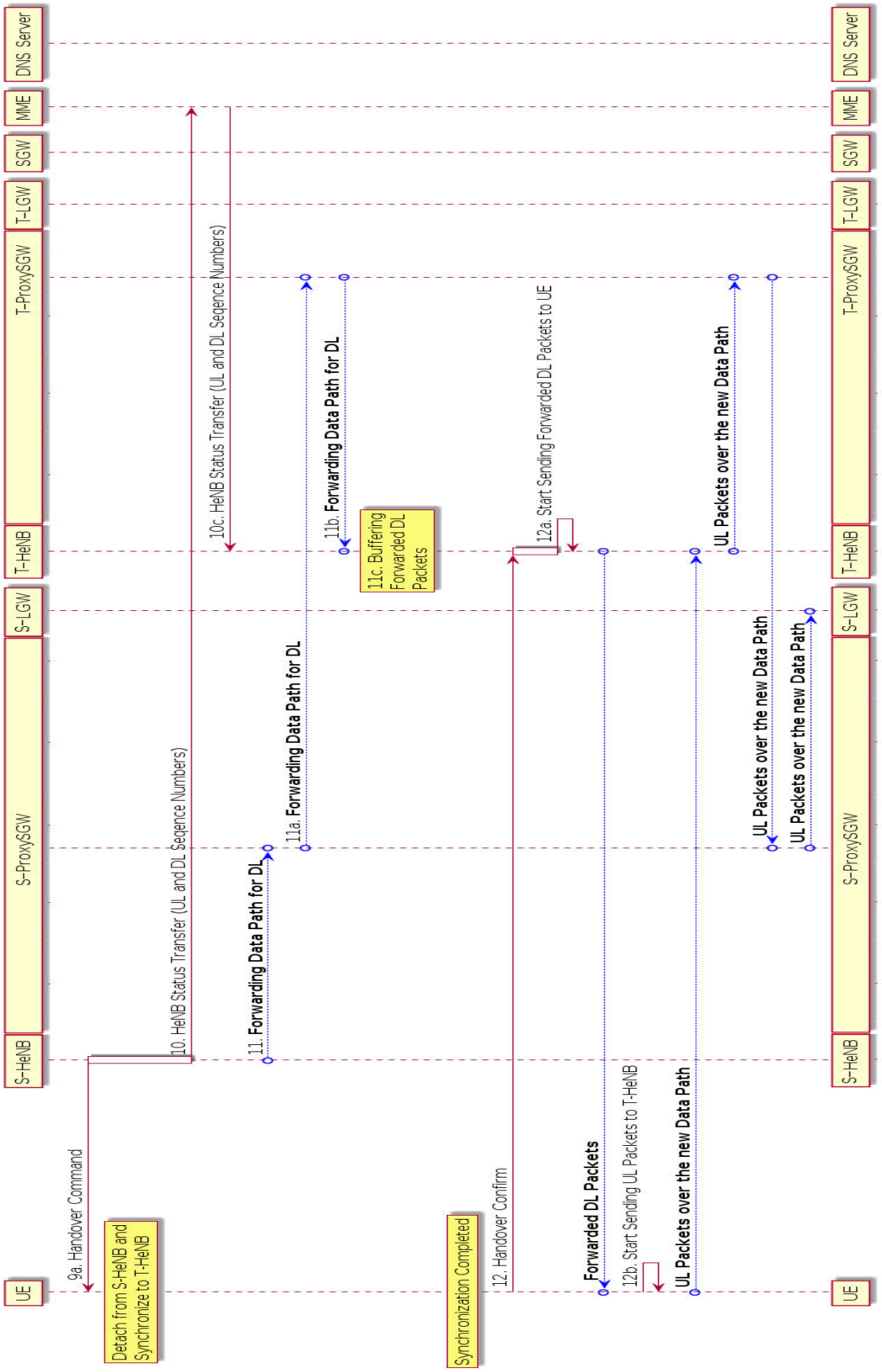


FIGURE 4.16: Handover Execution phase of smooth SIPTO for MC3

4.16. This message includes the uplink and downlink sequence numbers used to reorder the received packets and to prevent unnecessary packets re-transmissions.

Steps 11 to 11c. Forwarding DL Data Path:
At this stage, the downlink traffic forwarding between both the source and target HeNBs over the indirect SIPTO forwarding tunnel (via source and target Proxy-SGWs) begins. The target HeNB then start buffering the received packets until the UE's synchronisation is completed.

Steps 12 to 12b. [UE → T-HeNB] Handover Confirm:
The UE sends the target HeNB a Handover Confirm message to inform it that the synchronization to it is completed. At this stage the target HeNB starts sending the forwarded DL traffic to the UE. On its side, the UE also starts sending the UL traffic to the target HeNB over the radio interface between them. Both uplink and downlink forwarding paths used during the handover Execution phase are illustrated in Figure 4.17.

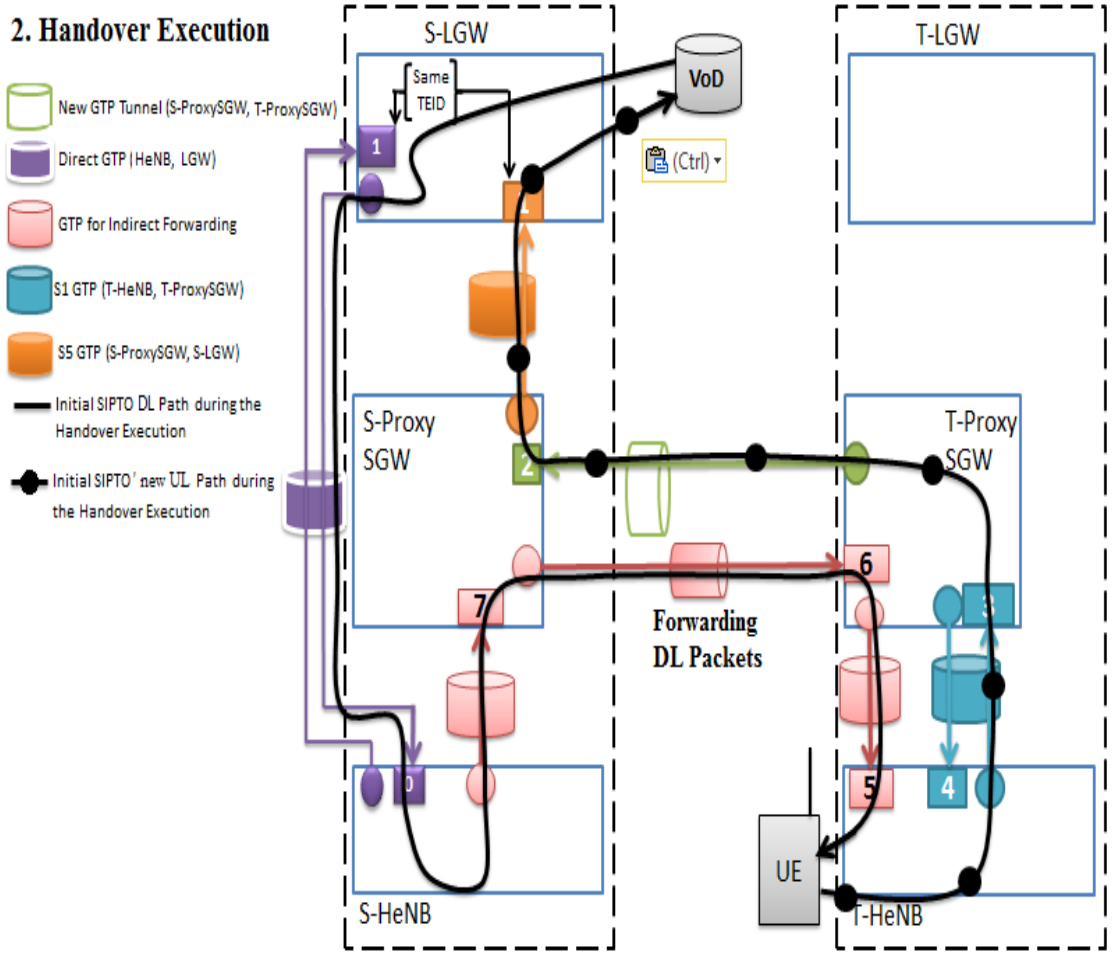


FIGURE 4.17: Network status during the handover execution of smooth SIPTO for MC3

Similar to smooth SIPTO for MC2, we note that the steps performed in the handover Execution phase of smooth SIPTO for MC3 are identical to those performed in the handover Execution phase of Classical LTE. The only difference is that, at the end of the handover Execution phase, the uplink data path of the current SIPTO session is routed indirectly to the source LGW (through the indirect forwarding tunnel between the source and target Proxy-SGWs).

4.2.2.4 Handover Completion

Step 13. [T-HeNB → MME] Handover Notify:
Target HeNB sends the MME a Handover Notify message informing it that the handover has successfully been executed. This message informs the MME that the handover Completion phase can now begin.

As pointed out in Section 4.1.3.1, the basic idea of smooth SIPTO for local mobility consists in handing over the initial SIPTO traffic, while establishing a new SIPTO data path. Note that in order to ensure that the initial SIPTO data path will not be broken before the new SIPTO data path is completely established, the standard handover process must be modified by increasing the delay of the Resource Release Timer at the MME's level. The additional delay should be larger than the delay required to establish the new local SIPTO connection. This is illustrated in Figure 4.18.

Once the Timer is launched, the MME establishes the new SIPTO connection in order to fulfill Requirement 2. A new local IP address is then assigned to the UE by the target LGW. Thanks to MPTCP features, the new IP address is communicated with the distant server and a new MPTCP subflow is created for the "new SIPTO path" using MP_JOIN option (see Figure4.13).

At this stage, the UE will have at least three available data paths to the server: the default (backup) data path towards the EPC, the indirect forwarding data path of the initial SIPTO towards the source LGW, and the new SIPTO data path towards the target LGW (see Figure 4.11-b).

Step 15. [MME → SGW → T-LGW → T-ProxySGW] Modify Bearer Request: The MME sends a Modify Bearer Request message to the target Proxy-SGW requesting it to prepare the network for switching the downlink path of the initial SIPTO.This message includes the T-HeNB address and S1 T-HeNB TEID for DL (4) allocated by the target HeNB in Step 5.

Upon receiving this message, the target Proxy-SGW establishes a downlink S1 tunnel connecting to the target HeNB, as requested. The target Proxy-SGW then allocates

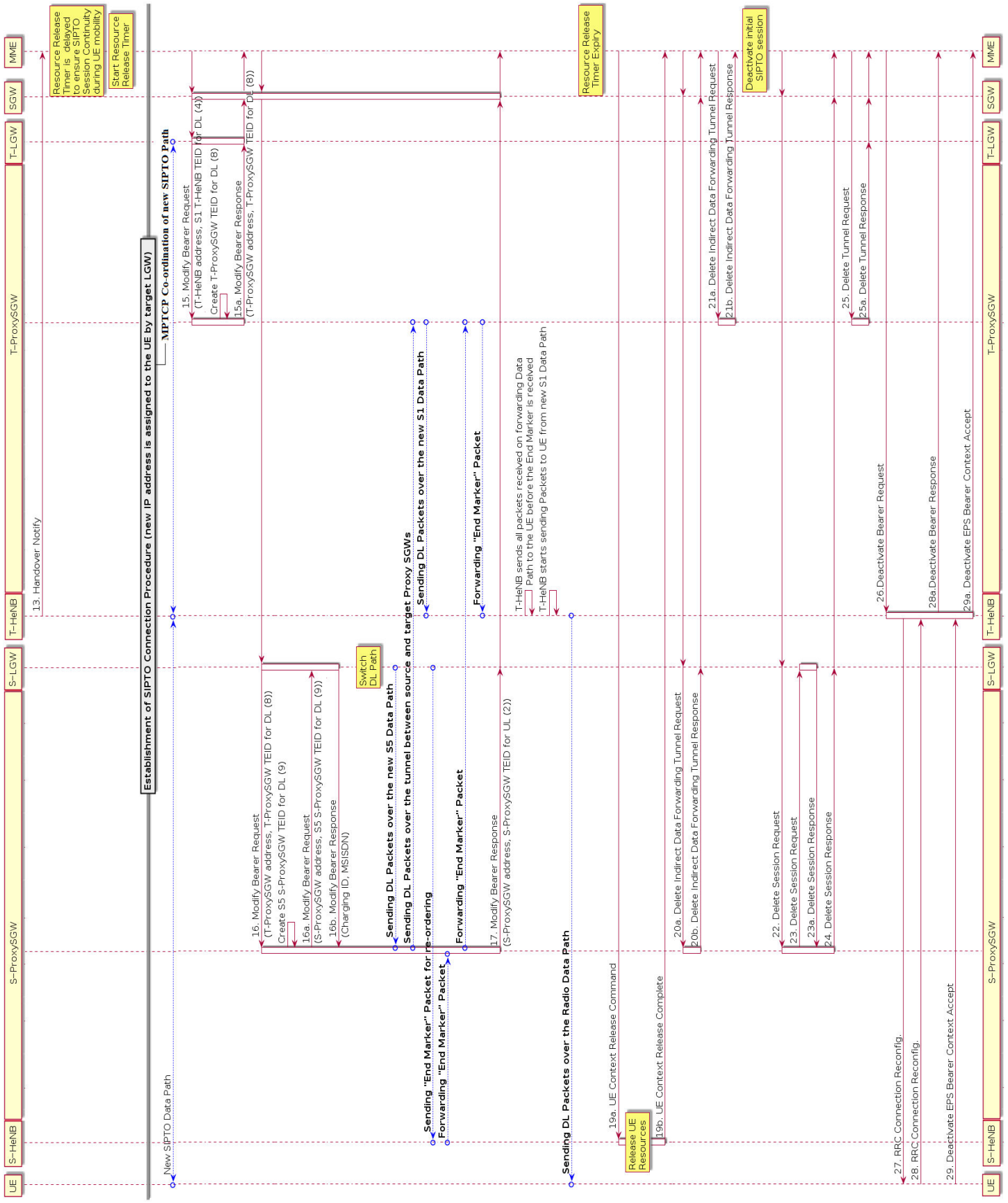


FIGURE 4.18: Handover Completion phase of smooth SIPTO for MC3

an IP address and TEID for downlink traffic coming from the source Proxy-SGW (T-ProxySGW address, T-ProxySGW TEID for DL (8)).

Step 15a. [T-ProxySGW → T-LGW → SGW → MME] Modify Bearer Response:
The target Proxy-SGW forwards the target Proxy-SGW's address and TEID for DL (8) allocated in Step 15 to the MME in a Modify Bearer Response message.

Step 16. [MME → SGW → S-LGW → S-ProxySGW] Modify Bearer Request:
The MME forwards the tunnels information received in Step 15a to the source Proxy-SGW in a Modify Bearer Request message. The source Proxy-SGW uses the received information to establish an indirect downlink path connecting to the target Proxy-SGW. It then allocates the downlink source ProxySGW address and TEID (S-ProxySGW address and S5 TEID for DL (9)) that would be used to modify the initial SIPTO downlink Path at the source LGW level.

Step 16a. [S-ProxySGW → S-LGW] Modify Bearer Request:
The source Proxy-SGW forwards the S-ProxySGW address and S5 TEID for DL (9) allocated in Step 16 to its co-located source LGW in a Modify Bearer Request message. This step aims to allow switching the downlink packet delivery from going from source LGW to source HeNB to go from source LGW to source Proxy-SGW.

Step 16b. [S-LGW → S-ProxySGW] Modify Bearer Response:
The source LGW acknowledges the bearer modification with a Modify Bearer Response message and switches the Path on the S5 interface to go towards the source Proxy-SGW. Simultaneously, the source LGW sends the target HeNB an "End Marker" packet over the indirect downlink forwarding path of the initial SIPTO connection. This packet allows coordinating the data at the target HeNB's level. Before receiving the End Marker Packet, the target HeNB continue to send all packets received on the forwarding data path to the UE. Once received, the target HeNB starts sending the UE packets coming from the new DL path of the initial SIPTO connection. Besides, the UE could also send/receive packets over the new SIPTO Connection as illustrated in Figure 4.19.

As shown in Figure , at this stage, the initial SIPTO traffic is routed over the new uplink and downlink paths. The source LGW first sends the new packets to its co-located source Proxy-SGW over the direct S5 interface between them. The source Proxy-SGW then, forwards the received packets to target Proxy-SGW, thanks to the indirect tunnel between them. The target Proxy-SGW finally forwards the traffic to its co-located target HeNB.

Step 17. [S-ProxySGW → S-LGW → SGW → MME] Modify Bearer Response:
The source Proxy-SGW confirms the switching by sending a Modify Bearer Response message to the MME.

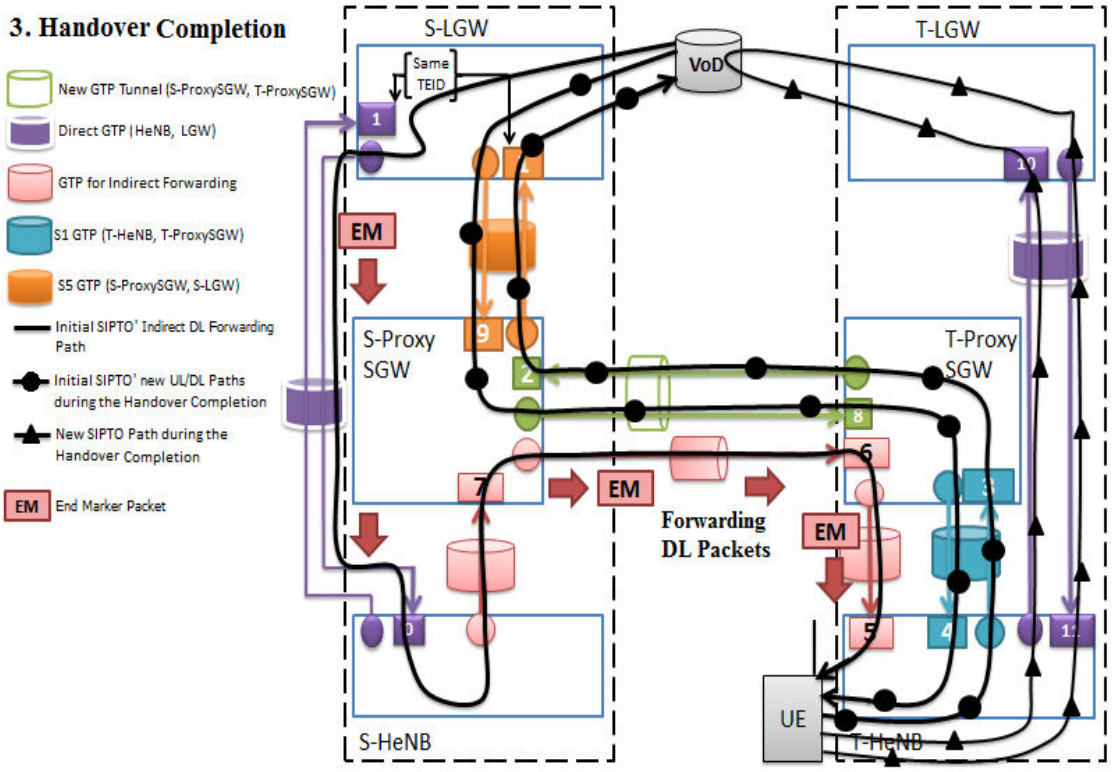


FIGURE 4.19: SIPTO Data Paths during of the handover completion phase of smooth SIPTO for MC3

Similar to smooth SIPTO for MC2, we note that in case the new (target) TAI that is sent to the user in Step 2 is not in the list of TAIs that the UE is registered with to the network, the UE initiates a TAU procedure presented by 3GPP in [4] and shown in Appendix A.

Steps 19a to 19b. UE Context Release:
Similar to 3GPP standards for S1-based handover, at the Expiry of the Resource Release Timer the MME commands the source HeNB to release the user's allocated resources as they are no longer in use. The source HeNB confirms the user's resource releasing to the MME with a UE Context Release Complete message.

Since the source LGW is still used to maintain the initial SIPTO connection after the handover, Steps 19c and 19d performed in 3GPP standards to release the initial path resources at the source SGW would not be performed during the handover procedure of our smooth SIPTO solution for MC3.

Steps 20a to 21b. Delete Indirect Data Forwarding Tunnels:
The MME deletes the indirect forwarding tunnels created in Steps 6 to 8a during the handover Preparation phase.

Steps 22 to 28b. Delete initial SIPTO Connection:
We now delete the initial SIPTO connection. First, MPTCP features are used to delete the MPTCP sub-flow originally set up for the initial SIPTO path. Then, the MME deactivates the initial SIPTO connection.

Unlike smooth SIPTO enabled handover for MC2 scenario, the connection deactivation of the initial SIPTO for MC3 differs from the PDN disconnection procedure illustrated in Figure A.16. In this solution, the MME first deletes the tunnel resources at the source co-located Proxy-SGW/LGW (steps 22 to 24). Then, it deletes the resources at the target Proxy-SGW with the Delete Tunnel Request/Response messages shown in Steps 24a to 24b.

Finally, Steps 25 to 28b are performed to deactivate the initial SIPTO connection.

At the end of the Handover Completion phase, the only path that should be used by the user is the new SIPTO path established between the co-located target HeNB and the target LGW (see Figure 4.20).

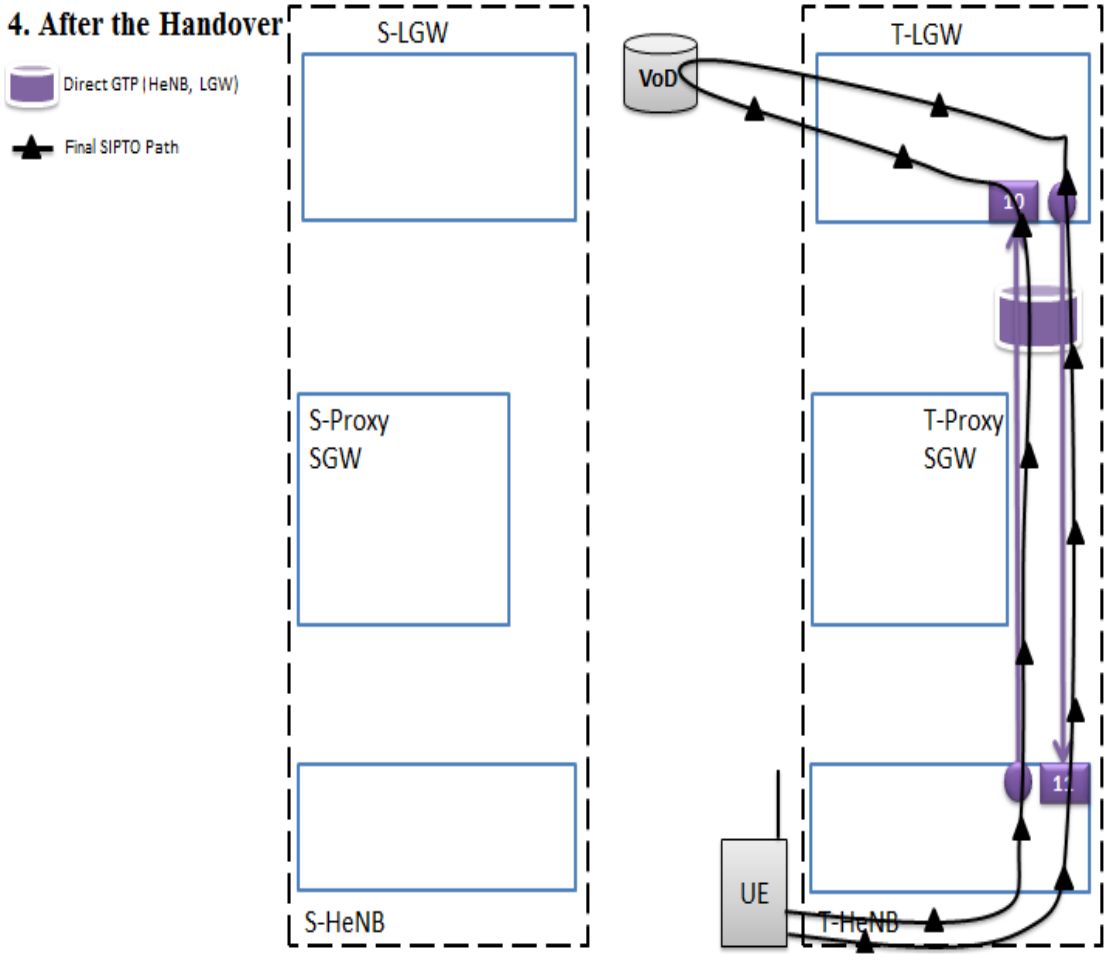


FIGURE 4.20: Final SIPTO Path at the end of smooth SIPTO handover procedure for MC3

Compared with the Completion phase of Classical LTE S1-based handover, our smooth SIPTO solution for MC3 proposes the use of MPTCP features to ensure the local mobility support. The main difference between the two Completion phases appears when we delay the resource release timer and establishes the new SIPTO Connection. Moreover, since there exists no direct interface between the target HeNB and the source LGW, and that the indirect forwarding tunnel ensures both the UL and DL paths of the initial SIPTO, the downlink path switching function of the initial SIPTO was then modified to ensure that the forwarding is always ensured thanks to the indirect tunnel of the initial SIPTO connection but the traffic is directly sent to the source Proxy-SGW instead of source HeNB.

Besides, steps 22 to 28b were introduced to the completion phase of our smooth SIPTO proposal for MC2 in order to delete the initial SIPTO path.

4.3 Qualitative Analysis of the Smooth SIPTO Handover Procedures

Generally, the new SIPTO Connection establishment procedure and the initial SIPTO Connection deactivation procedure are performed similarly to the standard 3GPP connection establishment/deactivation procedures. Therefore, the smooth SIPTO proposals for MC2 and MC3 sessions are fully compatible with 3GPP standards.

The above procedures allow a mobile UE to maintain the ongoing communication with the original server, i.e., the server with which the UE started the communication in the first place. However, this may not be the optimal server corresponding to the UE's new location (this is mainly considered for SIPTO above RAN sessions as illustrated in Figure 4.2-c).

On the other hand, a server relocation procedure would have broken the ongoing communication as the server would have changed, which is not allowed by MPTCP.

4.4 Global Picture of the Proposed Solutions

In this section we recap the global picture to our proposed smooth SIPTO solutions. As shown in Figure 4.21, when a UE connects to its application, a new session is established for this user by the network. In particular, a mobile session is carried over a TCP connection.

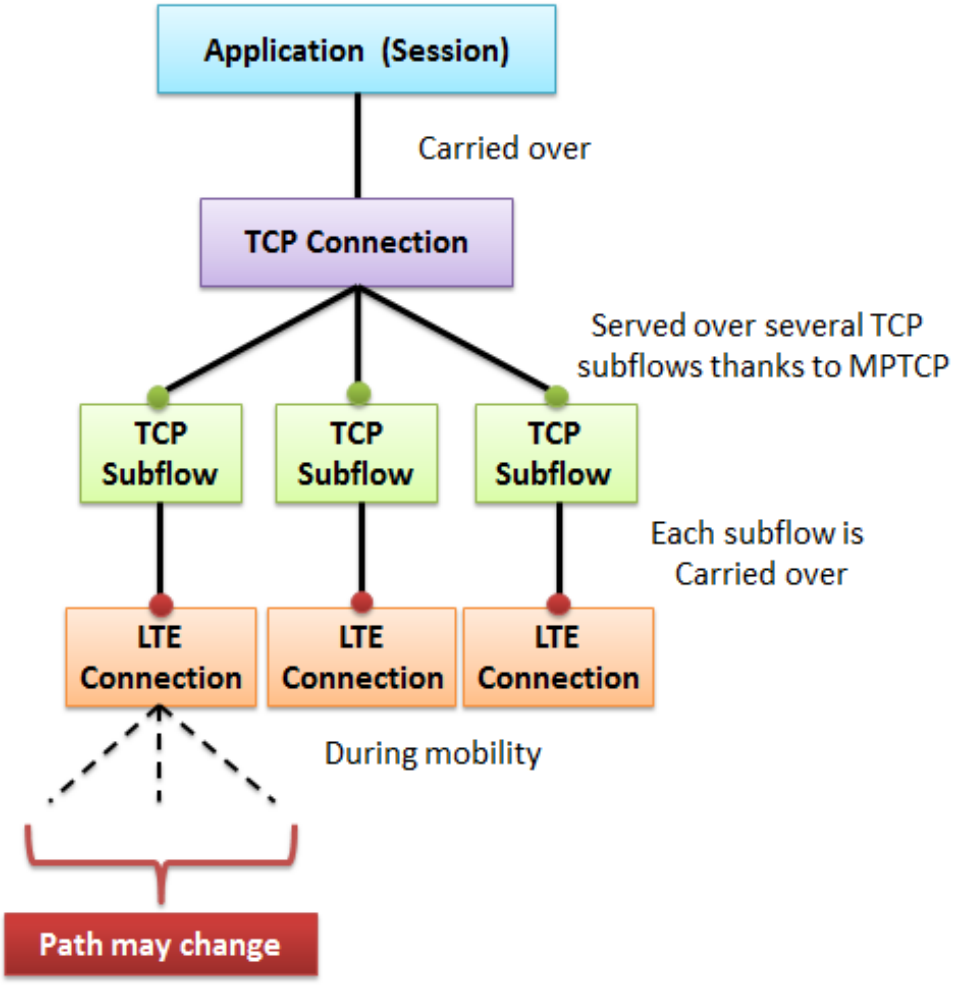


FIGURE 4.21: Global Picture of smooth SIPTO Proposals

Thanks to MPTCP, a TCP connection can be served over multiple TCP subflows simultaneously. Each subflow is carried over a single LTE Connection. During the user's mobility, thanks to 3GPP handover procedures, path may change.

4.5 Conclusion of Chapter 4

In this Chapter, we have addressed the session continuity issue in SIPTO above RAN and SIPTO at LN reference models presented in Chapter 2. We have proposed to blend the 3GPP handover procedures with the use of MPTCP to prevent traffic interruption due to PGW relocation.

Moreover, for SIPTO at LN, we proposed to enhance the LGW with a Proxy SGW function in order to overcome the fact that, at a given time, a UE can only be associated with a single SGW. This enhancement allows to go beyond the 3GPP study in [45] by using the LGW to offload both long-lived and short-lived data sessions.

The MPTCP protocol is used to maintain the ongoing session even after the previous IP address becomes unreachable and UE acquires a new address. We have described how to maintain MPTCP signalling for creating new sub-flows and discarding old ones over the backup path, thus relying on the legacy LTE mobile architecture.

The next Chapter presents the future FMC architecture proposed by COMBO Project during which this thesis is achieved. Potential mapping scenarios of smooth SIPTO proposals on future COMBO FMC network topology are then proposed.

Chapter 5

Smooth SIPTO as Part of the COMBO FMC Architecture

In this Chapter, we first present the motivations behind the Work Package three of COMBO project for functional convergence and the implementation models for future fixed and mobile converged network. Section 5.2 assess how Classical LTE architecture presented in Chapter 2 and smooth SIPTO architectures proposed in Chapter 4 could be mapped on uDPM functional blocks. Section 5.3 presents the mapping proposals of smooth SIPTO architectures on COMBO FMC network topology. Finally, Section 5.4 concludes the Chapter.

5.1 The COMBO Project

This Section summarizes the deliverables D3.2 [12] and D3.5 [13] of the COMBO Project. As pointed out in Section 2.4 of Chapter 2, the FP7 European collaborative project "COMBO" aimed to achieve a joint optimization of fixed and mobile networks by proposing a new broadband access/aggregation network architecture organized around the innovative concept of NG-POP, which is targeting two different aspects of network convergence: Functional Convergence and Structural Convergence.

The work done in this thesis is part of the COMBO Work Package three (WP3) for functional convergence. The objectives of WP3 "Fixed Mobile architectures Converged Architectures" is to propose, define and technically assess candidate architectures for FMC networks, both in terms of data plane and control plane. In that aim, COMBO project has defined two key intermediate goals to be achieved in WP3. The first one addresses the convergence of authentication and subscriber data management, whereas

the second one aims at providing the network operators with a dynamic control of data traffic on multiple available data paths (e.g., via fixed network or WiFi access), while maintaining session continuity whenever necessary.

To fulfil the requirements listed above and reach a true FMC network with global network control, COMBO project has proposed and developed in [12] a universal Subscriber and User AUThentication (uAUT) and a universal Data Path Management (uDPM). The main objective of providing uAUT for FMC architecture is to consolidate the "Authentication", "Authorization" and "Accounting" functions used in session establishment for Fixed, Wi-Fi or Mobile networks, i.e., FMC users should authenticate ONCE and have access to multiple existing networks and/or services. In particular, uAUT is linked to legacy subscriber data bases and plays the role of an intermediate server, which only receives the authentication requests of users and then forwards them to the adequate real servers (HSS for mobile, AAA for fixed and Wi-Fi). On the other hand, uDPM aims to provide the future FMC network with a dynamic control of mobile data paths when multiple routing interfaces are available (e.g., via Fixed, WiFi and Mobile networks), all with respect to session continuity support. Indeed, the deployment of uDPM aims to achieve seamless mobility between LTE and Wi-Fi technologies, allow mobile users to stream video traffic seamlessly thanks to LGWs, and ensure user's mobility when an IP edge gateway relocation is required by the network (e.g., LGW or PGW is relocated). More particularly, the work achieved in this thesis and represented in Chapter 4, in which we introduce a seamless mobility solution for users with ongoing SIPTO sessions (with IP edge gateway relocation), represents an integral part of the uDPM proposal of COMBO Project.

The following sub-sections present the several functional blocks involved in the deployment of uDPM and how uDPM and uAUT functional blocks should be implemented in COMBO architecture for future FMC network.

5.1.1 uDPM Design and Deployment Strategies

The uDPM should allow breaking out part of mobile data traffic from going through the LTE default data path to go over available Wi-Fi or fixed data Paths (e.g., using a LGW that is co-located with a fixed RGW, defined in Section 2.1.1 of Chapter 2). One of the main issues that could be faced by a FMC network operator during the deployment of such a solution is the maintain of session continuity support. uDPM also aims to allow multiple data paths to be simultaneously used for a given session. FMC network operators could benefit from such an implementation by being able to perform load balancing, avoid congestion and reduce costs and energy consumption.

The uDPM functional blocks could be triggered either by the network or by the UE. The uDPM would be triggered by any "session event". A session event can be represented by any action that would trigger or modify an activity of a particular user (e.g., a mobile device submit its measurement report to the base station, to which it is connected to). Typically, a session event is generated within the network either by monitoring functions or by subscriber's profiles and network rules.

- **Monitoring:** Overall, Telecom operators strength their networks with specific monitoring actions (e.g., scheduled signalling messages for requesting/sending information from/to the network). This allows operators to observe their network/user's behaviours and performances and to have a better control of their networks. Typical examples for network monitoring could be defined when the network or the UE measures a degradation of the received signal strength or when UE detects a new available Wi-Fi access point.
- **Subscriber Profile and Network Policies:** The set of rules as well as subscribers profiles are stored in a repository space named "Subscriber Profile and Network Policies". The application of such rules could trigger a session event (e.g., the network offload residential LTE traffic using Wi-Fi access points during business hours).

uDPM includes a **"Decision Engine"**, which hosts intelligence of the data path management. The Decision Engine is part of the control plan. The main task to be achieved for a decision engine is to decide how a session would be mapped on different available data paths. Indeed, the Decision Engine is responsible for: access selection, mobility management and content repository selection. These decisions rely on monitoring information, on user or subscriber's profile, on network policies and on UE configuration.

Indeed, a UE (i.e. mobile smartphone) may provide the user with the option of choosing the access interface to be used for a specific application at any time (e.g., activate the Wi-Fi and deactivate the 3G/4G and vice versa). A user could also configure its UE with the option to activate SIPTO or not, possibly dynamically. The main idea is to allow the user to choose again on which PDN connection he would like to transfer its packets. For example, SIPTO could be used particularly to offload "video traffic". As a result, in this case the uDPM Decision Engine would be distributed between the UE for interface selection, and the network (for example the MME as described later in Section 5.2).

As shown in Figure 5.1 for uDPM concept, there are three slave functional blocks depending on the Decision Engine:

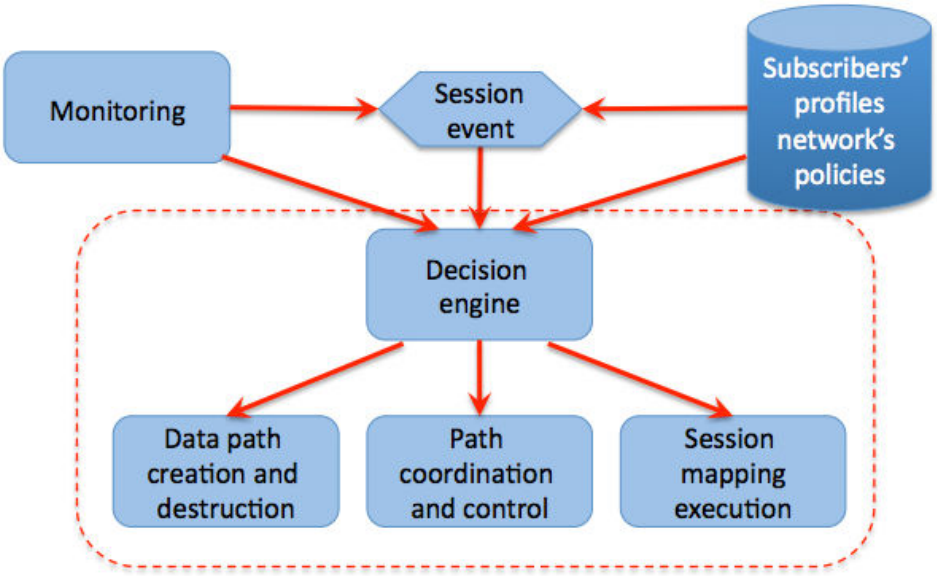


FIGURE 5.1: uDPM Functional Blocks [12]

1. **Data Path Creation and Destruction:** This functional block handles the control of data path creation and destruction on available interfaces. A given user can access its services through if he/she subscribes to different networks. For example, a mobile user could potentially receive traffic for a specific session over : default LTE data path, LIPA/SIPTO data paths [6] or even via Wi-Fi technology.

2. **Path Coordination and Control:** Path Coordination and Control is part of control plane. This functional block aims to ensure that concurrent data paths will smoothly deliver the packets corresponding to a given session over the different available interfaces, while maintaining session continuity especially during user's mobility.

3. **Session Mapping Execution:** unlike the rest of uDPM functional blocks, session mapping execution is a data plane function. Mainly, session mapping execution applies the decisions taken by the data path creation and destruction block described above.

5.1.2 How and Where Should Convergence Functions be Implemented?

To realize uAUT and uDPM functional blocks, COMBO has introduced the notion of Universal Access Gateway (UAG) [13] [9]. UAG is a functional entity in which both uAUT and uDPM functions will be implemented.

As any equipment in the network, UAG consists of Control Plane (CP) and Data Plane (DP)(or user plane in 3GPP terminology [74]). The UAG CP plays a significant role in

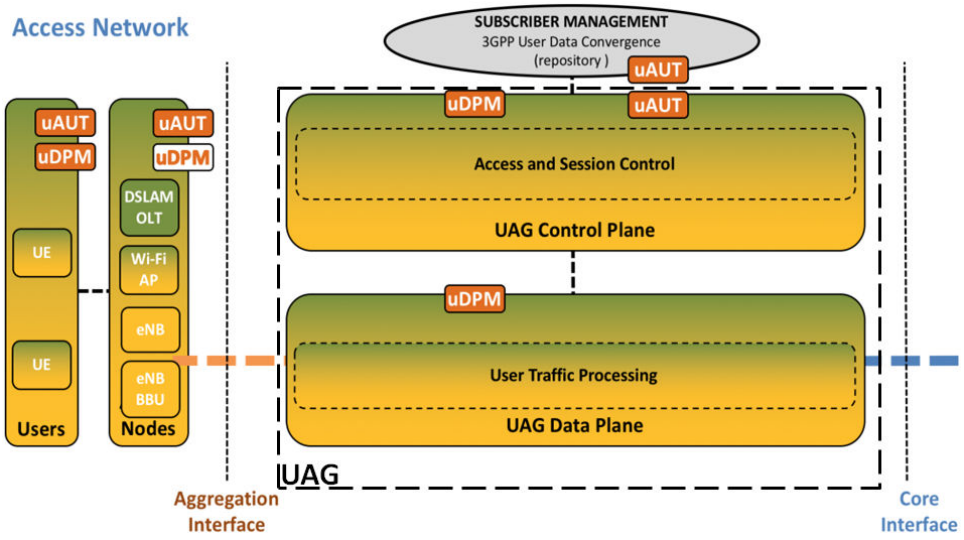


FIGURE 5.2: Splitting uAUT and uDPM in various locations [13]

the implementation of uAUT and uDPM functional blocks. This is due to the fact that most of the uAUT and uDPM functional blocks defined above are control functions (e.g., intelligence of interface selection and route control for uDPM and user's authentication and subscriber's data management for uAUT). As pointed out in Section 2.2.1 of Chapter 2, the MME function represents the mobile node, which is responsible of session and connection management in mobile networks. As a result, in the framework of COMBO FMC network, the MME function could be integrated within the UAG CP.

The UAG DP on the other side, it particularly enforces the uDPM functions. The main characteristic of UAG DP is the fact that it represents a common IP edge for fixed, mobile and Wi-Fi access networks; i.e., UAG DP should provide FMC users with an access to the external IP network regardless of their access network.

Besides, as both uAUT and uDPM rely on the UE collaborating with various network level entities (e.g., access nodes, subscriber data management), then, along with the UAG implementation, uAUT and uDPM functions should also be partly implemented within the UE and these network entities as well. Figure 5.2 depicts the high level functional view of the unified access gateway with split of uAUT and uDPM functions.

In this figure, we note that the proposed COMBO architecture for FMC network separates the UAG's DP from CP. This would allow FMC operators to consider different implementation options for each plane (e.g., CP in commodity servers and DP in specific hardware equipment). Moreover, such implementation would also improve the deployment flexibility of both plans. Particularly in COMBO, two deployment models have been considered: Standalone and Split UAG (CP/DP).

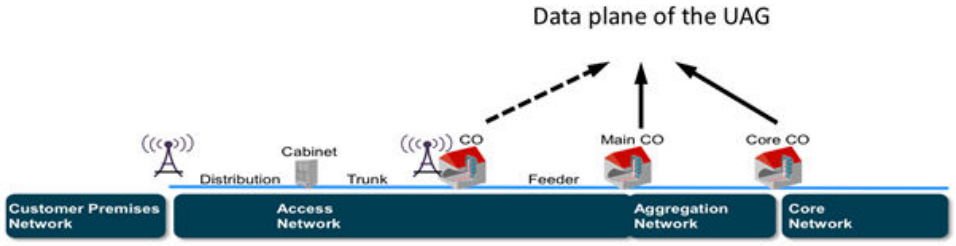


FIGURE 5.3: Reference locations for locating the data plane of the UAG [13]

In **standalone** model, both CP and DP functions have to be located in the same equipment. For instance, this could be achieved by integrating the current fixed, WiFi and mobile functional entities into a single node. As a result, the UAG can be considered as a kind of structural fixed–mobile converged subscriber IP edge.

In **split** model on the other side, CP and DP functions should be implemented in different equipment, and possibly in various topological or geographical locations in the network. However, in this case, a part of the UAG CP functions must remain in the same location as the UAG DP entity (at least, to manage the interface with the UAG CP). Mainly, the split model is designed to keep the maximum amount of control functions high in the network while distributing UAG DP functions close to the user.

In both models, the subscriber's common IP edge (UAG DP) should be located at the border of the access/aggregation and IP core network. The potential locations for the UAG DP deployment are shown in Figure 5.3. As shown in this figure, three locations were considered:

1. at the current CO, which represents the network operator building where traditional telco copper cables and fiber are terminated. In particular, COs host the DSLAM or the OLT and potentially a macrocell base stations (eNB). The distance between a CO and a UE is typically represented by few tens of meters (less than 200 meter).

2. at the Main CO, which represents a special CO with higher aggregation level than standard COs and not connected directly to the network's core. Main COs could be located far from the UE at a distance ranging that goes from 50 to 75 Km. Typically, Main COs aggregate passive infrastructure routes (e.g., fibre cable routes) of multiple COs.

3. at Core CO, which represents the network operator building that ends the current fixed aggregation network and connects it to the core network. Core COs are typically regional locations, which implies that they could be distant at approximately 300 Km from the UEs. It is in fact, the current location of the Broadband Network

Gateway (BNG), which is the IP edge for fixed and Wi-Fi traffic (Section 2.1 of Chapter 2). The IP edge for mobile traffic (PGW), on the other side, could be located even higher in the network than the Core CO (typically, at the IP Core).

Thanks to the deployment of fibre-based access networks, which allow increasing the distance between the customer premises and the CO without degrading fixed access performance (Chapter 2), it is now possible to reduce the number of COs, i.e., to move the CO up to the Main CO [13]. As a result, COMBO has chosen to deploy UAG DP only at the Main CO and Core CO levels of the network.

Regarding the control functions, the standalone and split UAG models lead to different degrees of flexibility for implementation and location of the UAG CP, which can be either co-located together with DP at the Main CO or at the Core CO or even located higher in the network (at the IP backbone). The multiple deployment options for UAG CP are discussed in more details in Section 5.2.

5.1.3 COMBO Scenarios for FMC Network Architecture

To realize the FMC network, network operators in [9] and [10] have decided to alleviate the load of the different segments of the network by deploying distributed IP edges and services closer to the user's location within a FMC network. The main idea is to co-locate the IP edges of different access networks that were integrated within a universal access gateway data plane with application servers and data centers within a Next Generation Point of Presence. As pointed out in Section 2.4 of Chapter 2, the NG-PoP is a location in the network, in which the operator could implement multiple services and functions (including the UAG DP). Considering the locations selected above in where a UAG DP could be deployed, NG-PoP could then be located in the network either at the Main CO or at the Core CO. The location of the NG-PoP would depend mostly on the population density of the region; e.g., NG-PoP is placed within the Core-CO for densely populated areas and within the Main-CO for rural or less densely populated areas. Based on the NG-POP locations and on the two potential deployment models (Standalone UAG and Split UAG) considered above, COMBO project analysed and compared potential alternative architectures for functional convergence in 5G networks, whereas two main FMC architectures have been identified, namely Centralized NG-POP architecture and Distributed NG-POP architecture.

5.1.3.1 Centralized NG-POP Architecture

The Centralized NG-POP architecture relies on a small number of NG-POP locations (particularly, at the Core COs level of the future FMC network). Compared with the current fixed and mobile network architectures, centralized NG-POP architecture enables a better use of network resources, a simplified deployment and an easy operation [13]. Figure 5.4 illustrates the potential deployment scenarios, which have been considered for the centralized COMBO architecture: Standalone UAG, Split UAG with co-located CP and DP and Split UAG with DP and CP distant from each other.

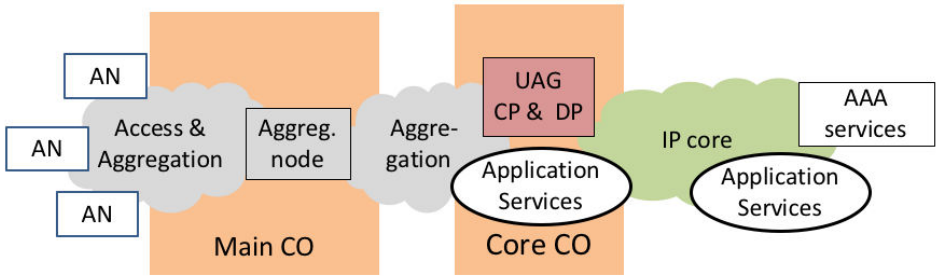
Globally, the centralization of UAG DP functions at the Core CO increases the flexibility and scalability of the DP and alleviates the load of the core network and thus reduces the potential congestion risks expected to occur at the core level of the network. Moreover, the centralization of UAG DP minimizes the frequent changes of mobility anchors for a given user (e.g., SGW/PGW for mobile users), and thus reduces the potential additional handover signalling costs. On the other hand, the centralization of the UAG CP alleviates the load of the control entities and allows a better thus minimizes the potential scalability issues resulting from mobility control (a large number of users/devices per MME in a typical 5G scenario may generate a burden in terms of control traffic, e.g. paging).

As a result, in the centralized NG-POP scenario, as far as mobility is concerned, it is thus recommended to locate both UAG CP and DP functions at Core COs. The deployment of the UAG DP at Core COs, does not require extension of the IP network. However, it increases the DP latencies.

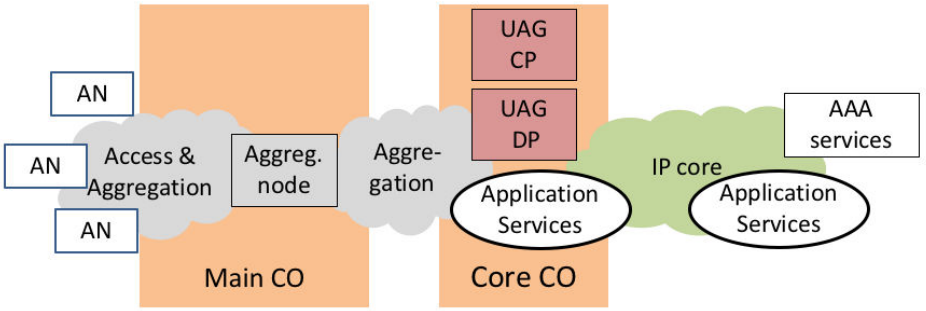
5.1.3.2 Distributed NG-POP Architecture

The Distributed NG-POP architecture relies on a large number of NG-POP locations, located mainly in the Main COs, leading to an extension of the IP backbone towards the access network. Such an extension (which may be considered as an IP aggregation network) can strongly impact routing management, and thus require reviewing the IP architecture. Moreover, this architecture requires more logical instances (with the same network functionalities) running in more physical network nodes inside multiple NG-POP locations.

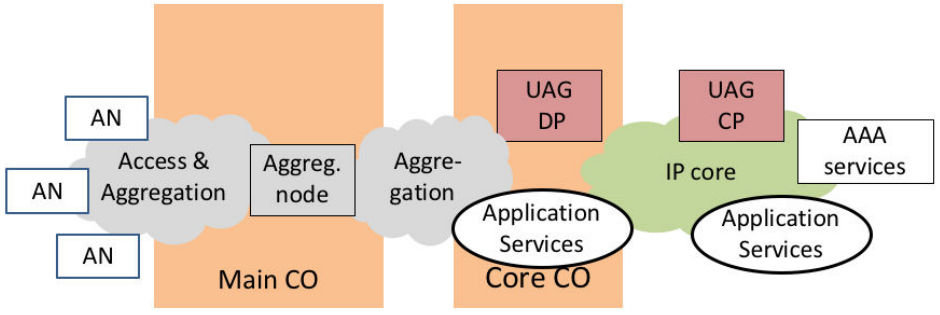
Figure 5.5 illustrates the different deployment scenarios for the distributed COMBO architecture. As shown in this figure, the UAG DP is located at Main COs, closer to the users/devices. On the other hand, the UAG CP is located either at the Main CO or at the Core CO.



Standalone UAG in COMBO centralised architecture



Split UAG with co-located DP and CP in COMBO centralised architecture



Split UAG with CP distant from DP in COMBO centralised architecture

FIGURE 5.4: UAG deployment with UAG DP at Core CO (centralized COMBO architecture) [13]

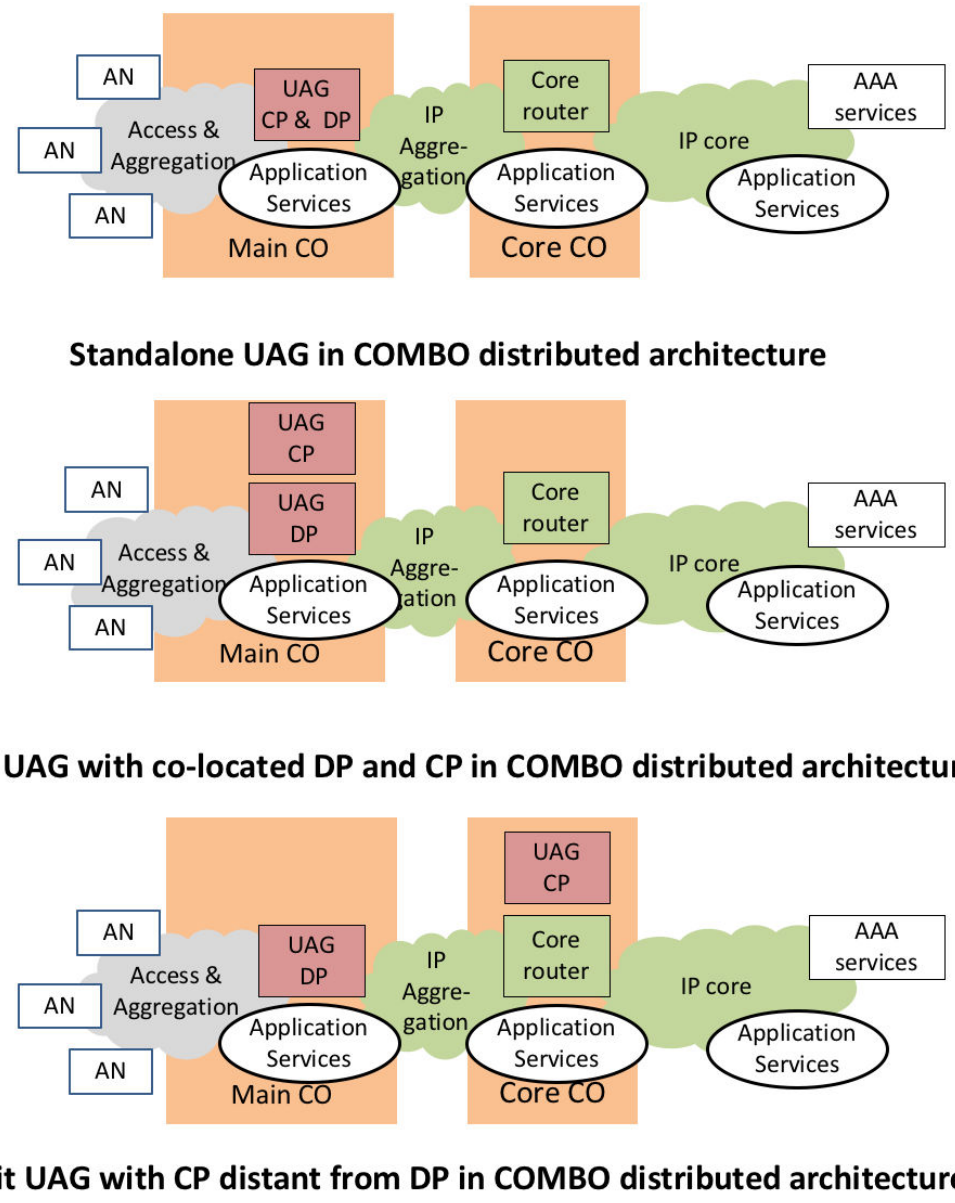


FIGURE 5.5: UAG deployment with UAG DP at Main CO (distributed COMBO architecture) [13]

Compared with Centralized NG-POP architecture, having mobility anchors in the UAG DP at Main COs allows a lower DP latencies, an improved flexibility and scalability of the DP and even lesser congestion risks by decreasing data traffic load not only in the core network but also in the aggregation network as well. For instance, as mentioned in Chapter 3, by distributing mobile LTE architecture at the border of Access/Metro and Metro/Core segments of the network, almost 40% of core and metro network bandwidths could be offloaded in 2017. This is a significant gain that can justify the CAPEX required for distributing the LTE architecture. However, as mobility anchors will be more distributed at Main COs, this will result in more frequent changes of mobility

anchors.

Furthermore, having the UAG CP in the Main CO will generate lots of additional signalling messages between different Main COs or between a Main CO and central entities like the HSS (e.g., for context and session management during handover procedures). Such signalling would increase loads not only on the communication links but also on the processing units of the devices. For that reason, it is always recommended to have the UAG CP at Core CO level or even higher in the IP core network. This represents a compromise between larger scalability of UAG mobility control functions and lower signalling traffic related to changes of mobility anchors.

5.2 Mapping Mobile Network's Functional Entities on uDPM Functional Blocks

As pointed out in Chapter 2, SIPTO architectures are designed to allow a UE to access the external IP services closer to the user's location. Considering the candidate COMBO FMC architectures presented in Section 5.1, we note that having the IP edges of the mobile network co-located together within the UAG (at Main CO or at Core CO) corresponds perfectly to the 3GPP architecture for SIPTO above RAN. Moreover, having the LGWs co-located with the HGWs could also corresponds particularly to the 3GPP architecture for SIPTO at LN.

The present section presents how smooth SIPTO solutions could be mapped on uDPM Functional Blocks and how they could be implemented on candidate FMC network architectures. As pointed out in Section 5.1.1, to provide future FMC network operators with a dynamic control of multiple available data paths for mobile users, COMBO project proposed four functional blocks enabling a universal Data Path Management. Nevertheless, the implementation of these functional blocks on future FMC network topology was not provided by COMBO.

5.2.1 Mapping Classical LTE Architecture on uDPM

We first represent how classical LTE handover can be mapped on uDPM functional blocks. In particular, we look into the detailed **S1-based handover procedure** with SGW relocation for classical LTE architecture.

Let us consider a scenario where a UE is in a car while having an ongoing classical LTE session (e.g., the user is watching a video on Youtube). As pointed out in Section 2.3.1 of

Chapter 2, due to the user's mobility an "inter eNB" handover procedure with/without SGW relocation could be performed [4].

In the following points we details each of the handover phases as defined in [4].

1. **Before Handover Preparation**

   During the handover Preparation phase, the source eNB monitors the user's radio connection in a periodic manner by sending a Measurement Control message. The UE then measures the signal strength of neighbour cells and replays back to the source eNB with a Measurement Report message. This action is considered as part of the network's monitoring function.

   The source eNB then selects a target eNB and decide on which interface this handover should be performed. This could be achieved using the handover decision control function within the source eNB. In this handover we assume that the handover is performed on S1 interface. Once the handover decision is taken, the source eNB sends the MME a Handover Required message requesting it to begin the handover preparation towards the target eNB. In particular, the Handover Required message represents the session event triggering the execution of the uDPM functional blocks.

2. **Handover Preparation**

   The handover Preparation in Classical LTE is performed similarly to the handover Preparation phase of smooth SIPTO for MC2 presented in Section 4.1.3.2 of Chapter 4, except that in Classical LTE, no PGW relocation is considered.

   Figure 5.6 shows how uDPM functional blocks are executed during the handover Preparation phase of Classical LTE.

   According to 3GPP in [4], the MME represents the network element that hosts the intelligence of data path management and connection management functions, which makes it the potential 3GPP control function that would implement the Decision Engine of the uDPM for the future FMC network.

   Upon receiving the Handover Required message and based on the target TAI information included in the message, the MME decides whether an MME relocation should be performed or not.

   Considering the COMBO FMC architectures proposed in Section 5.1.3 where the MME function could be integrated within the UAG CP e.g., at the Main CO, at the Core CO or even centralized very high in the network (at the IP Core), then the possibility of changing the MME would vary depending on the MME's

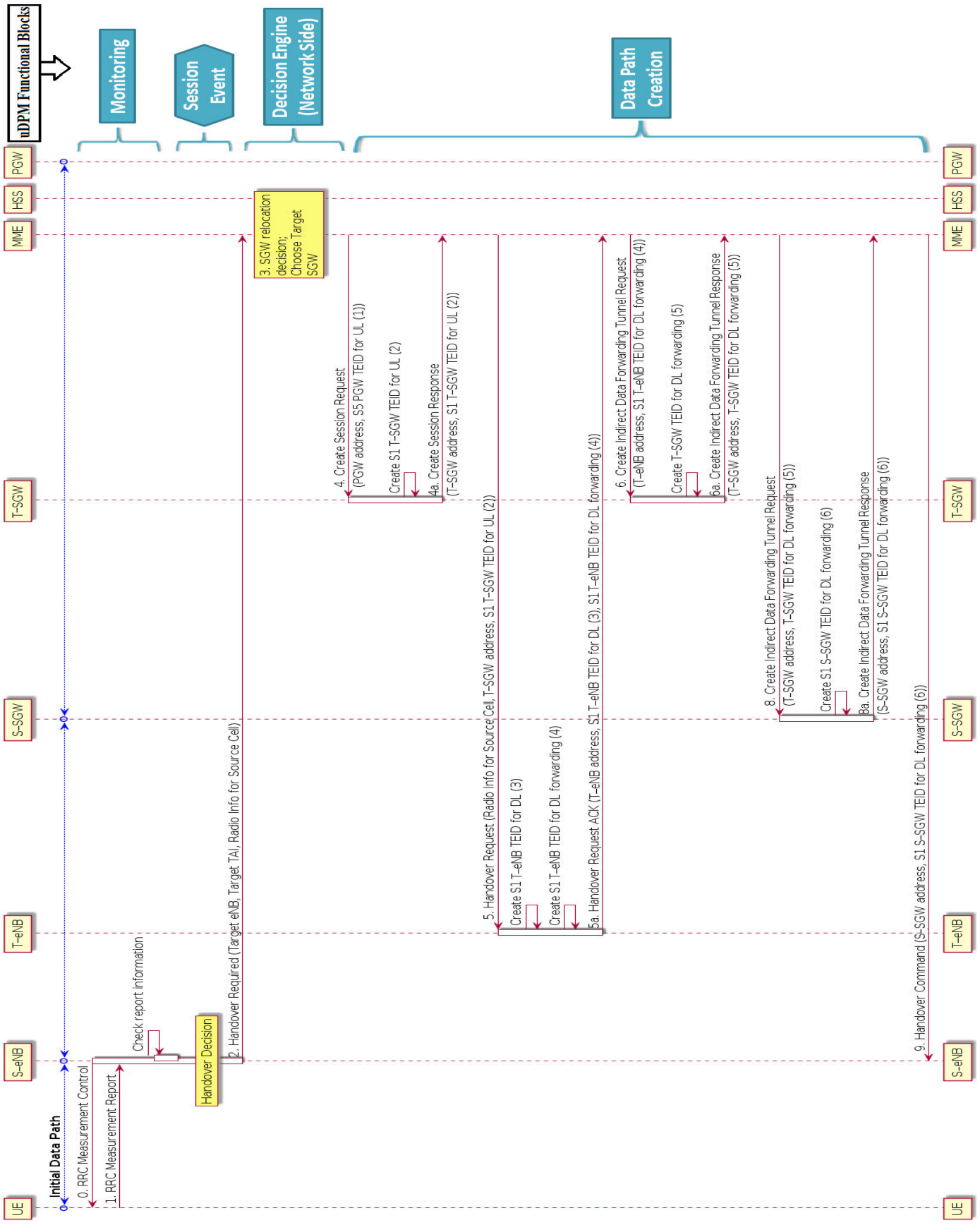


FIGURE 5.6: Mapping Classical LTE's handover Preparation phase on uDPM functional blocks

location. For instance, having the MME function within the UAG CP at the Main CO would lead to increase the potential relocation of the MME during user's mobility. This is due to the fact that having the MME very low in the network would reduce the number of users that could be managed by this MME. However, having the MME functions centralized high in the network (at the Core CO or at the IP Core) would result in maximizing the number of eNBs that could be connected to it and thus increasing the number of users managed by this MME. Consequently, the potential possibility of MME relocation in such scenario would be largely minimized, compared with the distributed scenario.

- In case that MME relocation is not required, the MME should use again the target TAI information to check whether a SGW relocation is required or not (we assume here that a SGW relocation is required). Following such decision, the MME must perform a SGW selection function to select the most adequate SGW to be used for this handover.

  Now that the target SGW is selected, the MME prepares the handover by ordering the initiation of the uDPM "Data Path Creation" functional block. This is done when the MME sends a Create Session Request to the target SGW requesting it to establish a new tunnel connecting to the PGW. Steps 4 to 9 presented in Figure 5.6 for Data Path Creation are then performed similarly to steps 4 to 9 presented in Section 4.1.3.2.

  We note that during the Preparation phase of S1-based handover, the MME forwards each of the created uplink and downlink TEIDs and IP addresses to their adequate network element. For instance, in step 4a the target SGW creates the TEIDs and the IP addresses to be used by the target eNB for uplink direction, the MME then transfers the received information to the target eNB in step 5 to ensure that the packets that could be sent by the target eNB to the target SGW arrives to the right destination and in the correct order. As a result, in this case, the MME represents the network element that applies the "Path Coordination and Control" functional block of uDPM. Mainly, the Path Coordination and Control function in mobile networks is ensured by the encapsulation (typically, GTP-U and IP protocols are used for mobile traffic encapsulation) and the packet-numbering during transport.

- In case the MME is relocated, the source MME should use the target TAI information to select a target MME. Then, the source MME is required to forward the user's context information to the selected target MME. This is represented in Step 3 of Figure A.14 shown in Appendix A. It should be noted here that the decision for SGW relocation would now be performed by target

MME instead of source MME. In this case, both source and target MMEs would proceed as Decision Engines to prepare the handover and initiate the uDPM "Data Path Creation" function. It is important to know that, upon an MME relocation, all signalling messages that are related to source eNB and/or source SGW are exchanged over S1 and S11 interfaces with the source MME, whereas the signalling messages related to target eNB and/or target SGW are exchange with the target MME.

For the sake of simplicity, we illustrates in Figure 5.6 the mobility case where no MME relocation is required.

3. **Handover Execution**

   Now that the two eNBs are ready to perform a handover, the handover Execution phase starts by detaching the UE from the source eNB and by synchronising it to the target eNB. As shown in Figure 5.7, and as pointed out in Section 4.1.3.3 of Chapter 4, the handover Execution phase of Classical LTE is performed similarly to handover Execution phase of the smooth SIPTO for MC2.

   During the handover Execution phase, the source eNB starts forwarding the downlink data traffic to the target eNB through the indirect data forwarding tunnel. Whereas, the target eNB, on its side, buffers all the received traffic until the user becomes completely synchronized with it. Once the UE is completely synchronized to the target eNB, uplink and downlink packets can be exchanged over the radio interface between them. Moreover, the new uplink path will finally be enabled.

   The uDPM's Session Execution functional block should be implemented within each data plane functional entity involved in the handover procedure for both the downlink forwarding path (source eNB $\rightarrow$ source SGW $\rightarrow$ target SGW $\rightarrow$ target eNB) and the new uplink path (target eNB $\rightarrow$ target SGW $\rightarrow$ PGW).

4. **Handover Completion**

   Now that the data sessions for both the downlink forwarding path and the new uplink path are in use, it is important to prepare the network for switching the downlink data traffic routing from going through the indirect forwarding path to go through the new downlink path (PGW $\rightarrow$ target SGW $\rightarrow$ target eNB). To begin this switching, the target eNB should inform the MME of the session execution status with a Handover Notify message. Figure 5.9 illustrates the mapping of Classical LTE Completion phase on uDPM.

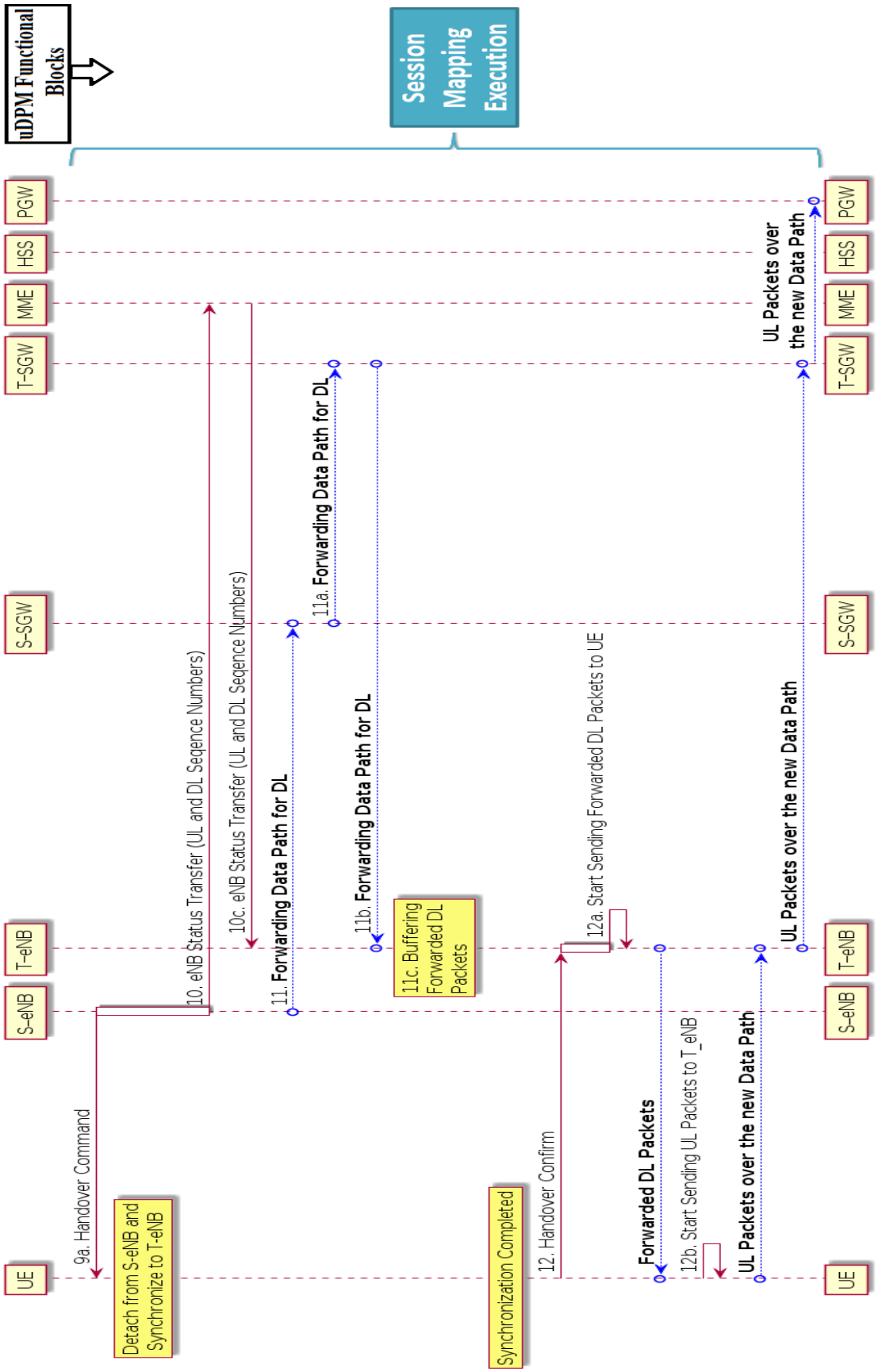


FIGURE 5.7: Mapping Classical LTE's handover Execution phase on uDPM functional blocks

As shown in this figure, at the reception of the Handover Notify message, the MME starts the Resource Release Timer and then proceed for the path switching. In that aim, Steps 15 to 21b for LTE Handover Completion are performed similarly to Steps 15 to 21b presented previously in Section 4.1.3.4 of Chapter 4 for smooth SIPTO handover of MC2. First, the MME sends a Modify Bearer Request message to the target SGW. Then, the target SGW creates on its side the downlink S5 TEIDs to be used by the PGW for the new downlink path and forwards it to the PGW within a Modify Bearer Request message. Here the target SGW represents the network element that ensures the uDPM path coordination and control functional block for the PGW functions on the downlink direction.

Now that the PGW and the target SGW are aware of the new downlink path information, the PGW will proceed for switching the downlink Paths. First, the PGW acknowledges the switching to the target SGW with Modify Bearer Response message and then it starts sending the downlink packets to the target SGW on the new downlink path. Finally, the target SGW replies to the MME with a Modify Bearer Response message and sends the downlink packets received from the PGW over the S5 interface to the target eNB over the S1 interface between them.

Simultaneously, the PGW sends an "End Marker packet" to the target eNB via the indirect data forwarding tunnel. The End Marker packet ensures the coordination of packets (coming simultaneously from the indirect data forwarding path and from the new downlink path) at the target eNB's level. As a result, the uDPM Path Coordination and Control functional block is ensured by the PGW function (thanks to the End Marker packet) and executed at the target eNB's level. Figure 5.8 illustrates how the downlink paths received by the target eNB are coordinated and sent to the UE.

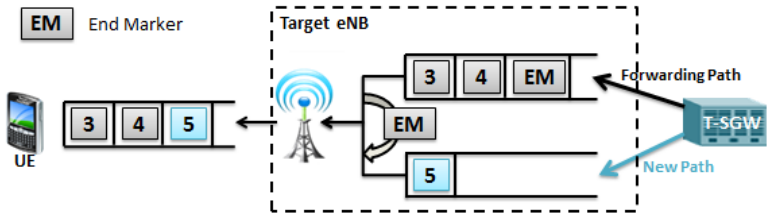


FIGURE 5.8: Path Coordination and Control ensured by the End Marker Packet

Indeed, before receiving the End Marker packet, the target eNB sends all received packets to the UE on the indirect data forwarding path. Once received, the target eNB starts sending packets to the UE that are received from the new downlink path. The packets transmission corresponds to the uDPM' Session Mapping Execution.

Finally, the MME checks the resource release timer expiry and proceed for the destruction of the data paths used during the handover procedure for the indirect forwarding path and the initial SIPTO path. Moreover, the MME deletes the user's context and releases the UE resources of the source eNB.

5.2.2 Mapping Smooth SIPTO Architecture on uDPM

We now define how the smooth SIPTO handover solutions proposed in Chapter 4 are mapped on uDPM functional blocks. As pointed out in Chapter 3 and Chapter 4, the main idea of implementing the smooth SIPTO handover solutions is to allow the network to optimize the risks of potential losses in terms of bandwidth and latency, while ensuring an always-on mobile connectivity.

Compared with the Classical LTE architecture, both proposed smooth SIPTO handover solutions relay on the use of MPTCP features to maintain the ongoing SIPTO session during users mobility. Considering the sequence diagrams illustrated in Figures 4.3 and 4.13 of Chapter 4, we note that the use of MPTCP features appears only during the Completion phase of smooth SIPTO handover solutions. Whereas the other phases (Preparation and Execution) remain unchanged. Figure 5.10 illustrates how smooth SIPTO for MC2 could be mapped on uDPM functional blocks. The same mapping applies to smooth SIPTO for MC3.

As shown in this Figure, the handover Completion phase of smooth SIPTO for MC2 starts by establishing a new SIPTO Connection during which the new SIPTO data path is created.

During the establishment procedure of the new SIPTO Connection, the UE initiates a session event by sending PDN Connectivity Request message to the target eNB, which forwards it to the MME. As illustrated in Figure A.1 in Appendix A, at the reception this session event, the MME, which represents the Decision Engine on the network side, checks the user's profile and the network rules in order to decide whether this PDN connection could be established for this user or not. Then, if the establishment decision is favourable, the MME selects the IP edges (target co-located SGW/PGW) to be used for this new connection and proceeds for the data path creation by sending a Create Session Request message to the target SGW. At the end of the establishment of the new SIPTO connection, the UE will have three different IP addresses: the IP address assigned by the default PGW for MPTCP and backup use, the IP address assigned by the source PGW for initial SIPTO session and the new IP address assigned by the target PGW for the new SIPTO connection.

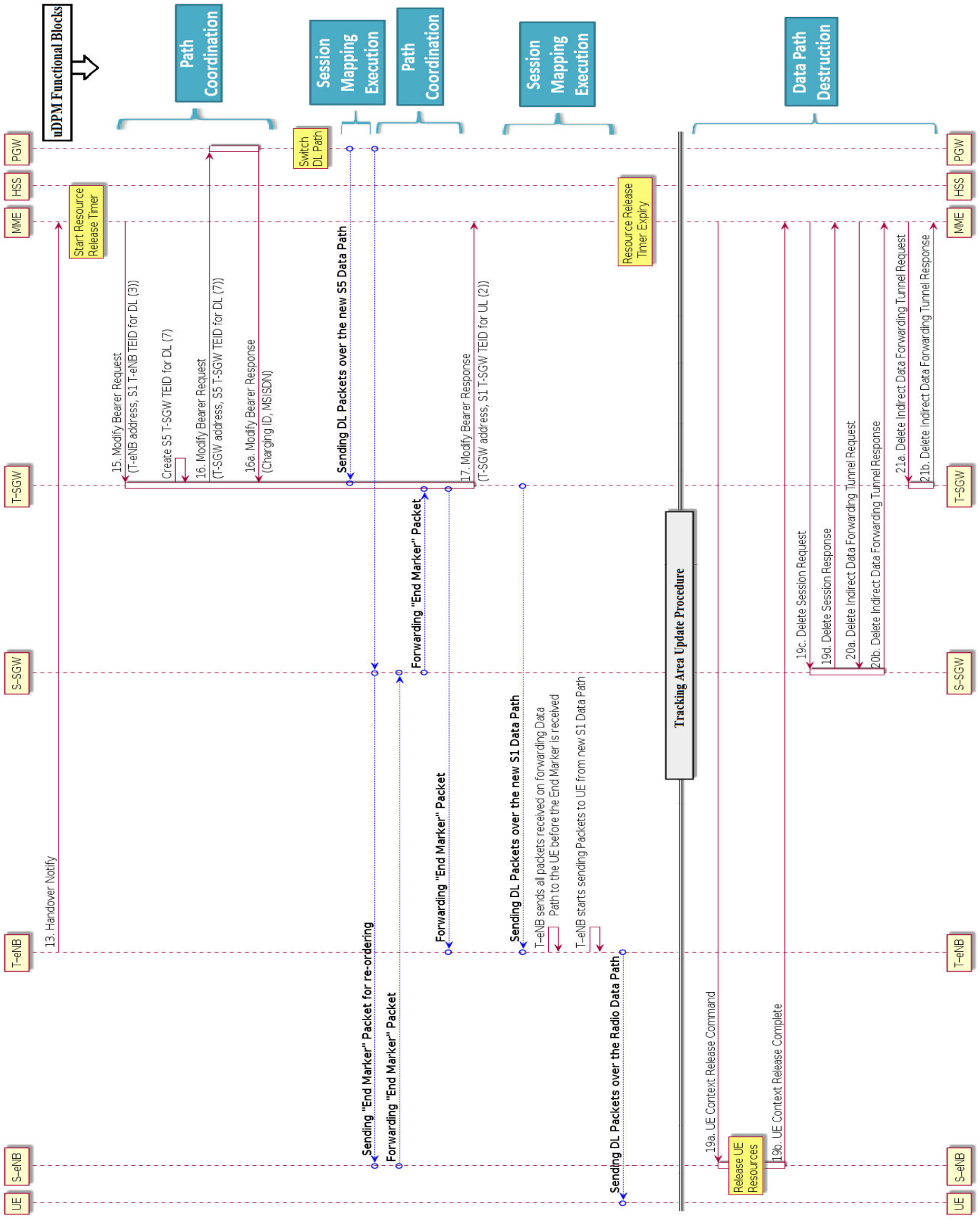


FIGURE 5.9: Mapping uDPM functional blocks on a Classical LTE's handover Completion phase

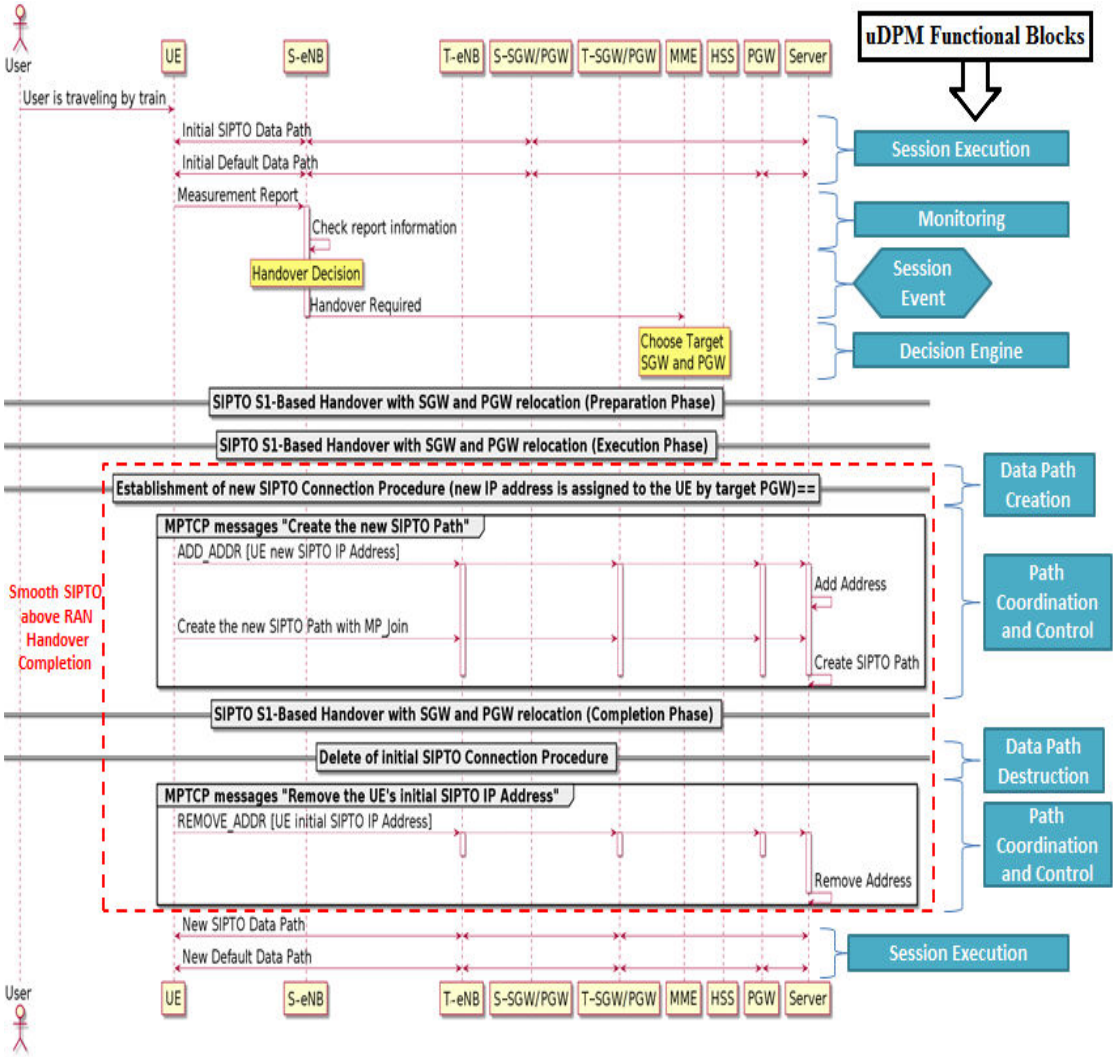


FIGURE 5.10: Mapping uDPM on the smooth SIPTO handover solutions

Now that the new SIPTO connection is established, the UE should communicates its new IP address with the remote host (e.g., VoD server). As shown in Figure 5.10, this would be achieved using the MP_Join option of the MPTCP features. MPTCP ensures simultaneously the co-ordination and the control of the multiple available SIPTO paths. As a result, in future FMC framework where smooth SIPTO is implemented, we can infer that Path Coordination and Control functional block of uDPM will be ensured, thanks to the MPTCP, by the network elements involved in data paths used for the MPTCP signalling, including the mobile host (the UE), the target SGW, the default PGW and the remote host (the server).

Basically, thanks to the mobile network's encapsulation (IP and GTP-U), mobility in Classical LTE is managed by the network directly at the IP layer. Whereas, in the framework of smooth SIPTO, where MPTCP multihoming features are used, mobility is managed by the network at the Transport layer instead of IP layer. This is due

to the fact that MPTCP enables any host to use multiple available network interfaces simultaneously for a single TCP session.

Following the creation of the new SIPTO connection, the MME performs the handover Completion phase similarly to Classical LTE. However, at the end of the handover Completion phase, since the PGW is relocated in smooth SIPTO, we propose that the MME proceeds for the destruction of the initial SIPTO data path by sending a Delete Session Request message to the target SGW. At the end of the handover, all resources related to the initial SIPTO connection would be deleted. The user then, uses the MPTCP features to delete the IP address assigned to it by the source PGW.

Considering the smooth SIPTO solution for MC3, uDPM functional blocks are mapped similarly to smooth SIPTO for MC2. Nevertheless, it should be noted here that, in smooth SIPTO for MC3 the regular SGW is not relocated. Instead it is the Proxy SGW, which is co-located with the source LGW, that would be relocated. According to 3GPP in [4], when the SGW is not relocates, it is the SGW that sends the End Marker Packet to the target eNB during the handover Completion phase. As a result, in case of smooth SIPTO enabled handover for MC3, the uDPM path coordination during the handover Completion phase is then ensured by the SGW function instead of PGW (thanks to the End Marker packet) and executed at the target eNB's level.

5.3 MPTCP Signalling Support

As pointed out in Section 5.1, the main idea for future FMC network is to integrate the IP edges of different access networks within a UAG on the one hand and to co-locate them with application servers and data centres within a NG-PoP on the other hand. This corresponds to distributing the EPC in many locations [9, 10], whereas the EPC is currently centralized in the legacy mobile network architecture [75].

As shown in Section 5.1.3, none of the FMC architectures proposed in COMBO has considered the implementation of a centralized PDN gateway. Whereas, in our proposed smooth SIPTO solutions, we considered that the use of the default PGW (PGW within the EPC) is mandatory to ensure session continuity during the users mobility.

In the following, we first introduce the notion of "Anchor PGW" and highlight its role in supporting MPTCP signalling in a COMBO architecture. Then, we define how smooth SIPTO could be implemented when either the UE, or the server is Non-MPTCP Capable.

5.3.1 Selecting an Anchor PGW in a Distributed EPC

In a FMC network topology, having PGWs functions within UAGs very low in the network (i.e., at the access network or beyond the RAN) makes it difficult to ensure session continuity during users' mobility as it yields potential loss of users' IP address. Smooth SIPTO procedures described in Chapter 4 have been designed to deal with this issue.

As shown in Chapter 4, smooth SIPTO relies on a "default PGW" which is used to carry MPTCP signalling, thus ensuring session continuity for SIPTO connections. It is necessary to ensure that the connection to this "default PGW" can indeed be maintained, even during UE's mobility, which is not an issue in a centralized EPC architecture, but may become one in a distributed EPC architecture.

We thus propose to introduce the notion of an **Anchor PGW**, on which the UE would be constantly connected to support MPTCP signalling used in the smooth SIPTO mechanisms.

Regarding smooth SIPTO at LN, it is possible to locate the Anchor PGW within the UAG; the connection to the server build with the UE's IP address provided by the Anchor PGW is used for MPTCP signalling whereas another connection, build with the UE's IP address provided by the LGW is used for MPTCP data flows.

However, smooth SIPTO above RAN may not be as easily supported. At the attachment of a UE in a distributed EPC architecture, the MME selects a "default PGW" ; this PGW is associated with the "NO-PGW Selection" during handover procedure [4]. In this particular case, the Anchor PGW is the "default PGW". If the MME has selected the "default PGW" close to the UE's initial location (e.g., within the UAG), it may also select the same PGW for its SIPTO above RAN sessions.

Assuming this is the case, an MPTCP connection will be established over the path to the Anchor PGW in order to ensure the MPTCP signalling and a smooth SIPTO connection will also be established towards the same PGW. This means that the UE uses the same IP address for both MPTCP signalling and smooth SIPTO paths.

The issue to be dealt with is that the path used within the mobile architecture for MPTCP signalling should be static, whereas the path used for data between the server and the UE could change due to user's mobility. Therefore, the selected PGW should be able to filter the traffic sent to the UE in order to forward signalling traffic on one mobile connection and data traffic on another mobile connection.

Specifying different bearers between a PGW and a UE within the mobile network is allowed by 3GPP [4]: each bearer is allocated with a different QoS and assigned with its appropriate QoS Class Identifier (QCI). Typically, the Classical LTE Default Bearer is assigned with best-effort QCI for non-Guaranteed Bit Rate (non-GBR), whereas Dedicated-Bearers for (video, voice, etc.) are assigned with different QCIs with GBR.

However, the PGW receives IP packets from the server, which carry the same source and destination IP addresses (respectively the server's IP address and the UE's IP address). Two options can be envisaged for the PGW to filter the traffic associated with the MPTCP connection:

1. the PGW implements Deep Packet Inspection (DPI) procedures in order to discriminate within an MPTCP connection signalling from data;
2. MPTCP packets are marked by the server with different Type of Service (ToS) indicators depending on whether they carry MPTCP signalling or data.

Furthermore, to achieve a smooth handover during user's mobility, the following requirements have to be ensured:

- Requirement 1': After the establishment of smooth SIPTO connection, the UE must not set the MPTCP path as backup path, as it is already used as regular path for SIPTO connection;
- Requirement 2': During handover, the MME must ensure a no-PGW relocation for the initial path used for MPTCP signalling and possible PGW relocation for SIPTO path;
- Requirement 3': At the end of smooth handover for SIPTO above RAN connection, the UE must not remove its "anchor" IP address from the server. Instead, the UE sends MP_Prio option of MPTCP to set the path with the "anchor" IP address as Backup-Path and the new SIPTO path as Regular-Path

Assuming that one of the two above options is indeed feasible, it can be seen that the three requirements for smooth SIPTO can indeed be fulfilled.

- Requirement 1': this deals with a procedure related to the MPTCP control part of smooth SIPTO and can thus be easily implemented as it only depends on UE configuration;

- Requirement 2' can be fulfilled if the MME considers the initial path as a "regular LTE" bearer, and the subsequent paths as SIPTO bearers.
- Requirement 3' is also related to MPTCP control and can thus be easily configured within the UE.

Considering MPTCP traffic filtering, the first option could be very demanding on the Anchor PGW, whereas the second is only possible if the network operator does not modify the ToS indicator between the server and the Anchor PGW.

Therefore, if none of the two options is indeed feasible, the MME should be configured to select an Anchor PGW, which should not be changed during the UE's activity; this Anchor PGW may not be close to the UE, and could e.g. be high in the network (at the Core CO or within the Centralized EPC). Then, when the UE requests a different connection, the MME could select a PGW close to the UE, in order to optimize data paths.

5.3.2 Dealing with Non-MPTCP Capable User/Server

The proposed smooth SIPTO solutions are applicable as long as both the UE and the servers are MPTCP capable. If the UE or the servers are non-MPTCP capable, we propose to use proxy MPTCP as a solution for a smooth SIPTO support. Proxy MPTCP is a fixed anchor that is used to enable the UE to initiate an MPTCP connection with the server. Figure 5.11 illustrates the placement of proxy component in the proposed smooth SIPTO architecture for MC2 within FMC network topology. Depending on the non-MPTCP client type, we propose the two scenarios explained below.

1. UE: In context of smooth SIPTO, a UE has to be MPTCP capable to achieve seamless mobility. This is due to the fact that changing the proxy MPTCP would break the ongoing session. Therefore, the proxy MPTCP has to be fixed relative to the UE. Thus, either a UE has inbuilt MPTCP support or it has installed the proxy such as a lightweight proxy MPTCP proposed in [76].
2. Server: If the server is non-MPTCP capable, the placement of the proxy MPTCP should ideally be close to the server. Therefore, the ideal place for the placement of proxy MPTCP would be within the NG-PoP (e.g. inside the UAG). This placement which ensures no relocation of proxy MPTCP during an ongoing session.

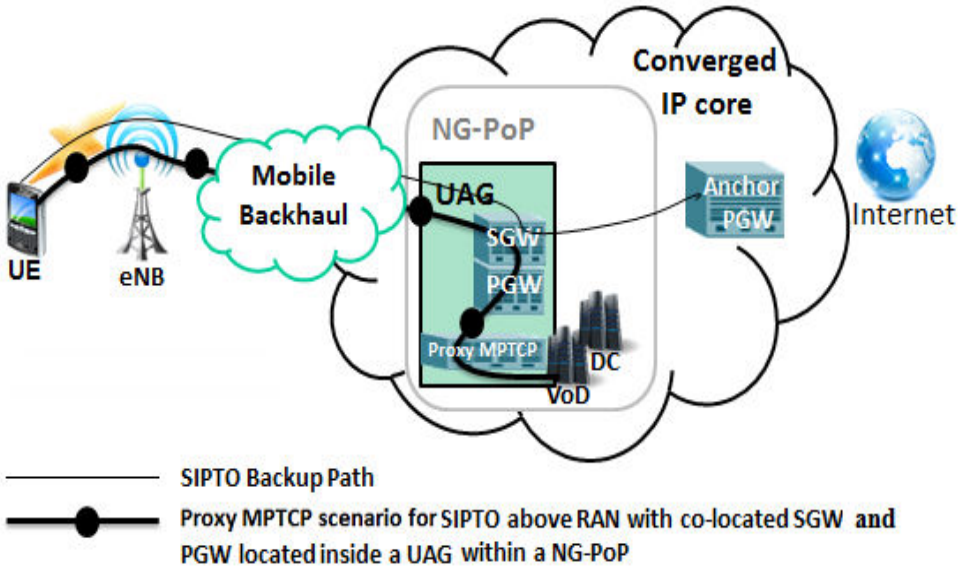


FIGURE 5.11: Proxy MPTCP placement for non-MPTCP capable servers

5.4 Conclusion of Chapter 5

In this Chapter, we first presented the work achieved in WP3 of COMBO Project. We then outlined how Classical LTE and smooth SIPTO functional entities is mapped on COMBO's uDPM functional blocks proposed in [12] and presented in Section 5.1. Next, we proposed different mapping scenarios of smooth SIPTO proposals presented in Chapter 4 on COMBO FMC network topology proposed in [12] and [13] where we introduced an Anchor PGW function to the COMBO architectures in order to maintain smooth SIPTO session continuity during user's mobility. Finally, as the support of MPTCP features represent one of the key elements required to apply smooth SIPTO approaches, we proposed the use of a proxy MPTCP as a solution to support mobility of users with SIPTO above RAN session in case the UE or the servers are non-MPTCP capable. As a result, our proposals for smooth SIPTO mobility support, are implementable on COMBO FMC architectures proposed in [12] and [13], as they could easily be mapped on uDPM functional blocks and on candidate FMC network architectures.

The next Chapter focuses on the mobility support of smooth SIPTO in candidate FMC architectures presented in Section 5.1.3.

Chapter 6

Smooth SIPTO Mobility Support in Candidate FMC Architectures

In this Chapter, we analyse how each implementation option of the UAG on FMC architectures impacts on the realization of smooth SIPTO. Section 6.1 focuses on the deployment options of smooth SIPTO architectures when a mobile UE connects to a fixed server. Sections 6.2 and 6.3 respectively address how mobility is supported in FMC networks when CDN-based services and mobile servers are considered. In Section 6.4, the applicability of smooth SIPTO mobility support within COMBO FMC architectures is evaluated in terms of interruption duration and signalling volume. Finally, Section 6.5 outlines how the proposed FMC architectures with respect to smooth SIPTO mobility support apply to advanced-4G and 5G architectures.

6.1 Supporting FMC Mobility of Mobile-User Connected to Fixed Server

In this Section we analyse the different mobility use-cases in FMC networks when a UE is connected to a Fixed server (e.g., VoD server). We assume here that the UE will be connected to the same server during all its session.

Considering the two UAGs' implementation models presented in Section 5.1.2 , we note that having the UAG DP and CP split from each other allows different degrees of flexibility for deploying the UAG CP. Therefore, in this section we only represent the mobility scenarios when split UAG model is implemented with CP and DP functions are either co-located together within a single UAG or separated from each other in different UAGs.

We recall here that in the proposed FMC networks, the logical elements of 3GPP mobile architecture are included within the UAG. In particular, the MME functionality will be implemented within the UAG CP, whereas SGW and PGW functionalities should be deployed within the UAG DP. In addition, as pointed out in Section 5.3, the co-located Proxy-SGW/LGW functionalities will be implemented at the fixed access level (e.g., co-located with the HGWs that host the HeNBs).

COMBO architectures introduces the concept of NG-POP, which represents a location in the network in which IP-edges (UAG DP) and services are co-located.

In particular, in distributed COMBO architecture, NG-POP is located in the Main CO, leading to an extension of the IP backbone towards the access network. Whereas, in centralized COMBO architecture, NG-POP is located high in the Core CO.

UAG CP functions on the other hand, are either co-located with the UAG DP within the NG-POP or implemented as a standalone function in the Core CO or very high in the EPC.

6.1.1 Smooth SIPTO Mobility Support in Distributed COMBO FMC architecture

In this section we present the mobility scenarios of a user with respectively, a smooth SIPTO above RAN session and a smooth SIPTO at LN session, in a distributed COMBO architecture.

6.1.1.1 Smooth SIPTO for MC2 in Distributed COMBO Architecture

Considering the scenario in Figure 6.1 where a FMC user is inside a moving vehicle while having an ongoing smooth SIPTO above RAN session.

As shown in this figure, a UE is attached to the network with an Anchor PGW, which is located within the closest UAG DP in the Main CO. The UE is also using the same PGW to offload SIPTO data to the server co-located with its associated PGW. The anchor PGW then ensures both MPTCP signalling and MPTCP data paths. This is achieved using one of the options proposed in Section 5.3.1 of Chapter 5.

Figure 6.1-a illustrates a distributed COMBO scenario where UAG CPs are split from the UAG DPs and located at the Core CO. As shown in this figure, having the UAG CP functions split in the Core CO allows limiting the potential UAG CP (MME) relocation in case of user's mobility. This is due to the fact that, in general, it is very rare that

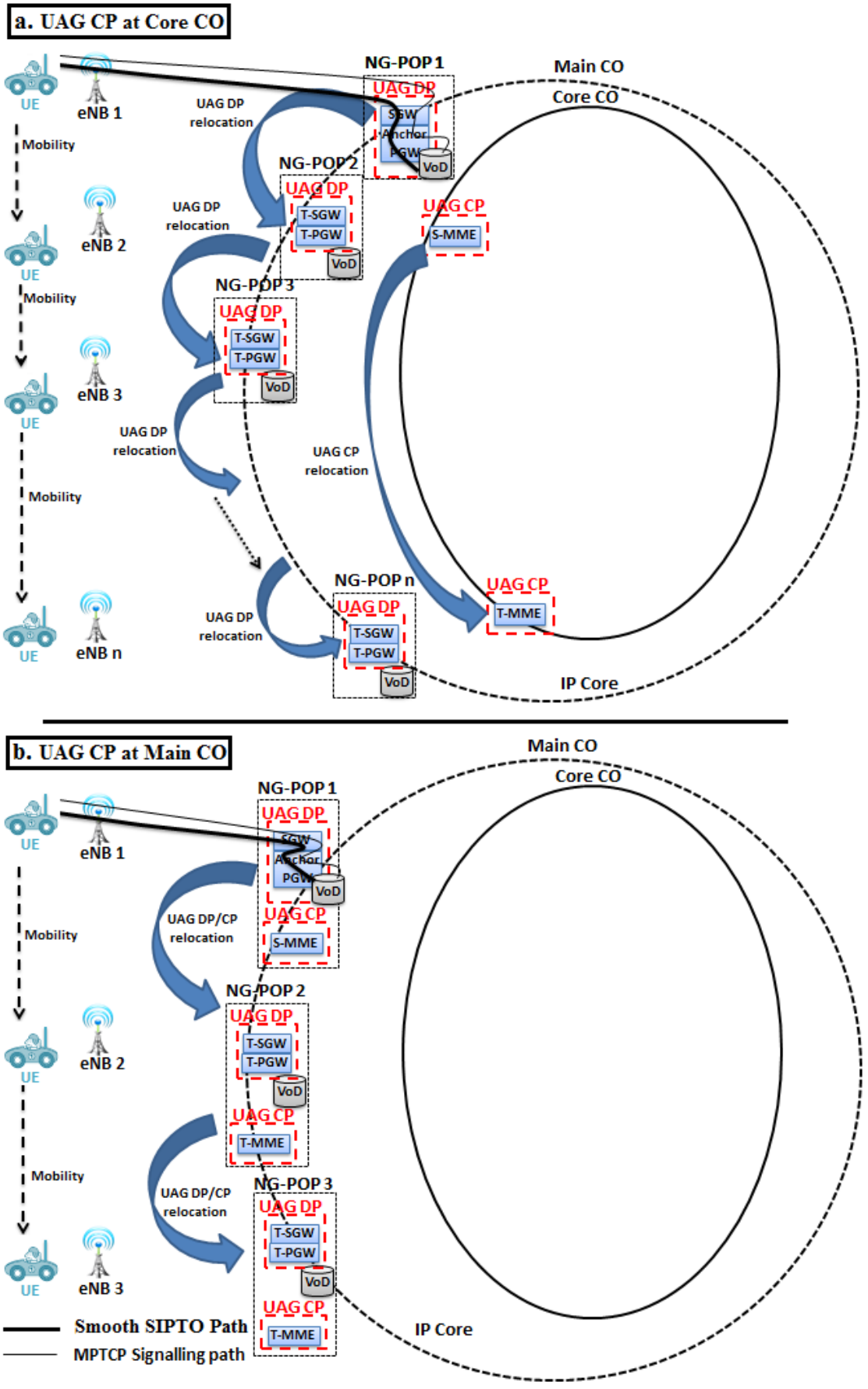


FIGURE 6.1: Smooth SIPTO for MC2 in Distributed COMBO Architecture

a user makes a journey of hundreds of Kilometres (KMs) in his usual days. Whereas, compared with UAG CP, having the UAG DP close to the user leads to a much more frequent relocation of the UAG DP. For instance, in this figure we have: "*one*" UAG CP relocation per "n" UAG DP relocations.

Figure 6.1-b, on the other side, illustrates a scenario where UAG CP and UAG DP are split yet co-located together within the same NG-POP. Even though having the MME functions within the UAG CP at the Main CO would improve the data paths management (reducing the number of controlled UEs per UAG CP), this scenario is considered impractical to manage user's mobility. This is due to the fact that the complexity of mobility management is much higher when UAG CP is closer to the UE. Indeed, as shown in this figure, such implementation would result on relocating the UAG CP (MME functions) whenever its co-located UAG DP (SGW/PGW) is relocated.

As a result, when distributed COMBO architecture is implemented, smooth SIPTO for MC2 is supported regardless of the UAG CP location. The operator can decide where to implement the UAG CP according to its requirement for mobility control and/or to the number of users to be controlled in that zone.

Furthermore, in both scenarios, the anchor PGW remain the same during the lifetime of the user's session. The only PGW that is relocated is the PGW used for offloaded traffic.

We should note here that for static UEs or for UEs with low velocity (e.g., walking or biking user), session continuity is not an issue. Indeed, session continuity is considered an issue only for UEs with high velocity (e.g. a user in a car, a bus or a train) where user's journey includes long distances.

6.1.1.2 Smooth SIPTO for MC3 in Distributed COMBO Architecture

In this section we consider the scenario in Figure 6.2 where an FMC user is having a smooth SIPTO at LN session while walking around a large University Campus, and attaching its device to different HeNBs subsequently.

Similar to smooth SIPTO for MC2 in distributed COMBO architecture, an Anchor PGW is selected for this user to maintain the user's session continuity by ensuring MPTCP signalling path.

Figure 6.2-a and Figure 6.2-b illustrate a distributed COMBO architecture where a split UAG model is implemented with, respectively, separate and co-located UAG DP and UAG CP. As shown in these figures the user's mobility is limited by distance. Therefore,

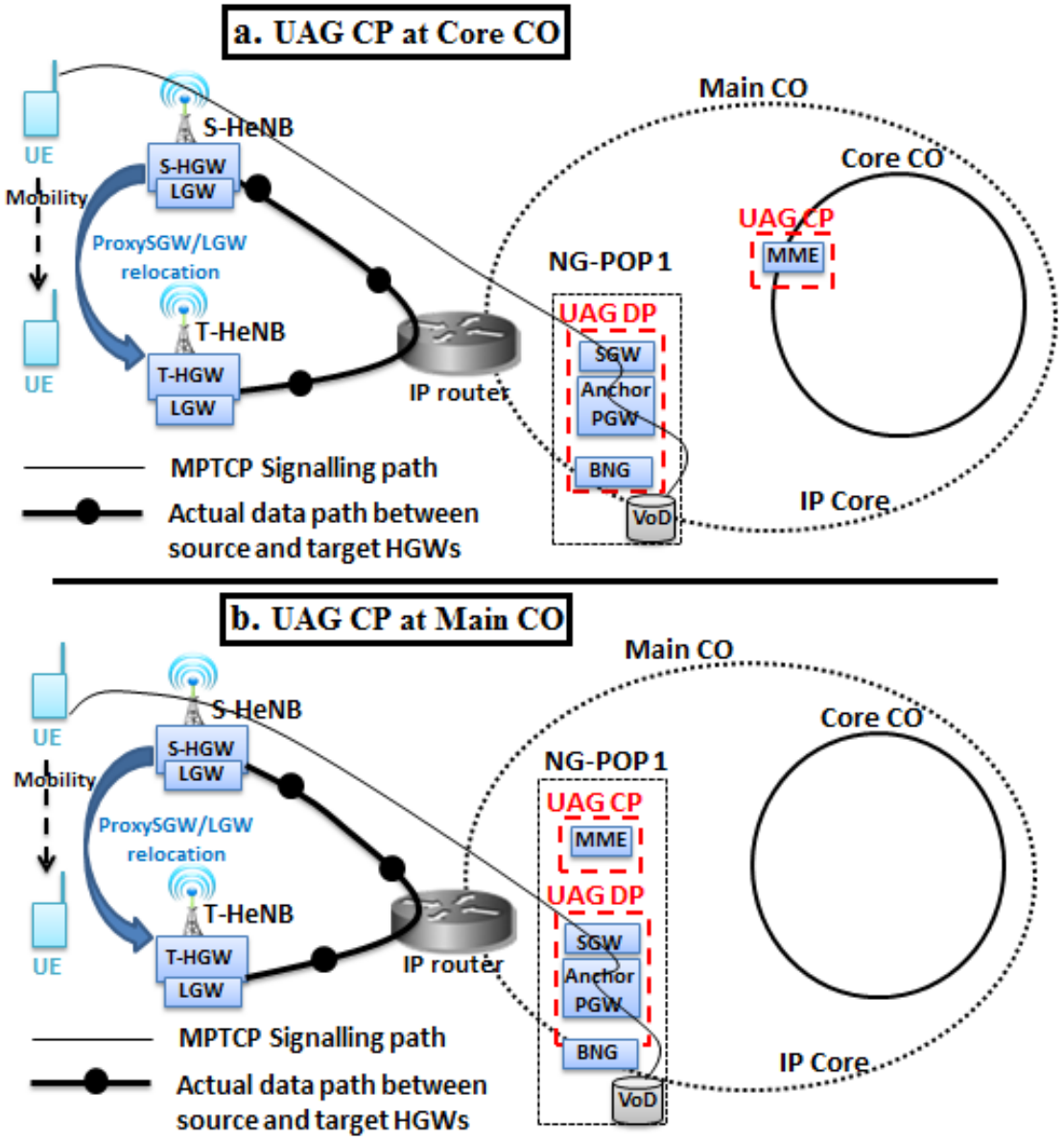


FIGURE 6.2: Smooth SIPTO for MC3 in Distributed COMBO Architecture

when the user is on a move, neither the UAG CP nor the UAG DP would be relocated. Instead, the LGW, which is co-located with the user's associated HeNB that would be relocated; i.e., during user's mobility, a handover procedure of smooth SIPTO for MC3 would is performed each time the UE changes its associated HeNB/LGW.

We should point out here that, although the IP edge in smooth SIPTO at LN sessions is located at the fixed access level (i.e., co-located with the HGW), the local data traffic in real field deployment can reach the external IP network only through an IP router located at the border of the Main CO's level (in the IP Core). For instance, considering the scenario in Figure 6.2, in real field deployment, when the FMC user moves and a smooth handover for local traffic is performed with LGW relocation (see Section 4.2

of Chapter 4), the forwarded downlink data traffic, which uses the indirect forwarding tunnel between the source and target co-located ProxySGWs/LGWs, should be routed first from the source co-located ProxySGW/LGW within the source HGW to the IP router located at the border of the Main CO, then form the IP router to the target co-located ProxySGW/LGW within the target HGW.

As a result, similar to smooth SIPTO for MC2, local mobility in distributed COMBO architecture is supported for FMC users regardless the UAG CP location.

Besides, compared with Classical LTE architectures, selecting the default PGW (Anchor PGW) within the Main CO, improves the network performance in case SIPTO path breaks. Thus, in this case, the users IP session is resumed directly in the default path with the initial Anchor PGW within the distributed EPC (instead of Centralized EPC as presented in Classical LTE architecture). Indeed, selecting a distributed Anchor PGW, would allow keep serving the user's requests from the same distributed VoD server even if the user's connection with the LGW used for smooth SIPTO at LN breaks. A significant amount of bandwidth is then saved at both the core and the metro segments of the network.

6.1.2 Smooth SIPTO Mobility Support in Centralized COMBO FMC Architecture

In this Section we present the mobility scenarios of a user with respectively, a smooth SIPTO above RAN session and a smooth SIPTO at LN session, in a centralized COMBO architecture.

We assume in this section that, the UE in both scenarios is attached to the network using an Anchor PGW located within the UAG DP in the Core CO.

6.1.2.1 Smooth SIPTO for MC2 in Centralized COMBO Architecture

Considering the scenario in Figure 6.3 where the FMC user is inside a moving high-speed train while having an ongoing smooth SIPTO above RAN session towards the IP services located within the closest NG-POP at the Core CO.

Similar to the scenario in Section 6.1.1.1, we assume here that the UE is using its initial PGW (Anchor PGW) to ensure both MPTCP signalling and MPTCP data paths.

As shown in Figure 6.3, due to user's mobility, a smooth handover for MC2 with UAG DP relocation may be performed. Smooth SIPTO mobility in this case is supported

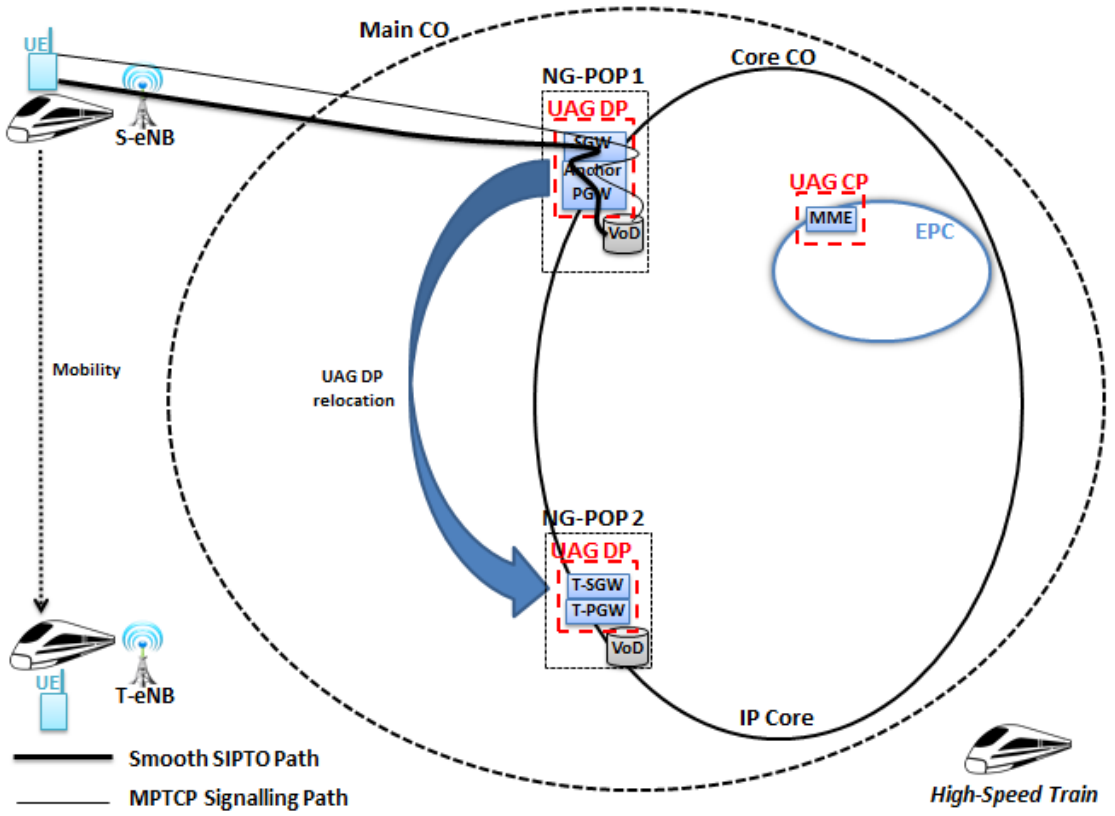


FIGURE 6.3: Smooth SIPTO for MC2 in Centralized COMBO Architecture

using the requirements descried in Section 5.3.1 of Chapter 5. In this figure we consider that the UAG CP is centralized very high in the IP Core (within the centralized EPC). Compared with the legacy mobile networks, this corresponds to the current location of MMEs functions.

As shown in this figure, compared with distributed COMBO scenario presented in Section 6.1.1.1, having the IP edges at the Core CO allows limiting the potential UAG DP (SGW/PGW) relocation in case of users' mobility. Such implementation would then improve the mobility management and minimize the signalling load of the different network elements (e.g., during the user's mobility, when no UAG DP relocation is required, the handover procedure can simply be handled thanks to X2 interface between the source and destination eNBs).

A network operator may also decide to deploy the UAG CP in the Core CO (co-located with UAG DP within the NG-POP). It should be noted here that the deployment of such scenario, will result on relocating the UAG CP functions whenever the UAG DP functions are relocated during the user's mobility.The probability of relocating network functions located in the Core CO is although very low, compared to the distance required to perform this relocation.

Finally, noting that the main benefit from implementing smooth SIPTO above RAN with co-located SGW/PGW architecture is to allow mobile users to access the IP services very close to the user's location, then the deployment of smooth SIPTO above RAN architecture within Centralized COMBO FMC architecture where IP services are co-located with the UAG DP in the Core CO will limit the applicability of SIPTO above RAN approach.

6.1.2.2 Smooth SIPTO for MC3 in Centralized COMBO Architecture

Similar to the local mobility in distributed COMBO architecture presented in Section 6.1.1.2, in a scenario where smooth SIPTO at LN is considered in centralized COMBO architecture, as the mobility of users is limited by distance, then neither the UAG CP nor the UAG DP would be relocated.

In general, local mobility of users with ongoing smooth SIPTO at LN sessions in Centralized COMBO architecture is mainly supported similarly to the local mobility in distributed COMBO architecture. Although, in centralized COMBO architecture, the IP router to be used in the real field deployment, is now located at the edge of the Core CO (instead of the Main CO).

6.2 Supporting FMC Mobility of Mobile-User Connected to CDN-based Service

In Section 6.1.1.1 and Section 6.1.2.1 we note that, since the content servers are co-located with the UAG DP within the NG-PoP, potential service relocation could then occur whenever a UAG DP relocation is performed. This could correspond to a scenario when the content-server connected to the user is part of a CDN-based service.

Assuming that in the scenario shown in Figure 6.4, the user is having a smooth SIPTO above RAN session towards a CDN-based content service, while being in a moving vehicle. As shown in this figure, due to user's mobility UAG DP relocation is performed and a handover procedure of smooth SIPTO for MC2 is performed.

From the content distribution service point of view, it could happen that the new IP address allocated to the UE belongs to a location that is geographically closer to another cache of the CDN-based service (e.g content-server in NG-POP2 as in Figure 6.4).

However, a single MPTCP connection cannot connect the UE to multiple servers. If the content distribution service decides to attach the UE to the closer server, the initial

MPTCP connection is terminated and session continuity is not assured by the smooth SIPTO procedure.

It does not necessarily mean that the content distribution service is degraded as some content distribution services, relying on adaptive bit rate streaming, could include procedures maintaining session continuity even when the network does not maintain it [73].

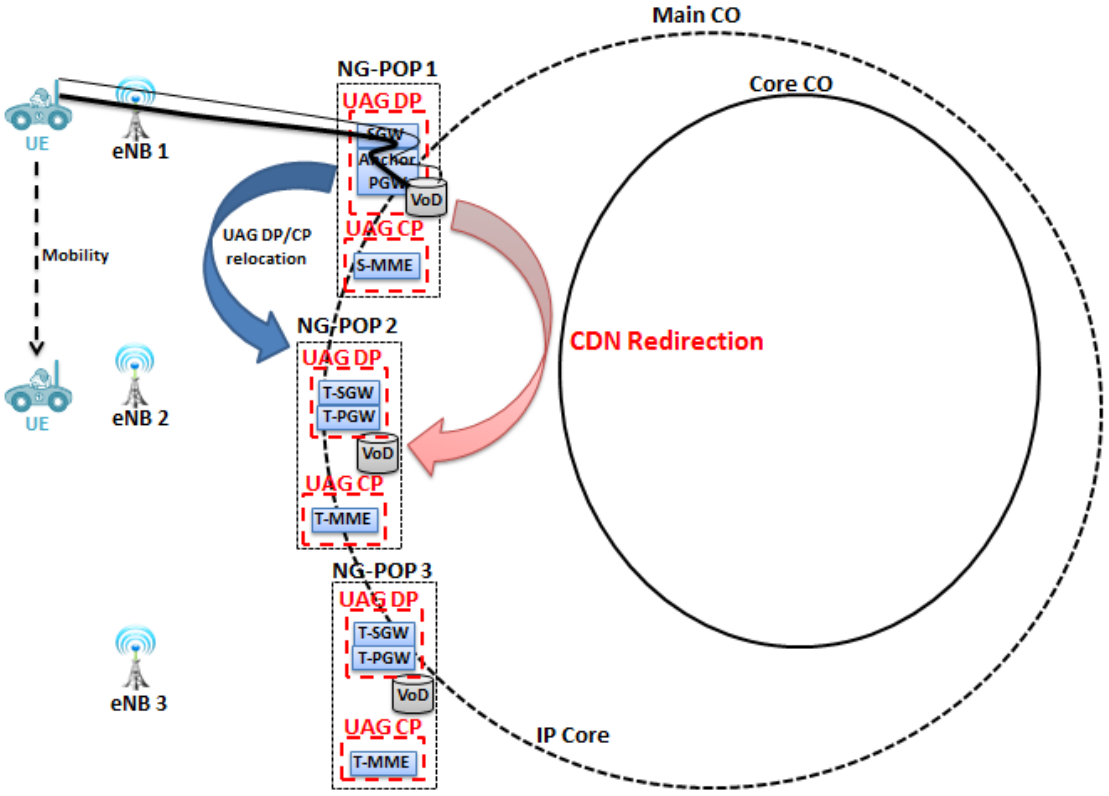


FIGURE 6.4: CDN Redirection in Distributed COMBO Architecture

Therefore, in order to rely on the smooth SIPTO architecture for CDN services, we do not recommend content-server's relocation in such user's mobility use cases. The user can obviously connect with the new content-server for its new sessions, i.e., new data-traffic demands will be served from the closest server to the user's location (VoD within NG-POP2 in Figure 6.4).

6.3 Supporting FMC Mobility of Mobile-User Connected to another Mobile-User

In smooth SIPTO proposals presented in Chapter 4, we only focus on mobility scenarios when a user is connected to a fixed-server. In this Section we discuss the potential mobility use cases when two FMC mobile-users are communicating with each other using e.g., Video over LTE (ViLTE) or Voice over LTE (VoLTE) services.

Assuming the scenario in Figure 6.5 where UE1 and UE2 are having a ViLTE session using smooth SIPTO above RAN approach in a distributed COMBO architecture. As shown in this figure, a user's data traffic is first routed to a distributed IMS architecture within the IP Core for media processing (traffic is provided with the adequate QoS and transcoded) and then is forwarded to the second user. In this scenario we also assume that both users are connected with the same Anchor PGW located within the centralized EPC for MPTCP signalling.

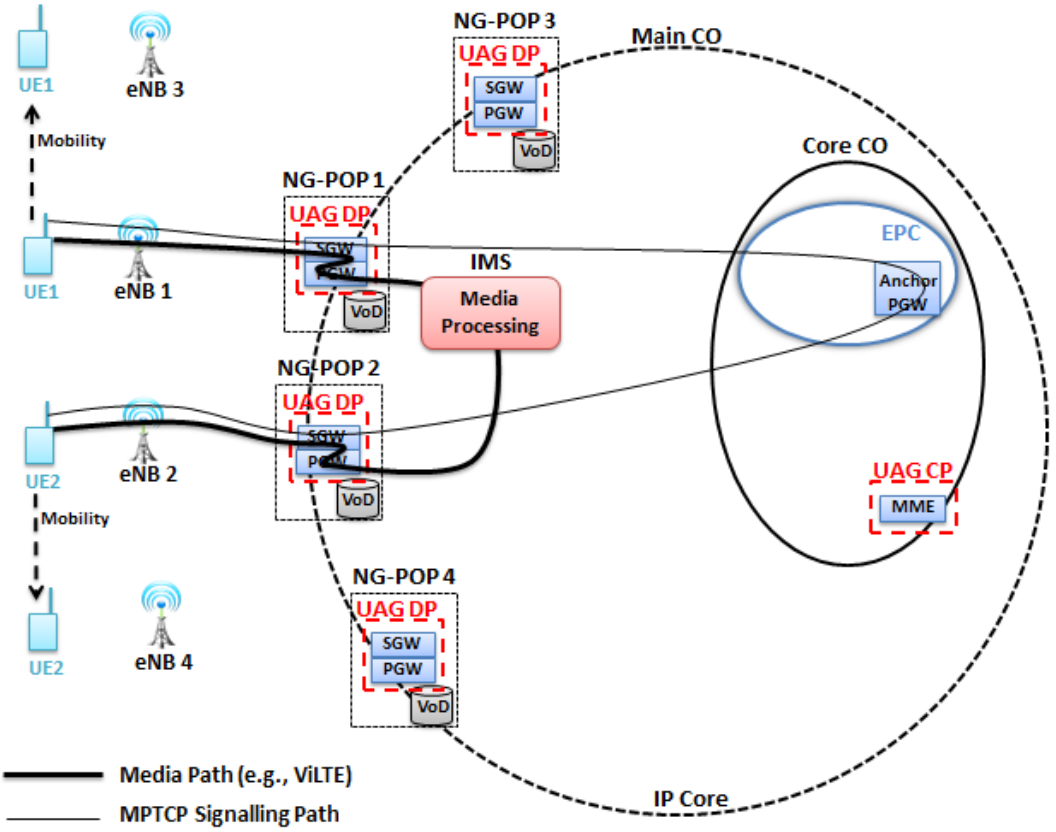


FIGURE 6.5: FMC mobility support for two mobile users with ongoing ViLTE session

In a scenario where one of the users is mobile (with UAG DP relocation) while the other user is static, the FMC mobility will then be supported similarly to the FMC mobility use case presented in Section 6.1.1.1 (of smooth SIPTO for MC2 in distributed COMBO architecture), where a UE is connected to a fixed server. The same FMC mobility use case can also apply to a scenario where UE1 and UE2 moves in sequence (i.e, UE2 starts its handover procedure once that UE1's handover is completed).

Considering now a scenario where both UEs are moving simultaneously. Assuming that, due to users mobility, a smooth SIPTO handover procedure with UAG DP relocation is then performed for each UE.

In order to maintain the users session continuity, each UE establishes an indirect forwarding path with its associated (source) PGW and the target SGW it's heading to (see initial media path in Figure 6.6).

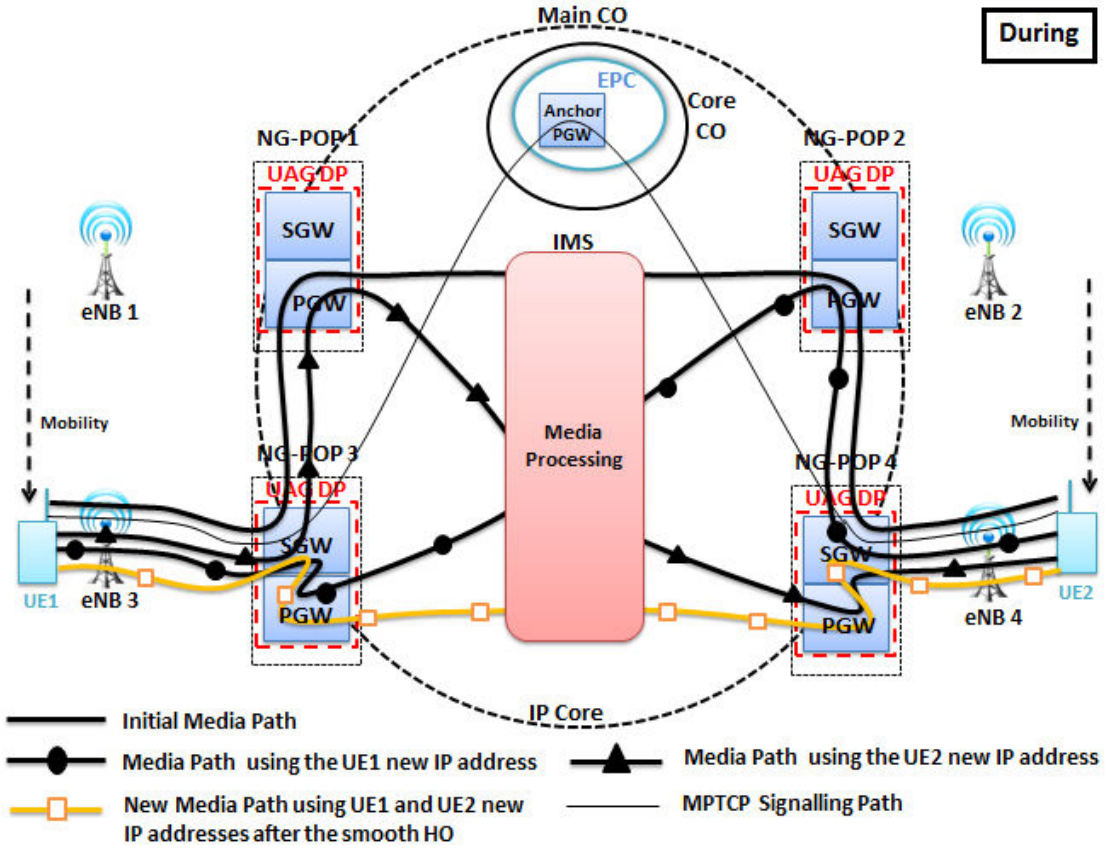


FIGURE 6.6: Media path during users mobility

Assuming that the UE1 starts its handover Completion phase before UE2. Then, during the completion phase of UE1, once its new smooth SIPTO connection is established, UE1 sends its new IP address to UE2 over the MPTCP signalling path ensured via the Anchor PGW. Once received, UE2 will join the new SIPTO path to its session using its IP address allocated by its source PGW (within NG-POP2). This new SIPTO path corresponds to the "media path using the UE1 new IP address". As mentioned in Chapter 4, the connections binding is performed using a three-way-handshake with the MPTCP option MP_JOIN.

In parallel, during the completion phase of UE2, the latter also establishes a new smooth SIPTO connection with its target UAG DP (within NG-POP3). A new IP address is then allocated for this user as well. Similar to UE1, UE2 now sends its new IP address to UE1 over the MPTCP signalling path ensured via the Anchor PGW.

Once received, UE1 has to initiate on its side, two additional sub-connections using its both IP addresses (the IP address allocated by its source PGW (within NG-POP1) and

the IP address allocated by its target PGW (within NG-POP3)) with the IP address received from UE2 (IP address allocated by its target PGW (within NG-POP4)). This respectively corresponds to the "media path using the UE2 new IP address" and the "new media path using UE1 and UE2 new IP addresses" illustrated in Figure 6.6.

At this stage, thanks to MPTCP, the established subflows form a full mesh using all available paths between the two mobile users.

At the end of the handover procedures of both smooth SIPTO connections, the only paths remaining that will be used by the users are the new smooth SIPTO paths with the target UAG DPs and the MPTCP signalling paths with the Anchor PGW.

The FMC mobility use case presented above can easily be applied to local mobility between two users with smooth SIPTO for MC3. As a result, FMC mobility when two users are having a media session (ViLTE or VoLTE) can also be supported thanks to our smooth SIPTO proposals.

6.4 Performance of Smooth SIPTO

In this section, the performance of the smooth SIPTO handover scenarios in the framework of COMBO architectures is evaluated in terms of interruption time duration and signalling traffic volume. The evaluation is focused on finding out whether the scenarios proposed in Chapter 4 can reduce the interruption time of SIPTO traffic during mobility of UEs between two cells when a PDN gateway (PGW or LGW) relocation decision is triggered by the MME. Besides, the evaluation of the volume of signalling traffic, would help us decide which location would be better to deploy the UAG CP (MME) within the COMBO FMC architecture.

6.4.1 Interruption Time

Let PD and LD respectively represent the nodes' processing delay and the links' transmission time. Let also SDD and SED respectively represent the SIPTO connection deactivation delay and the SIPTO connection establishment delay.

The graph in Figure 6.7 illustrates the standard SIPTO traffic interruption time during user's mobility scenarios (MC2 and MC3) with no handover support.

As pointed out in [4], when PGW or LGW is relocated, the interruption time is given by:

$$IT_{standardS} = SDD + SED + X \tag{6.1}$$

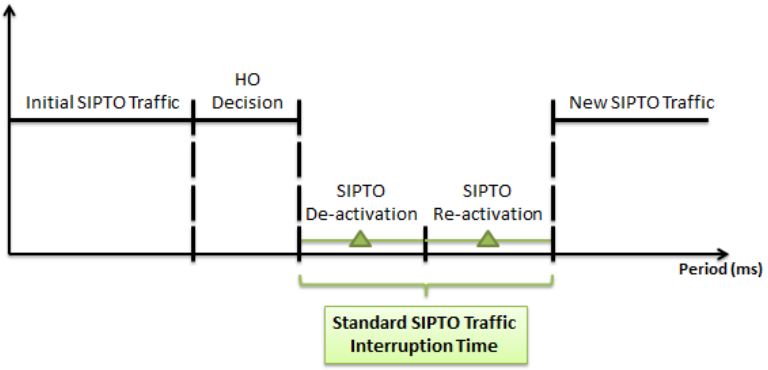


FIGURE 6.7: Standard SIPTO Interruption Time for MC2 and MC3

X is the total propagation and processing delay from MME to HSS, MME to DNS and PGW to PCRF. Typical values for those delays are respectively 100 ms, 50 ms and 100 ms [77].

Thanks to the transition diagrams for activation and deactivation procedures in([4], [7]), we have illustrated in Figure 6.8 the standard handover procedure for SIPTO above RAN session with co-located SGW/PGW.

As shown in this figure, due to the PGW relocation triggered by the MME, the handover is failed and a deactivation procedure with reactivation is performed. From this Figure, the SDD and SED can directly be derived and the interruption time for MC2 is computed as follows:

$$\begin{aligned} IT_{standardMC2} = 2\{&2\ PD_{(\mathrm{UE})} + 3\ PD_{(\mathrm{eNB})} \\ &+ 2\ PD_{(\mathrm{MME})} + 2\ PD_{(\mathrm{SGW})} + PD_{(\mathrm{PGW})} \\ &+ 3\ LD_{(\mathrm{UE\_eNB})} + 3\ LD_{(\mathrm{eNB\_MME})} + 2\ LD_{(\mathrm{MME\_SGW})} \\ &+ 2\ LD_{(\mathrm{SGW\_PGW})}\} + X \end{aligned} \tag{6.2}$$

Furthermore, with the help of the transition diagrams for the activation and deactivation procedures shown in [7] and [78], we illustrated in Figure 6.9 the standard handover procedure for for users mobility with ongoing SIPTO at LN sessions (MC3) described in Chapter 2. From this figure, the SDD and SED for standard MC3 are derived, yielding:

$$\begin{aligned} IT_{standardMC3} = 2\{&2\ PD_{(\mathrm{UE})} + 3\ PD_{(\mathrm{HeNB})} \\ &+ 2\ PD_{(\mathrm{MME})} + 2\ PD_{(\mathrm{SGW})} + PD_{(\mathrm{LGW})} \\ &+ 3\ LD_{(\mathrm{UE\_HeNB})} + 3\ LD_{(\mathrm{HeNB\_MME})} + 2\ LD_{(\mathrm{MME\_SGW})} \\ &+ 2\ LD_{(\mathrm{SGW\_LGW})}\} + X \end{aligned} \tag{6.3}$$

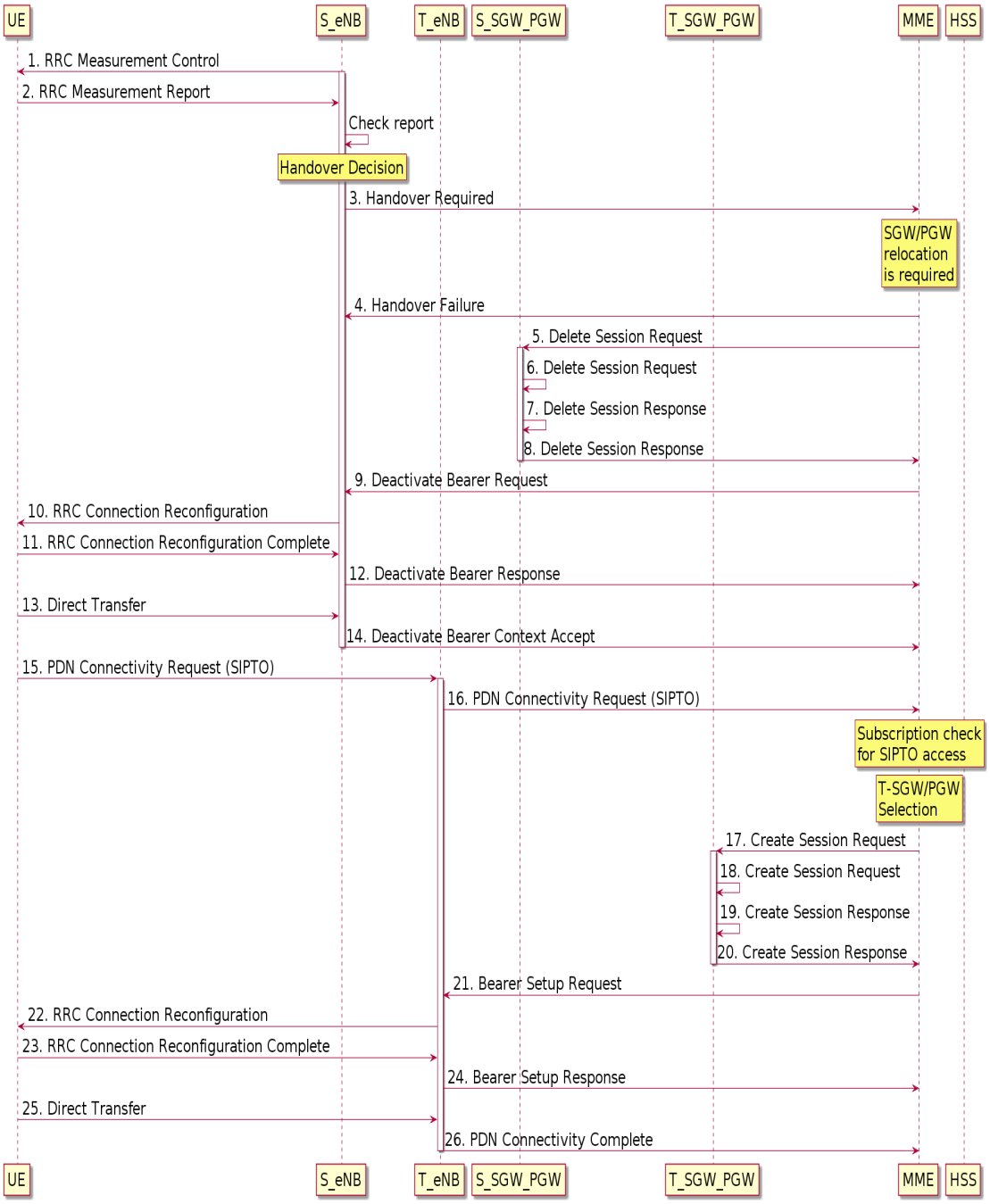


FIGURE 6.8: Deactivation and Re-activation Procedure for Standard MC2 Scenario

Compared with standard handover scenarios for MC2 and MC3, the smooth SIPTO proposals, described in Chapter 4, enables an always-on mobile connectivity by initiating an S1-based handover procedure for the initial SIPTO Connection, while establishing a new SIPTO data path to be used for the ongoing session before deactivating the initial SIPTO data path. Indeed, during smooth SIPTO handover procedure, both SIPTO paths should be active, joint and used together for the ongoing session.

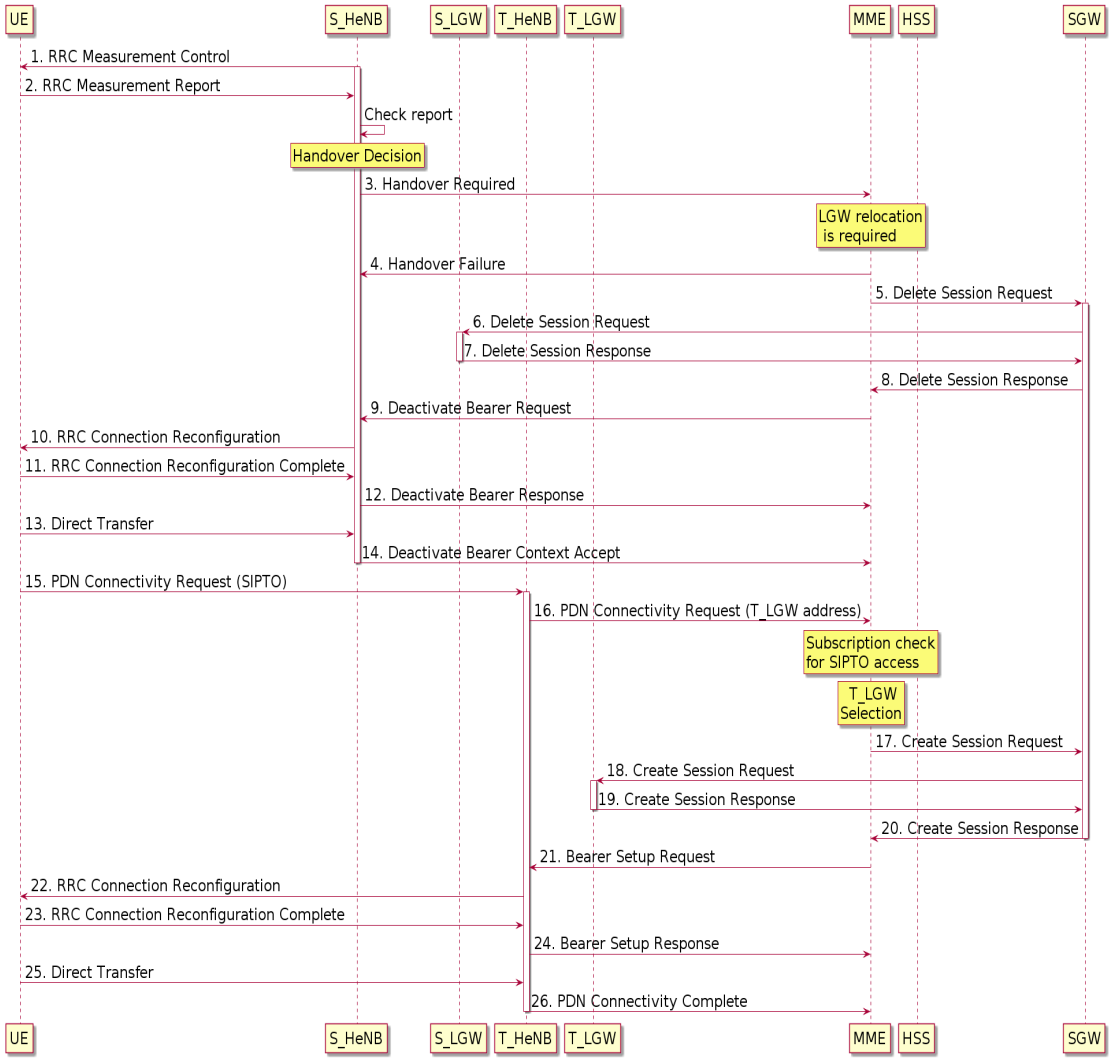


FIGURE 6.9: Deactivation and Re-activation Procedure for Standard MC3 Scenario

Figure 6.10 illustrates the different steps to be performed during the smooth SIPTO handover procedure. As shown in this figure, the interruption time in smooth SIPTO proposals is present only during the Execution phase of the initial SIPTO handover.

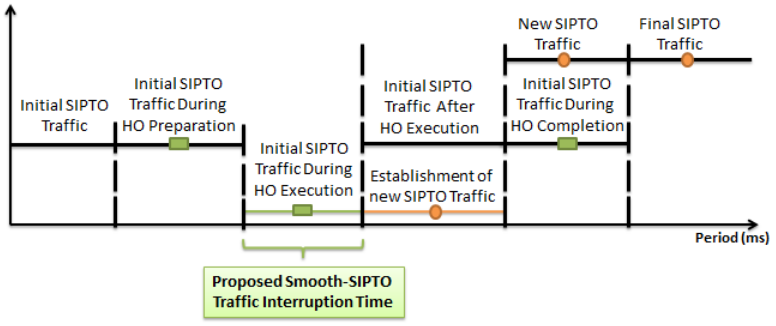


FIGURE 6.10: Smooth SIPTO Interruption Time for MC2 and MC3

Based on the mapping scenarios of smooth SIPTO proposals on distributed COMBO

and Centralized COMBO architectures, shown in Section 6.1.1 and Section 6.1.2, and according to the MME's location in each scenario, the following potential options of the smooth SIPTO handover procedures could be considered:

- **smooth SIPTO handover without MME relocation:** This scenario could be considered for both smooth SIPTO handover proposals.
- **smooth SIPTO handover with MME relocation:** Unlink the first scenario, this scenario is only considered during smooth SIPTO handover with SIPTO above RAN sessions.

The interruption time in each of these scenarios differs depending on the UAG CP (i.e., MME) location. The generic Execution phase of smooth SIPTO handover procedures proposed in Chapter 4 is shown in Figure 6.11, with associated delays encountered in the procedures. Let (a), (b) and (c) be defined as in Table 6.1 according to the notations in 3GPP standard [79].

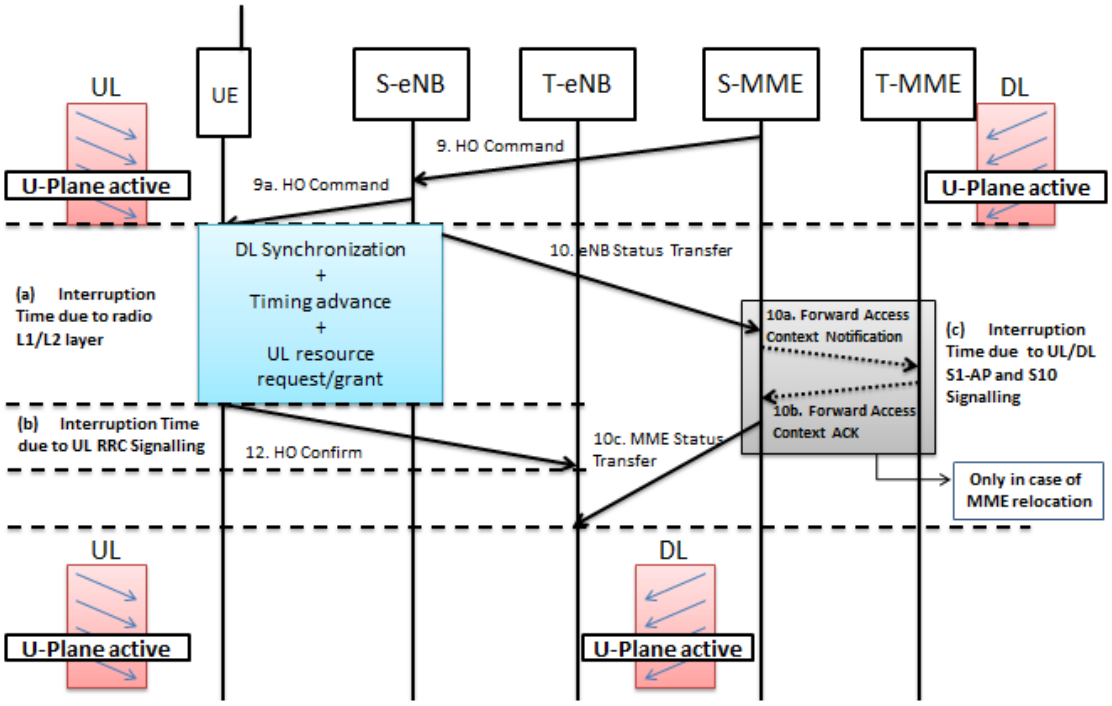


FIGURE 6.11: Data-Plane interruption involved in the smooth SIPTO handover procedures proposed in Chapter 4

The components (a) and (b) are explained in details in [79], whereas the component (c) is elaborated below.

TABLE 6.1: Constituent for the DP interruption delay

(a)	Radio Layer Process
(b)	Uplink RRC Signalling
(c)	Status Transfer Signalling
(c1)	S1-C Signalling for Status Transfer
(c2)	S10 Signalling for Access Context Forward

- Status Transfer Process (c)
  This is the delay between the handover command to the MME status transfer, hence consisting of these two elements:

  1. S1-C signalling for status transfer (c1): The time taken for status transfer depends on the distance between the MME and the source and target eNBs. According to 3GPP in [79], the S1-C transfer delay varies between 2 ms to 15 ms, depending on the MME's location.
  2. S10 signalling for access context forward (c2): This delay is included in the DP interruption delay for smooth SIPTO above RAN proposals, if and only if, an MME relocation is performed during the users MC2. Indeed, this delay depends particularly on the distance between the source and target MMEs and the location of the source MME.

According to this model, the total interruption delay of Data-Plane in smooth SIPTO handover procedures equals (a)+(b) or (c), whichever is larger. As a result, the interruption delays for smooth SIPTO MC2 and MC3 could be derived as follows:

- Interruption time for proposed MC2 "without MME relocation" equals:

$$IT_{proposedMC2} = MAX[(a) + (b), (c1)] \tag{6.4}$$

with $(c1) = PD_{(\text{MME})} + LD_{(\text{S-eNB\_MME})} + LD_{(\text{MME\_T-eNB})}$,

- Interruption time for proposed MC2 "with MME relocation" equals:

$$IT_{proposedMC2'} = (c1 + c2) \tag{6.5}$$

with
$(c1 + c2) = 2\ PD_{(\text{S-MME, T-MME})} + LD_{(\text{S-eNB\_S-MME})} + LD_{(\text{T-MME\_T-eNB})}$
$+ 2\ LD_{(\text{S-MME\_T-MME})}$

- Interruption time for proposed MC3 "always without MME relocation" equals:

$$IT_{proposedMC3} = MAX[(a) + (b), (c1)] \tag{6.6}$$

with $(c1) = PD_{(\text{MME})} + LD_{(\text{S-HeNB\_MME})} + LD_{(\text{MME\_T-HeNB})}$

We used typical 3GPP values [79] for the processing delay in different nodes and for the link delay between two different nodes. The propagation delay between different nodes is taken to be equal to 5μsec/km. In SIPTO at LN the LGW and Proxy-SGW must be co-located together with the base station (HeNB) while in SIPTO above RAN, the co-located SGW/PGW must be located within the metro/core segment of the network (within the Main CO in COMBO FMC architecture). Based on this information, we assume that in today's mobile architecture, the distance from the eNB to the co-located SGW/PGW and the distance from the SGW to the MME typically equals 150 km. We also assume that the distance from the co-located HeNB and LGW to the default SGW and the distance from the HeNB to the MME equals 180 km. Therefore we obtain the results reported in Table 6.2. The diagonal entries in this table represent the processing delay of nodes whilst the other values correspond to link delays.

TABLE 6.2: Delay Budget for processing and links in today's (in ms)

	UE	HeNB	eNB	MME	SGW	Proxy SGW	LGW
UE	3	0	2	-	-	-	-
HeNB	0	4	-	9	9	0	0
eNB	2	-	4	7.5	7.5	-	-
MME	-	9	7.5	15	7.5	-	-
SGW	-	9	7.5	7.5	10	-	9
Proxy SGW	-	0	-	-	-	10	0
LGW	-	0	-	-	9	0	10

Now, let us consider the information given in Section 5.1.2, where we identified the distance from the UE to the Main CO and to the Core CO to be, respectively, less than 75 Km and approximately 300 Km. Considering also the scenarios presented in Section 5.1.3 for UAG CP and DP implementations, in split UAG model within both centralized and distributed COMBO architectures with respect to the scenarios represented in Section 6.1.1 and Section 6.1.2 for potential mapping options of smooth SIPTO proposals on both COMBO FMC architectures. We would then be able to identify the potential delay budget to be considered for the three interruption time components (a), (b) and (c). In order to have a realistic values for our evaluation, we used the delay ranging given in 3GPP standard [79] for S1-C transfer delays (min 2 ms, max 15 ms). Based on these delays, we proposed the delay budget reported in Table 6.3.

TABLE 6.3: Delay Budget for links in COMBO FMC (in ms)

	S1-C	S10
UAG CP in IP Core	15	20
UAG CP in Core CO	7.5	10
UAG CP in Main CO	2	3

Here, we assumed that, when the UE is accessing the external IP network at the edge of the Main CO's level where the maximum distance would not exceed 75 Km, the delay budget would be minimized to the least potential value (in this case S1-C equals 2 ms). Then, since the distance from the UE to the Core CO is 4 times the distance to the Main CO, we assumed that the average delay budget from the UE to the Core CO would be approximately the four times the delay budget required on the S1-C from the UE to the Main CO and thus, we assumed that the S1-C from the UE to the MME within the Core CO equals 7.5 ms. Similarly, since the IP Core is the farthest distance from the UE, we assumed that when the UAG CP is located at the IP Core, the delay budget on the S1-C interface would be maximum (i.e., 15 ms). On the other hand, we assumed that the distance between two MMEs is a little superior than the distance from a UE to the MME. As a result, we assumed that, the delay budget on the S10 interface would equals 3 ms, 10 ms and 20 ms, respectively when UAG CP is deployed within the Main CO, Core CO and IP Core.

Given the numbers in Table 6.2 applied to equations (6.2) and (6.3), and the numbers in Table 6.3 applied to equations (6.5) and (6.4)) we obtain the following results:

- the interruption time for standard scenarios ($IT_{standardMC2}/IT_{standardMC3}$) approximately equals to 500 ms.
- the interruption time for the smooth SIPTO handover scenarios when no MME relocation is required ($IT_{proposedMC2}/IT_{proposedMC3}$) given in equation (6.4) and (6.6) approximately equals to:
  - 21 ms when UAG CP is located at the Main CO.
  - 30 ms when UAG CP is located at the Core CO.
  - 45 ms when UAG CP is located at the IP Core.
- the interruption time for the smooth SIPTO handover in MC2 scenario when MME is relocated ($IT_{proposedMC2}$) given in equation (6.5) approximately equals to:
  - 55 ms when UAG CP is located at the Main CO.
  - 80 ms when UAG CP is located at the Core CO.
  - 120 ms when UAG CP is located at the IP Core.

From these results, we can state that the smooth SIPTO handover solution enables operators to significantly reduce delay during the mobility of UEs having a SIPTO at LN and/or SIPTO above RAN services. This allows seamless session continuity for users having ongoing gaming or video-streaming data taking into consideration that the delay budget for conversational video is estimated by 150 ms and non-conversational video is estimated by 300 ms [77]. Thus, obtained results using our method are fully compliant with the level of quality of service required for this kind of traffic.

6.4.2 Signalling Volume

In the present section we evaluate the signalling volume due to the application of smooth SIPTO handover proposals presented in Chapter 4. The amount of signalling volume of smooth SIPTO handover differs from one scenario to another depending on the following two factors:

- MME function is relocated during the handover or not
- TAU procedure is performed during the handover or not

As presented in Section 6.1.1 and Section 6.1.2, the MME function included within the UAG CP would strongly be relocated when deployed in a distributed COMBO architecture with smooth SIPTO above RAN application. Whereas, this is not the case in a centralized COMBO architecture with smooth SIPTO above RAN. Indeed, in this scenario, having the MME function within the UAG CP very hight in the network (at the Core CO or at the IP core) would reduce the potential relocation of this function.

Moreover, in 3GPP handover scenarios [4], TAU procedure is performed whenever the SGW and/or MME functions are relocated. As a result, when a distributed COMBO architecture with smooth SIPTO handover scenario is deployed, TAU procedure should always be performed. Whereas, it is less frequent to perform the TAU procedure in centralized COMBO architecture with smooth SIPTO handover scenario (e.g., it is less frequent to relocate the SGW and the MME functions).

Compared with the centralized and distributed COMBO architectures with smooth SIPTO above RAN, in centralized and distributed COMBO architectures with smooth SIPTO at LN, where no SGW and/or MME relocation is performed, the TAU procedure is never performed.

Based on 3GPP documentations in ([80], [81], [82], [83], [84] and [85]) and on the values given in [86], we evaluated the signalling volume in handover procedures of smooth SIPTO proposals as follows:

- Signalling volume of the handover procedure of smooth SIPTO for MC2 with TAU and without MME relocation is less than 15 KiloBytes (KB)
- Signalling volume of the handover procedure of smooth SIPTO for MC2 with TAU and with MME relocation is less than 25KB
- Signalling volume of the handover procedure of smooth SIPTO for MC3 without TAU and without MME relocation is less than 19KB

The signalling volume of the handover procedure of smooth SIPTO proposals could then be represented by only few tens of KiloBytes (less than 30 KB). As a result, the location of the MME function in COMBO architectures with smooth SIPTO handover solutions does not affect the signalling volume.

6.5 Application to Advanced-4G and 5G Mobile Architectures

The deployment scenarios of smooth SIPTO handover solutions on future FMC network architectures presented in Sections 6.1.1 and 6.1.2, naturally apply to the new EPC architectures for advanced-4G and 5G [10, 87]. Globally, the present work is related to how future 5G architecture shall be efficiently implemented.

6.5.1 Localizing the UAG DP

With respect to the quantitative evaluation presented in Chapter 3, it is expected that the deployment of centralized COMBO architecture represents an improvement to today's (4G) mobile architecture as it helps saving around 40% of the bandwidth currently used in the Core network.

The centralized COMBO architecture already presents a significant gain in terms of bandwith decrease in the core network. This is due to the fact that in case of centralized COMBO architecture, the SGW and PGW would be located closer to the user than in legacy architectures.

In a scenario where smooth SIPTO proposals are assessed in a distributed COMBO architecture, with application services co-located with the IP edges (SGW/PGW) within the NG-POP at the Main CO, the gain of bandwidth could include both the Core and Aggregation segments of the network.

Finally, deploying distributed COMBO architecture allows a better usage of network resources by saving a major part of bandwidth both for the Core and the Metro networks. Session continuity would be maintained thanks to the smooth SIPTO proposal.

6.5.2 Localizing the UAG CP

With 5G, hundreds of millions of mobile devices will be connected to the network in the next few years. For instance, in France, Orange network has exceeded the one million fiber customer in 2016 by covering over than five million connectible homes with Fiber services [88]. As a result, more than two in three homes in Lyon and Paris become eligible for 100% Fiber. Moreover, by 2022, Orange network is targeting a set of 20 million connectable homes with FTTx services. As the capacity of core nodes is limited, then such a dramatical increase in numbers of devices could lead to scalability issues for mobility control at core network's level. Regarding the centralized COMBO architecture, it is thus not recommended to have UAG CP functionalities implemented very high in the network, i.e. at the IP core level of the FMC network architecture.

We have shown in Sections 6.4.1 and 6.4.2 that both the Main CO and the Core CO levels represent appropriate locations for UAG CP placement for respectively signalling volume levels and traffic interruption delays.

From a network operator's point of view, having the UAG CP functions at the Main CO level would potentially allow a better control of mobility management (e.g., less controlled users per MME). However, such an implementation could possibly require higher CAPEX and OPEX as multiple servers and/or multiple virtual machines would have to be placed closer to the user's location. Moreover, it is well known that, when it comes to data and control function implementations, one of the operators main rules is to "distribute when necessary and centralize whenever possible". Therefore, it seems preferable to implement the UAG CP at core CO rather than at main CO, as long as the degradation in terms of mobility control is acceptable.

6.5.3 Application to 5G Use Cases

In the framework of the FMC broad mobility scenarios described in Section **??**, some are especially relevant, i.e. those defined as 5G use cases.

Indeed, according to ITU in [16] and 3GPP in [89], 5G target use cases include enhanced Mobile Broadband (eMBB) services (UHD, 3D video, 4K, Augmented Reality, etc.), massive Internet of Things (IoT) communications (billios of connected devices) and critical Machine Type Communications (MTC), also known as Ultra-Reliable Low

latency Communications (URLLC), all in frequencies both above and below 6 GHz (see Appendix B). Moreover, 5G network are also expected to provide satisfactory services to users travelling at high speeds mobility, up to 500 km/h.

The deployment of centralized COMBO architecture would particularly allow session continuity even for fast moving users as the UAG DP is implemented high in the network (at the Core CO), which reduces the potential need to change the PGW (i.e., the user's IP address) during a session, and thus limits the risk of traffic interruption.

The deployment of distributed COMBO architecture with SIPTO above RAN would potentially not support session continuity with fast moving users as well as the centralised COMBO architecture since having the UAG DP above the RAN at the Main CO increases, in a limited way, the potential PGW relocation, and thus also the risk of traffic interruption. This would e.g. be the case for a long session (more than one hour) carried out in a moving vehicle, which is probably not the major part of mobile traffic. On the other hand, having the UAG DP co-located together with the IP services at the Main CO would provide users with better QoS such as higher throughput and lower latency. Having IP services co-located with IP edges at the Main CO will thus allow network operators to implement mobile edge computing.

This is even truer for the distributed COMBO architecture with SIPTO at LN. Indeed, the enhancement of the radio access with combined millimetre Wave (mmWave) radio frequencies and massive Multiple-Input Multiple-Output (MIMO), in addition to the deployment of small cells everywhere (femtocells at home/enterprise and picocells within dense and urban areas) will enable 5G network to deliver packets at a very hight data rates (20 Gbps on downlink and 10 Gbps on uplink [16]), which allows 5G to compete with wireline broadband services by 2030 [90]. Taking advandage of these technical developments, SIPTO at LN would thus provide users a very high throughput, and no session continuity issues as long as the user is static, or moves slowly.

5G could also take advantage of distributed COMBO architecture with smooth SIPTO at LN to support massive IoT communications required by e.g. "Wearables" and eHealth applications. This could be achieved by co-locating an IoT service Gateway (IoT-GW) with the widely distributed small cells (femtocells and picocells). Indeed, having IoT-GWs co-located together with femtocells and LGWs/HGW as well as picocells would allow seamless mobility to these specific users. Moreover, in case the sensors deployed for IoT communications were provided with IP, our smooth SIPTO handover solutions would not only provide humans with seamless mobility, but also it would provide any connected mobile device with session continuity.

Finally, when critical communications are considered and ultra-low radio latency becomes indispensable to deliver the IoT data, it is recommended to use a distributed architecture (either smooth SIPTO at LN or smooth SIPTO above RAN) to achieve a minimal end-to-end latency. Indeed, with millions of small cells implemented within a distributed COMBO architectures, backhaul and core delays are minimized, thus the support of critical IoT communications would no more be an issue. This is due to the fact that all of the base stations, IP edges and services are close to the user's location (low radio latencies are achieved thanks to the deployed small cells and low end to end latency communications are achieved thanks to the distributed IP edges and services).

6.6 Conclusion of Chapter 6

In this Chapter, we presented several potential deployment scenarios of smooth SIPTO proposals in candidate FMC network architectures. We showed that FMC mobility can be supported for users that are either connected to a fixed server or connected with other users. We also presented the use case where FMC mobility is not ensured when users are connected to a CDN service, unless the UE does not change its content server. We then evaluated the performance of the smooth SIPTO handover scenarios in terms of interruption time duration and signalling traffic volume. Herein, we proved that our smooth SIPTO handover solutions proposed in Chapter 4 allow seamless mobility to the users by significantly reducing the handover delay compared to the standard 3GPP SIPTO architectures. Besides, the evaluation allows an operator to decide which of the COMBO scenarios represents the best implementation in case of smooth SIPTO deployment, as addressed in the previous Section.

Chapter 7

Conclusion

In the last few years, fixed and mobile networks were subject to a significant traffic increase. For instance, carrying on the same momentum that we have witnessed since the end of the 20th century, mobile data traffic is expected to grow rapidly during the next decade at a CAGR of over 50% [91]. This is due to the increasing bandwidth demand resulting from the evolution of devices and the introduction of UHD services such as 3D video, 4K and 8K, which is expected by 2020. With the advent of these services, operators are enhancing their network capacities by deploying optical access networks (e.g., 10G-PON) on the one hand, and implementing a more distributed network/service architecture on the other hand (e.g., deployment of CDN and IP edges close to the user). Mobile data offloading represents a major step towards a distributed mobile network architecture that could accommodate increasing mobile traffic demands. It potentially facilitates load sharing on different available access technologies such as: WiFi, Femtocells or Macrocells. The offload of mobile data may however increase traffic interruptions whenever the mobile user is on the move.

The present PhD thesis is an attempt to help solving the traffic interruption problems that occur during the mobility of users when a distributed mobile network architecture is deployed. After a detailed description of the existing mobile data offloading approaches in Chapter 2, we have selected the SIPTO solution as it provides its users with access to both local and external IP services using either a Macrocell (eNB) with separated SGW and PGW, a Macrocell (eNB) with co-located SGW and PGW above the RAN or a femtocell (HeNB) co-located with a LGW.

This thesis assesses whether the SIPTO approach is worth being deployed. With this objective, Chapter 3 focuses on evaluating the quantitative gain of bandwidth, in terms of percentage of offloaded traffic, that can be expected assuming that SIPTO offloading technique is actually implemented. With a gain of bandwidth reaching 40% at both the

Core segment and the Metro/Core segment of the mobile network, SIPTO architecture was proven to be worth being deployed.

Like any new technology that is still being standardized, SIPTO has some technical issues. In particular, session continuity has been considered an issue to be solved within 3GPP in [4] and [6], when users with either SIPTO above RAN or SIPTO at LN sessions are on the move. In Chapter 4, we propose a novel method to support seamless mobility for sessions carried by SIPTO connections under the theme of "smooth SIPTO handover solution". Initially, when SIPTO relies on using separated (standalone) SGW and PGW (MC1 in Chapter 2), the users session continuity is supported as long as the PGW is not changed. However, when SIPTO relies on using co-located SGW and PGW (MC2) or on using LGWs (MC3), it is necessary to change the IP address allocated to the UE, an active SIPTO session may then be interrupted. To avoid such problem, in this contribution I introduce a new mechanism to support users mobility without modifying SIPTO approach in itself is considered. The proposed mechanism focuses only on modifying the application of SIPTO during the users mobility. To achieve this goal, it is first required that both the user and the server are MPTCP-capable. Thanks to MPTCP features, it is possible to maintain a single session, initially carried over a given SIPTO connection, and then carried over another SIPTO connection initiated due to the mobility of the UE. The basic idea is, instead of deactivating the initial SIPTO connection, to handover the initial SIPTO traffic, carrying on the same steps performed in 3GPP S1-based handover procedure [4], to the target cell while introducing additional delay to the resource release timer. This delay is required to be equal to the delay necessary to establish a new SIPTO connection. This new connection should carry the initial SIPTO traffic at the end of the handover procedure. I also present detailed description on how MPTCP is blended with SIPTO in both MC2 and MC3 scenarios.

The main point is that the SIPTO 3GPP architectures are only slightly modified (Delaying the Resource Release Timer, Modifying the MME, LGW and HeNB behavior, etc.) while session continuity is ensured by MPTCP running at the end-points (UE and server).

This thesis was performed within the framework of the COMBO project [11]. COMBO focused on fixed and mobile convergence. The work reported in Chapter 5 addresses the functional convergence of fixed and mobile network architectures. In this contribution, we propose potential mapping scenarios of the Classical LTE architecture and the smooth SIPTO architectures (3GPP SIPTO architectures with respect to smooth SIPTO handover solutions proposed in Chapter 4) on COMBO's functional blocks proposed in [12]. Several mapping scenarios of smooth SIPTO architectures on the future

COMBO FMC network topology have been proposed. Indeed, smooth SIPTO handover is one of the solutions proposed by COMBO to facilitate functional FMC.

In Chapter 6, we propose several deployment options for smooth SIPTO handover architectures on both distributed and centralized COMBO scenarios presented in Chapter 5. We then evaluate the applicability of smooth SIPTO handover solution on future COMBO FMC architectures with respect to the interruption time duration and signalling volume. Herein, we showed how the centralized COMBO architecture could be considered as an important evolution to the current Classical LTE architecture (in particular, co-locating the SGW and PGW at the Core CO allow to reap a major part of the benefits identified in Chapter 3). We then showed that for a best applicability of smooth SIPTO approaches, distributed COMBO architecture would be the best choice of implementation. Indeed, when it comes to smooth SIPTO above RAN and smooth SIPTO at LN implementation, it is best to have the UAG DP at the Main CO and the UAG CP at the Core CO. This allows more flexibility and better usage of the network resources. Next, we proved how centralized COMBO architecture can provide mobile users with very-high-speeds mobility support in a context of 5G mobile architecture. Finally, we presented how 5G mobile architecture could take advantage from the deployment of distributed COMBO architecture with respect smooth SIPTO at LN to provide IoT users with seamless mobility (e.g., wearables) and to allow the support of critical IoT communications by co-locating an IoT gateway with the millions distributed smallcells.

Results of this work could be further improved by considering some of the following issues :

- Shall further services require network level handover or could they rely on transport or application level support of mobility?
- In a scenario where the user's session is not served by TCP, can Quick UDP Internet Connections (QUIC) Protocol be used as an alternative to MPTCP to support session continuity? How would it be implemented?
- How does smooth SIPTO, and more generally generalized handover, can be realized thanks to Network Function Virtualization (NFV)? Where should the virtual network functions be implemented? How should they be designed? How would they interact with the MPTCP logics within end-points?
- In a network slicing scenario, often considered for implementing future 5G networks, are smooth SIPTO mechanisms specific to some slices or would they be applicable to all slices?

Appendix A

Sequence Diagrams for Classical LTE Architecture

The following figures represent the sequence diagrams of the main 3GPP procedures for the classical LTE architecture defined in 23.401 [4]. The detailed information used in each diagram could be found in the 3GPP standards 23.401 [4] and 36.300 [45].

The establishment procedure of SIPTO above RAN Connection is shown in Figure A.1 where the following steps are performed:

- When the UE requests the establishment of a SIPTO connection towards a content server, the MME checks the UE's subscription information to verify whether this user has the right for SIPTO access or not. If so, the MME performs a selection function in order to choose a PGW that is close to the UE's location. In Figure A.1 PGW is co-located with the serving gateway "SGW" currently used by the UE for its default data path.
- Now that the target PGW for SIPTO connection has been selected, the MME initiates the creation of the SIPTO Connection by sending a "Create Session Request" message to the SGW including the UE's identity and the selected PGW IP address. The SGW then forwards this request to the selected (co-located) PGW including SGW's TEID and IP address in addition to the UE's identity.
- When the PGW receives the request message, it creates a new entry in its EPS bearer context table and generates a Charging Id for the Default Bearer. The new entry allows the PGW to route user plane between the SGW and the packet data network, and to start charging. Further, the PGW allocates a new IP address to the UE and sends it to the SGW in a "Create Session Response" message. This

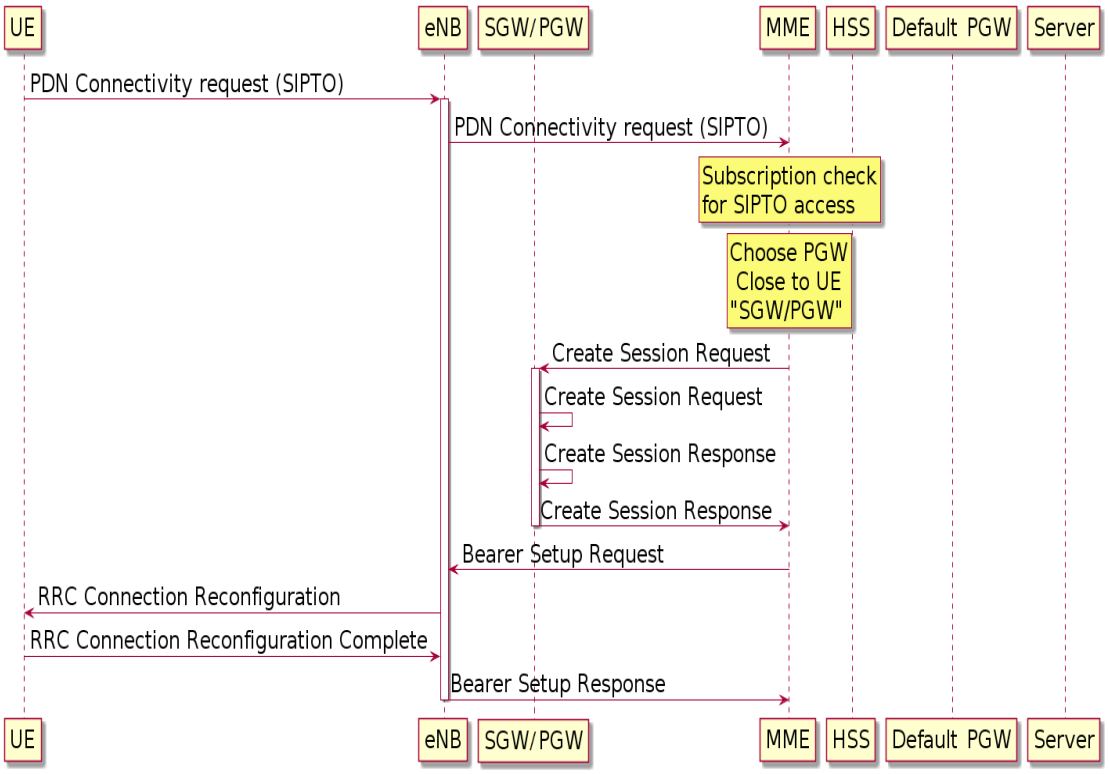


FIGURE A.1: Establishment Procedure of SIPTO above RAN Connection with co-located SGW/PGW

message is then forwarded by the SGW to the MME. This message also serves as an indication to the MME that the GTP-U tunnels over S5 interface are successfully established.

- Next, the MME requests the setup of the SIPTO default EPS bearer to the eNB, which performs an RRC Connection Reconfiguration procedure with the UE.
- Finally, the eNB sends a "Bearer Setup Response" to the MME followed within a "PDN Connectivity Complete" message indicating that now the new SIPTO above RAN Connection is established.

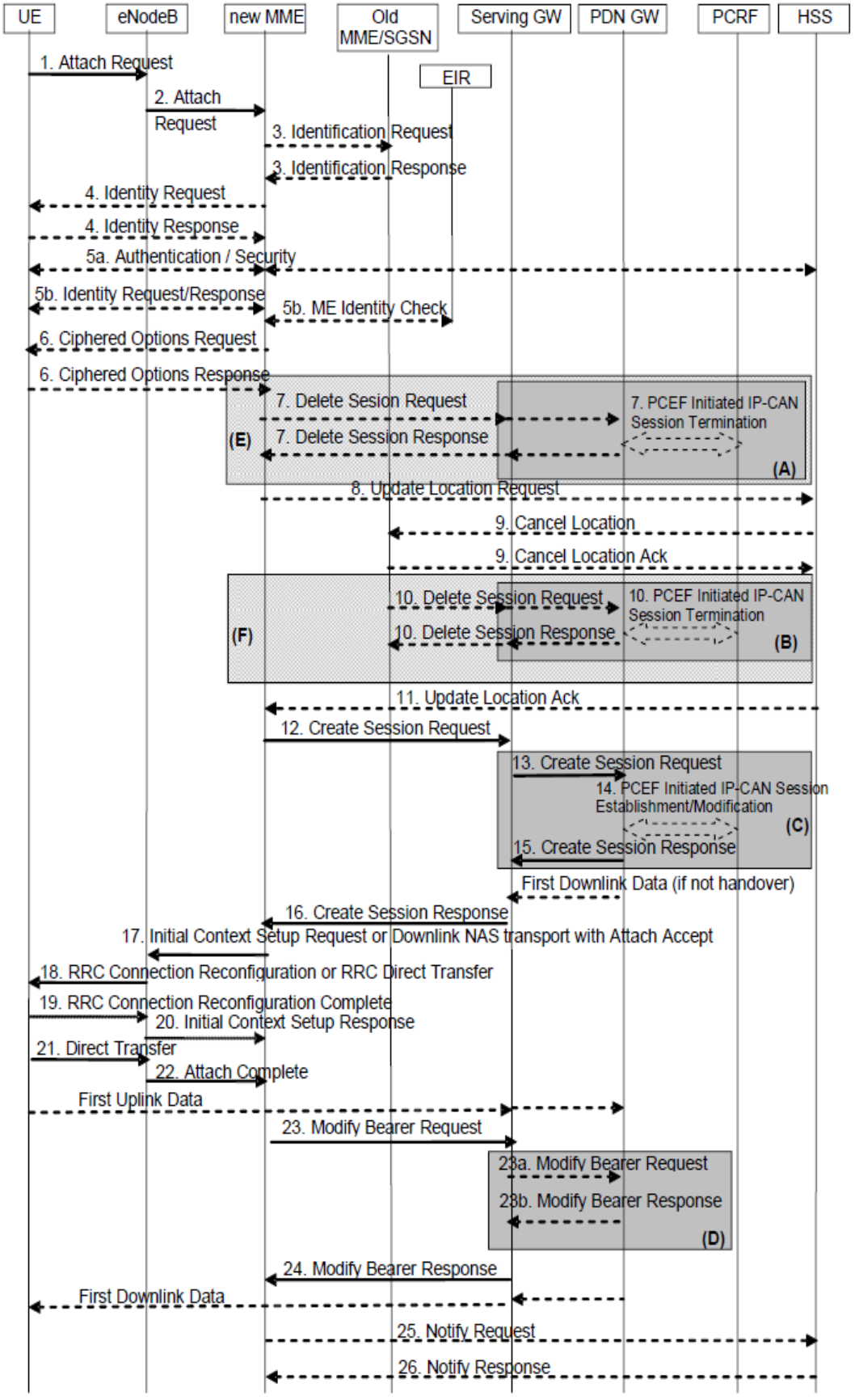


FIGURE A.2: Attach Procedure

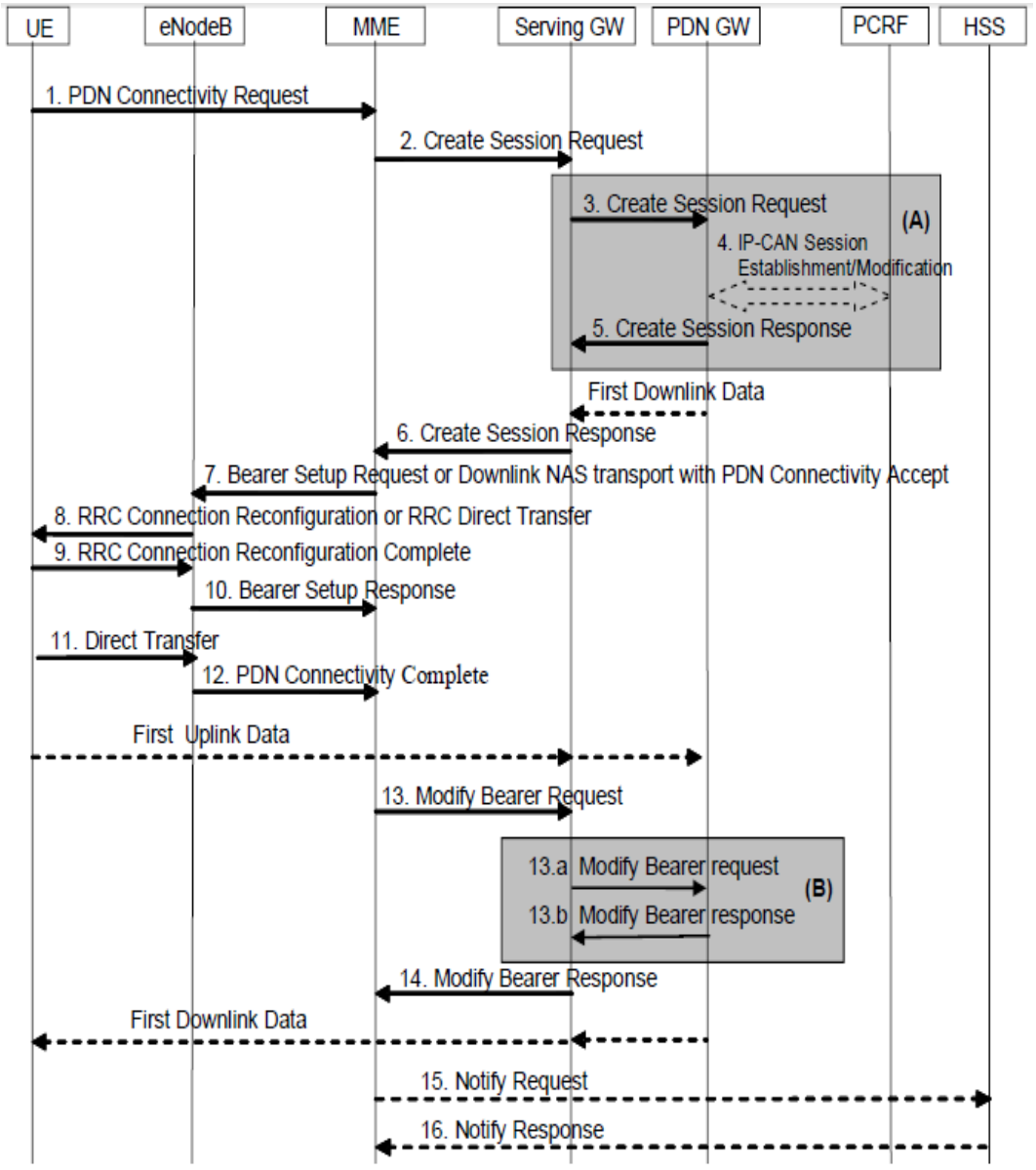


Figure A.3: UE Requested PDN Connectivity Procedure

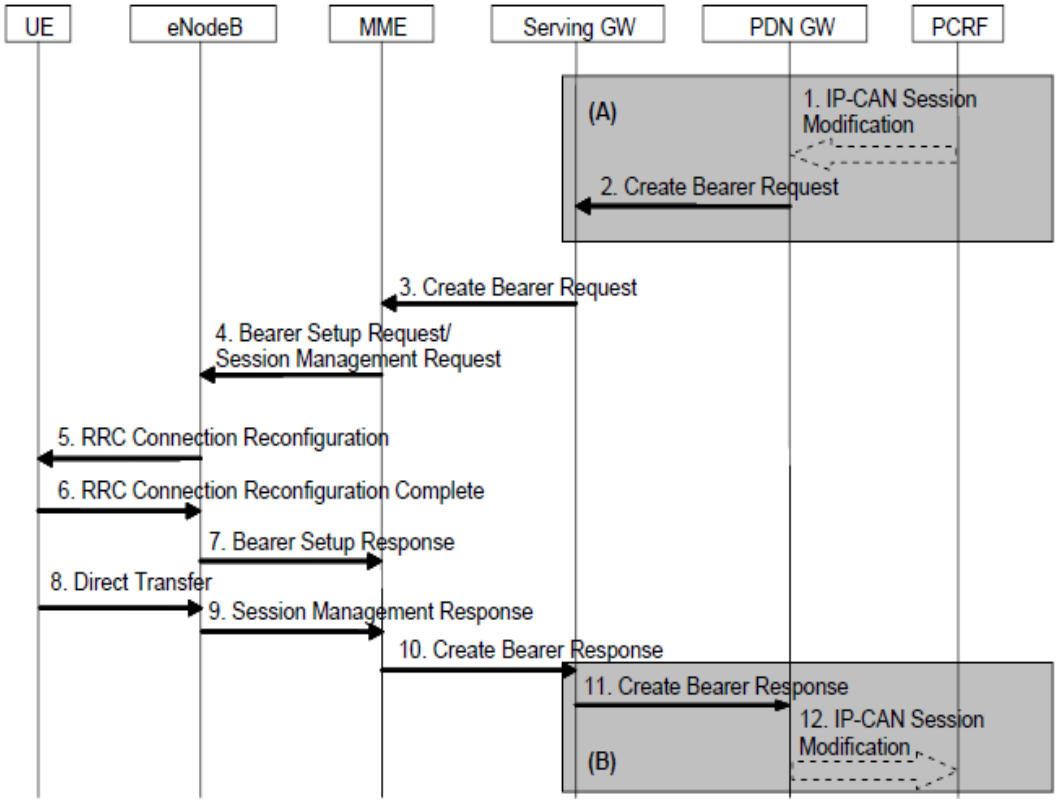


FIGURE A.4: Dedicated Bearer Activation Procedure

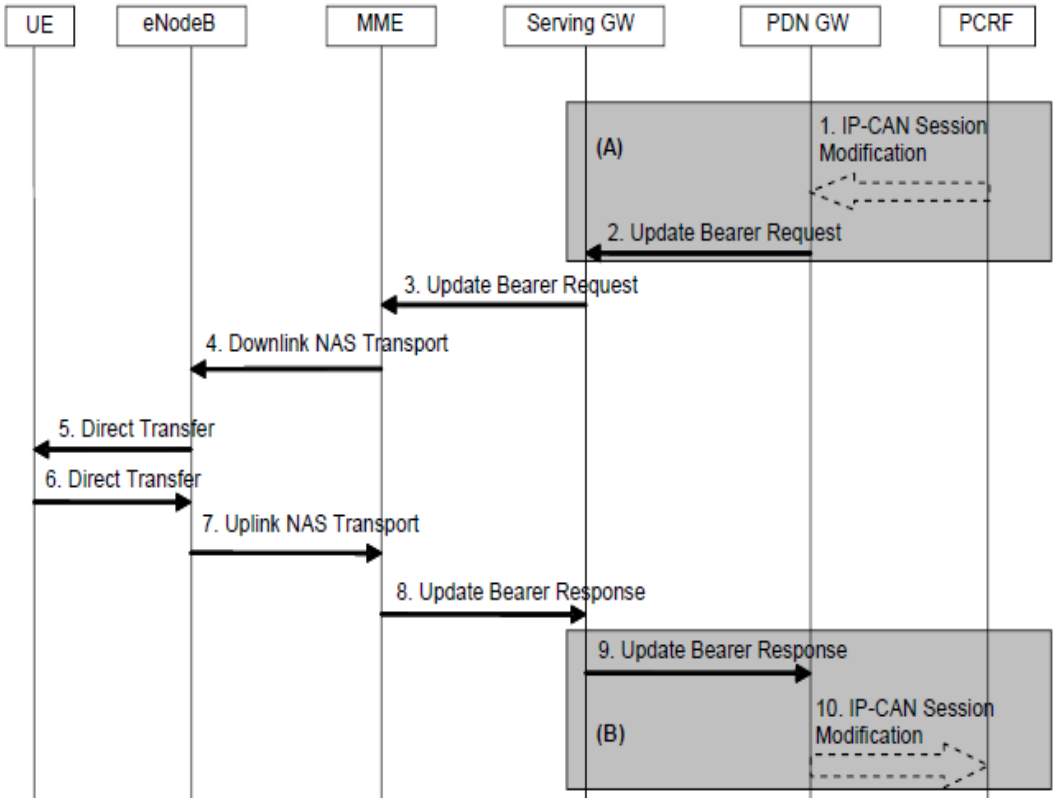


FIGURE A.5: Bearer Modification Procedure without Bearer QoS Update

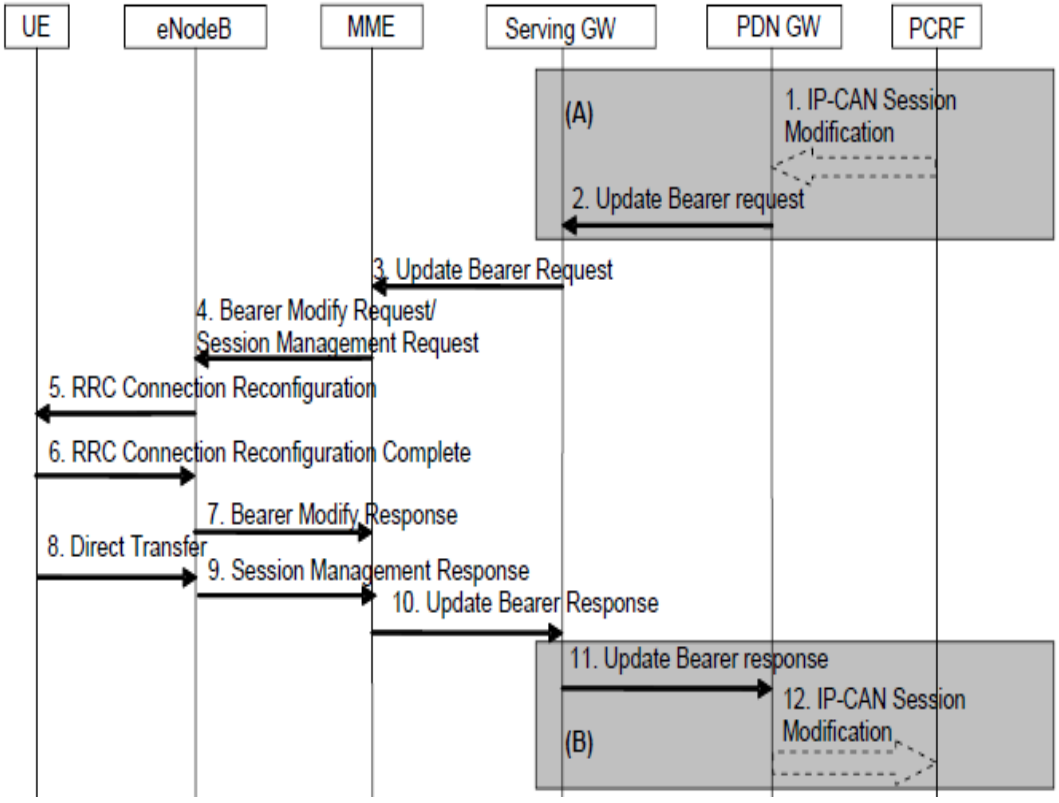


FIGURE A.6: Bearer Modification Procedure with Bearer QoS Update

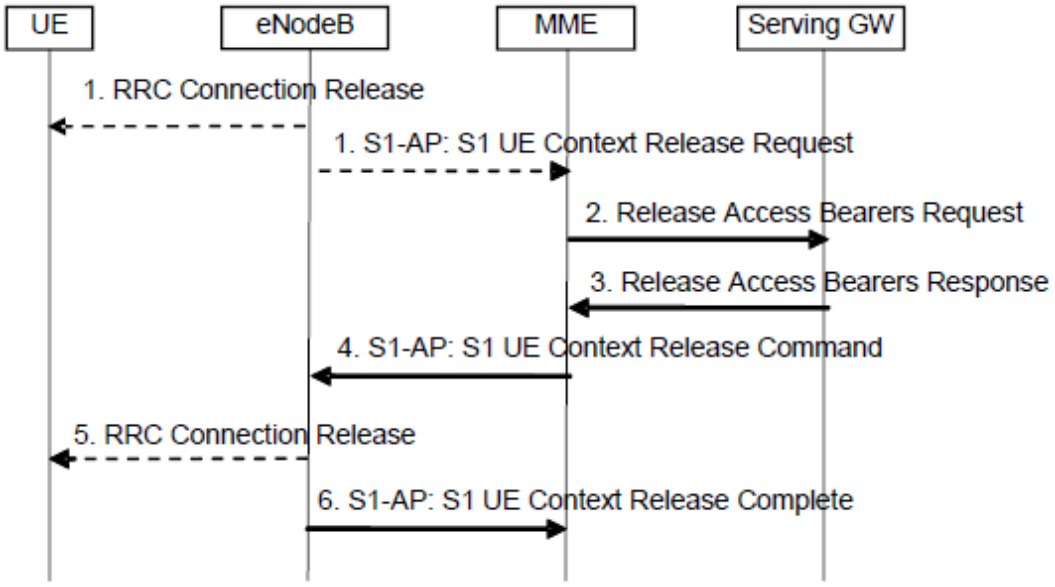


FIGURE A.7: S1 Release Procedure

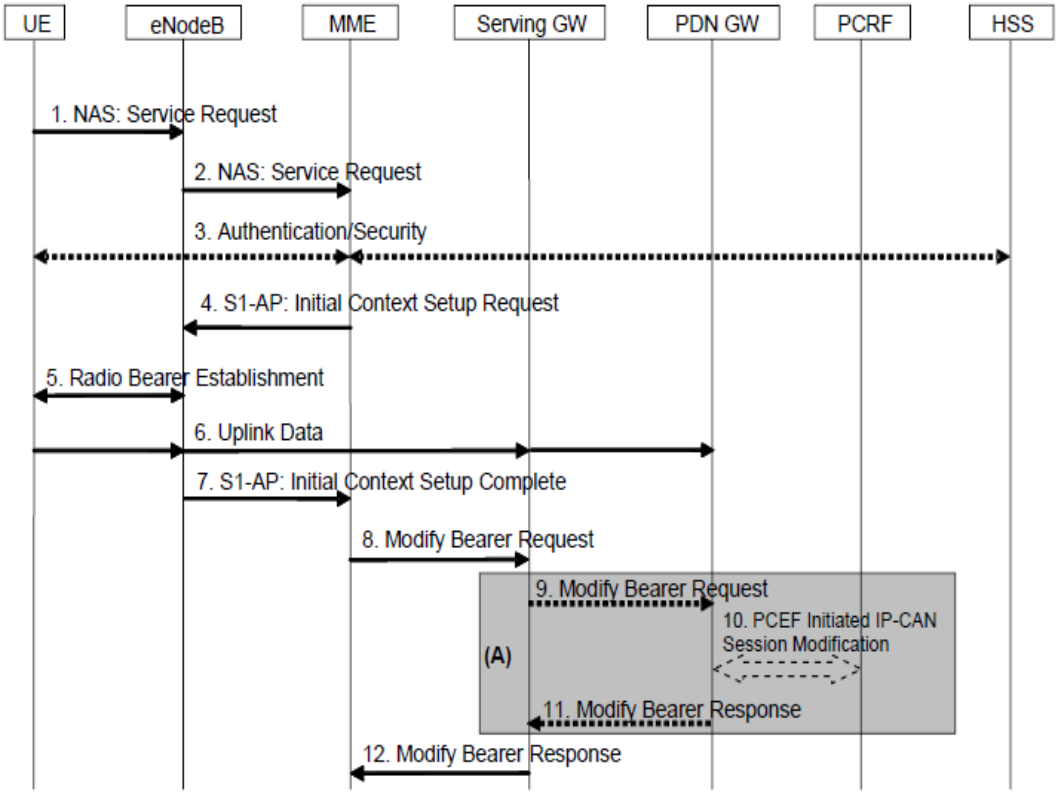


FIGURE A.8: UE triggered Service Request Procedure

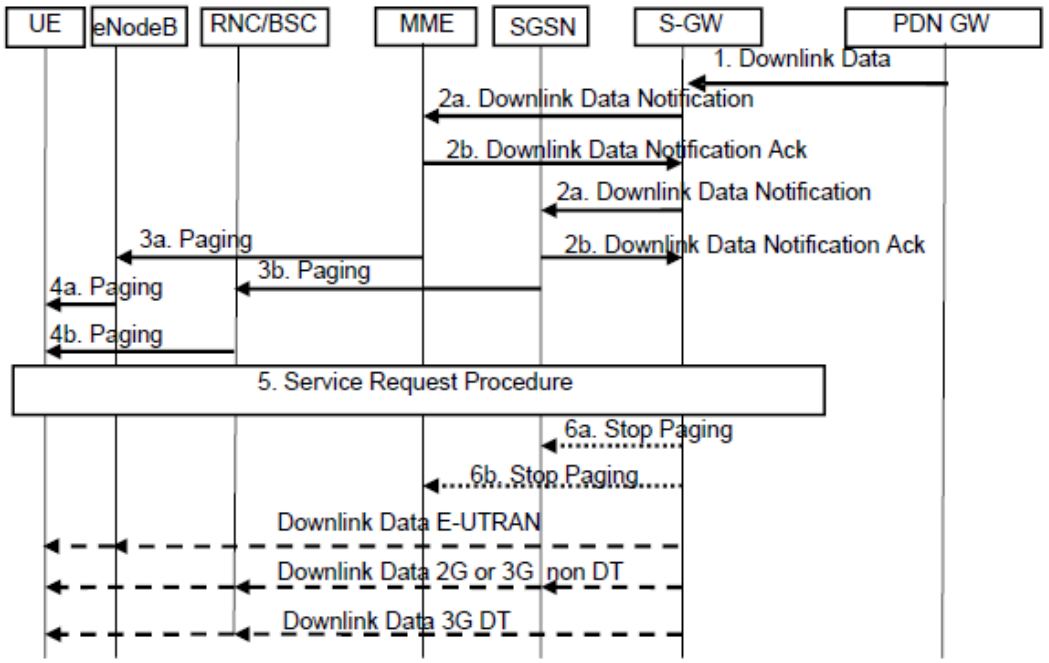


FIGURE A.9: Network triggered Service Request Procedure

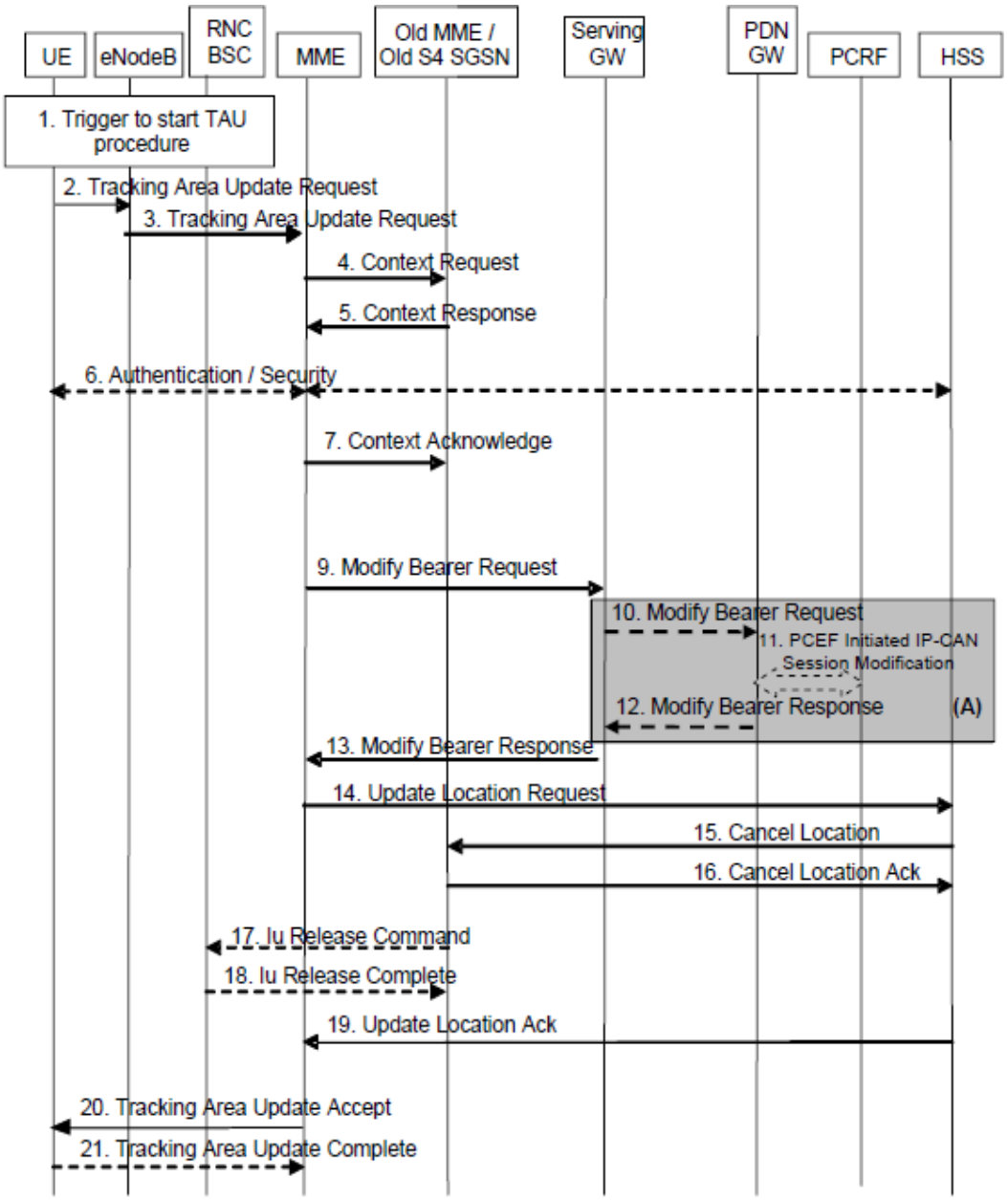


FIGURE A.10: Tracking Area Update Procedure without SGW change

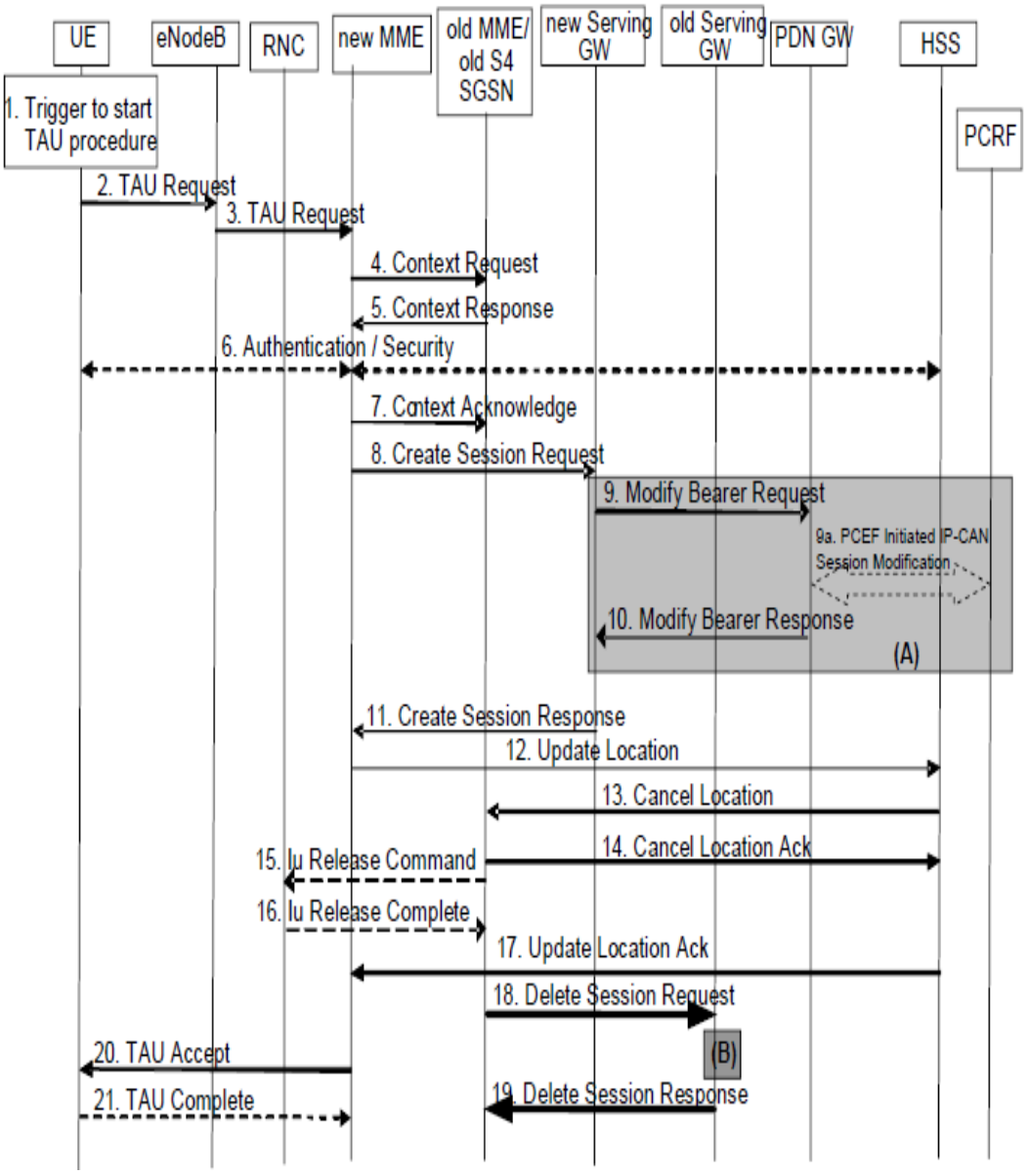


FIGURE A.11: Tracking Area Update Procedure with SGW change

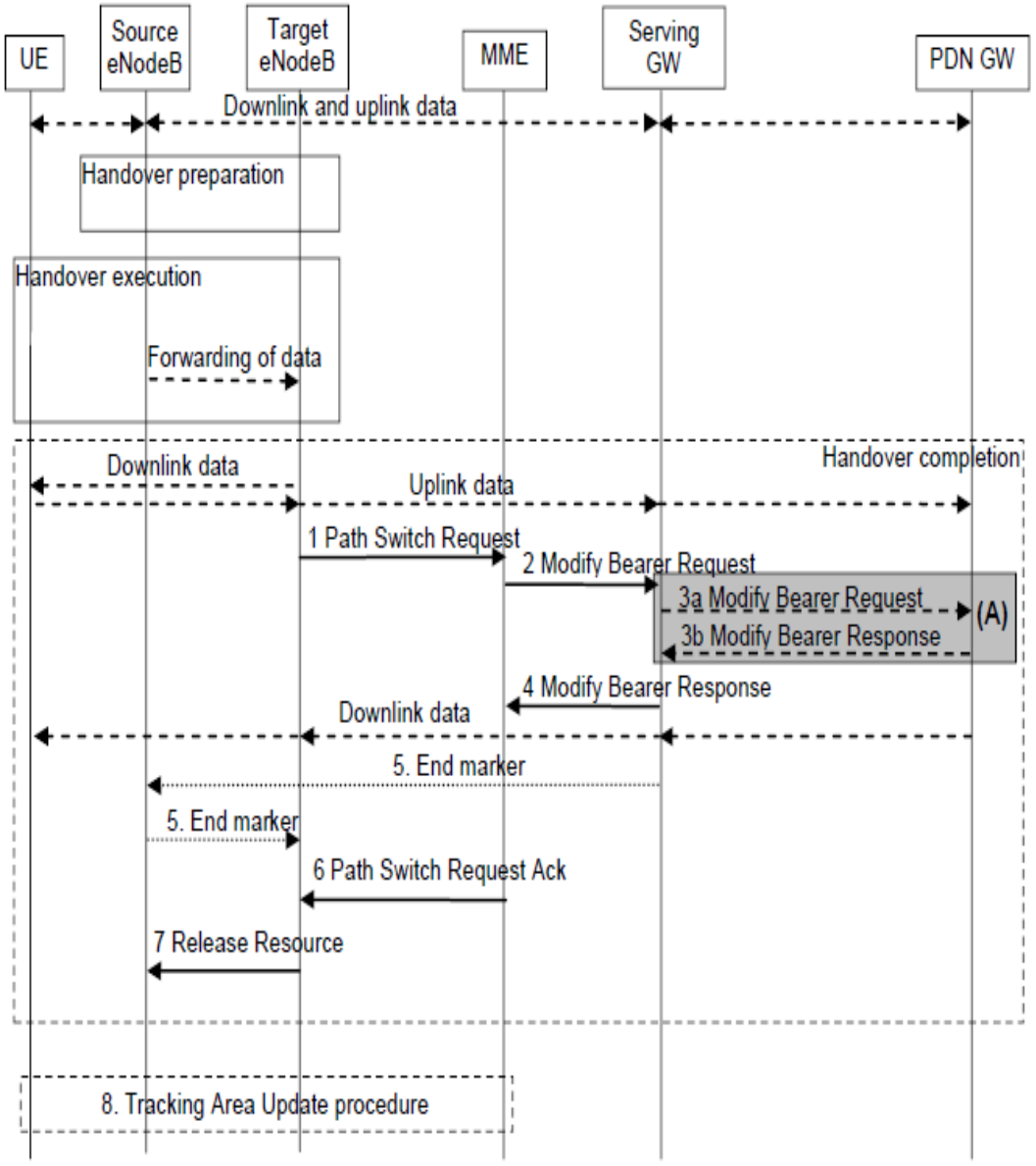


FIGURE A.12: X2-based handover without SGW relocation

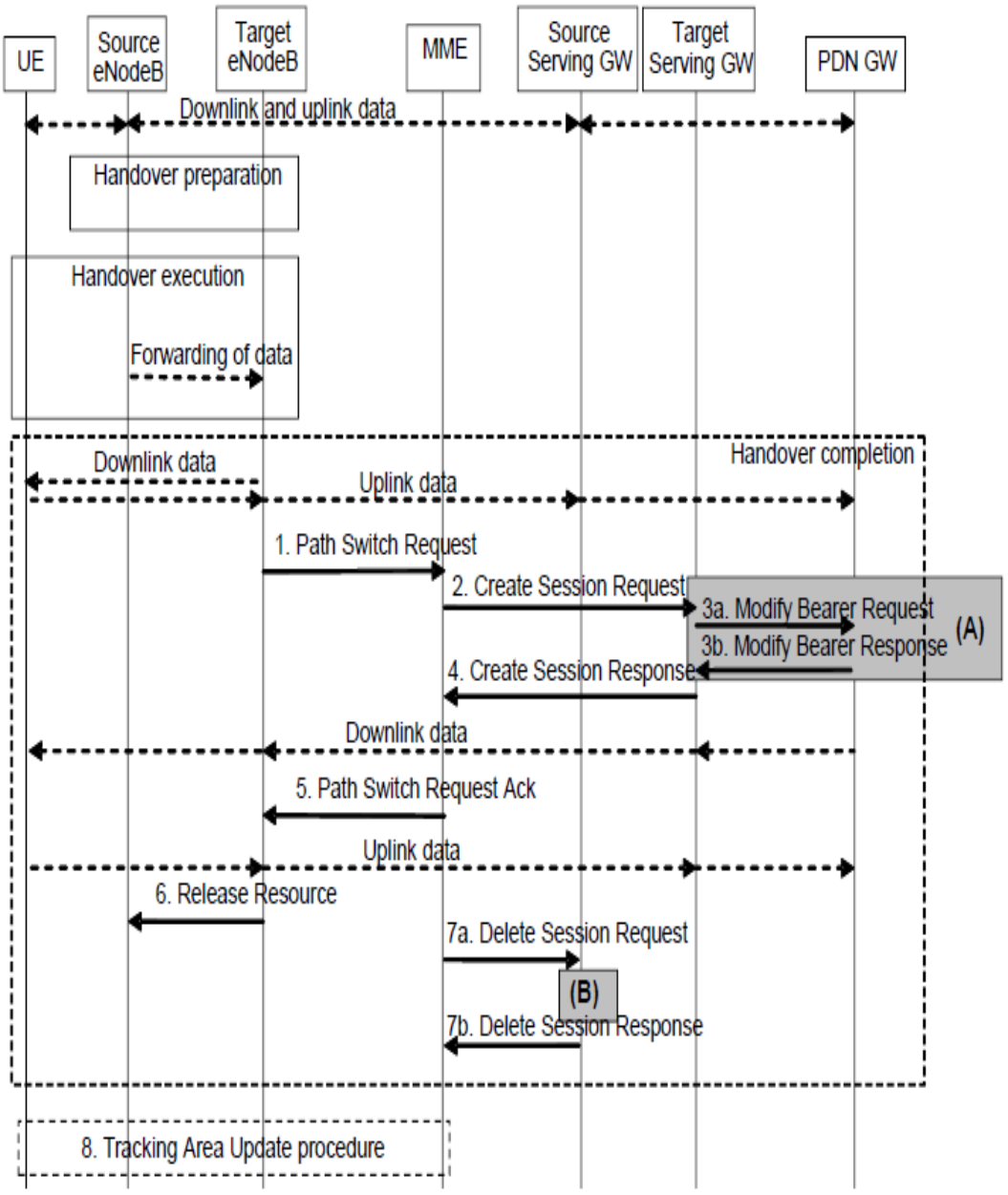


FIGURE A.13: X2-based handover with SGW relocation

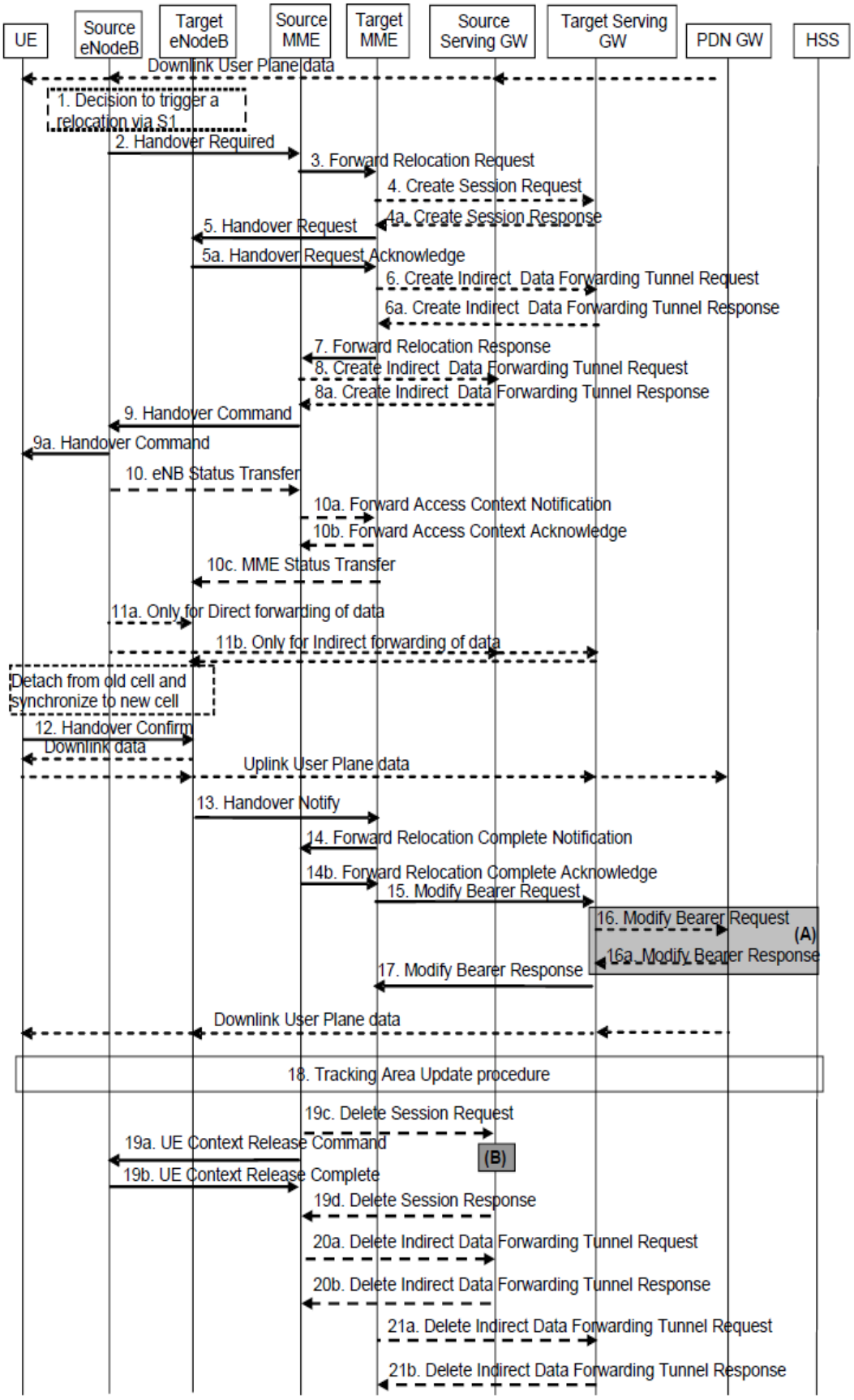


Figure A.14: S1-based handover with SGW relocation

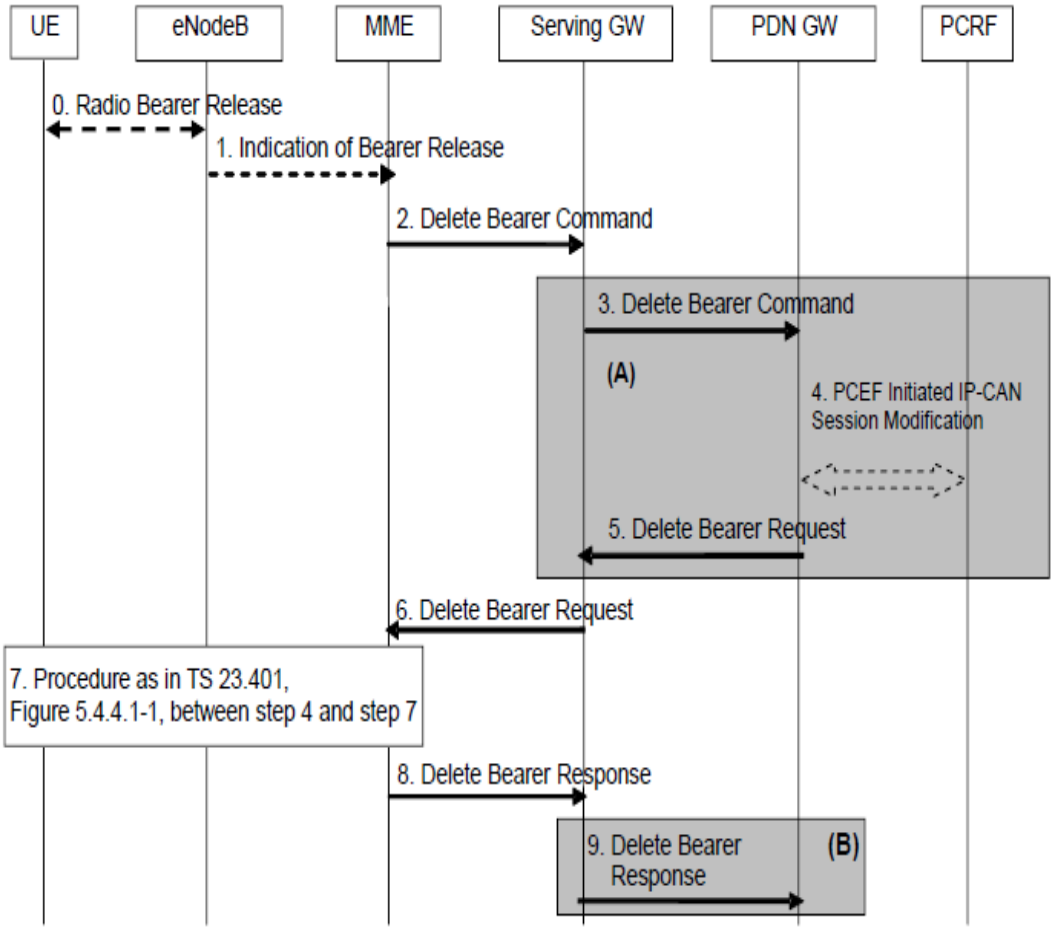


FIGURE A.15: MME initiated Dedicated Bearer Deactivation

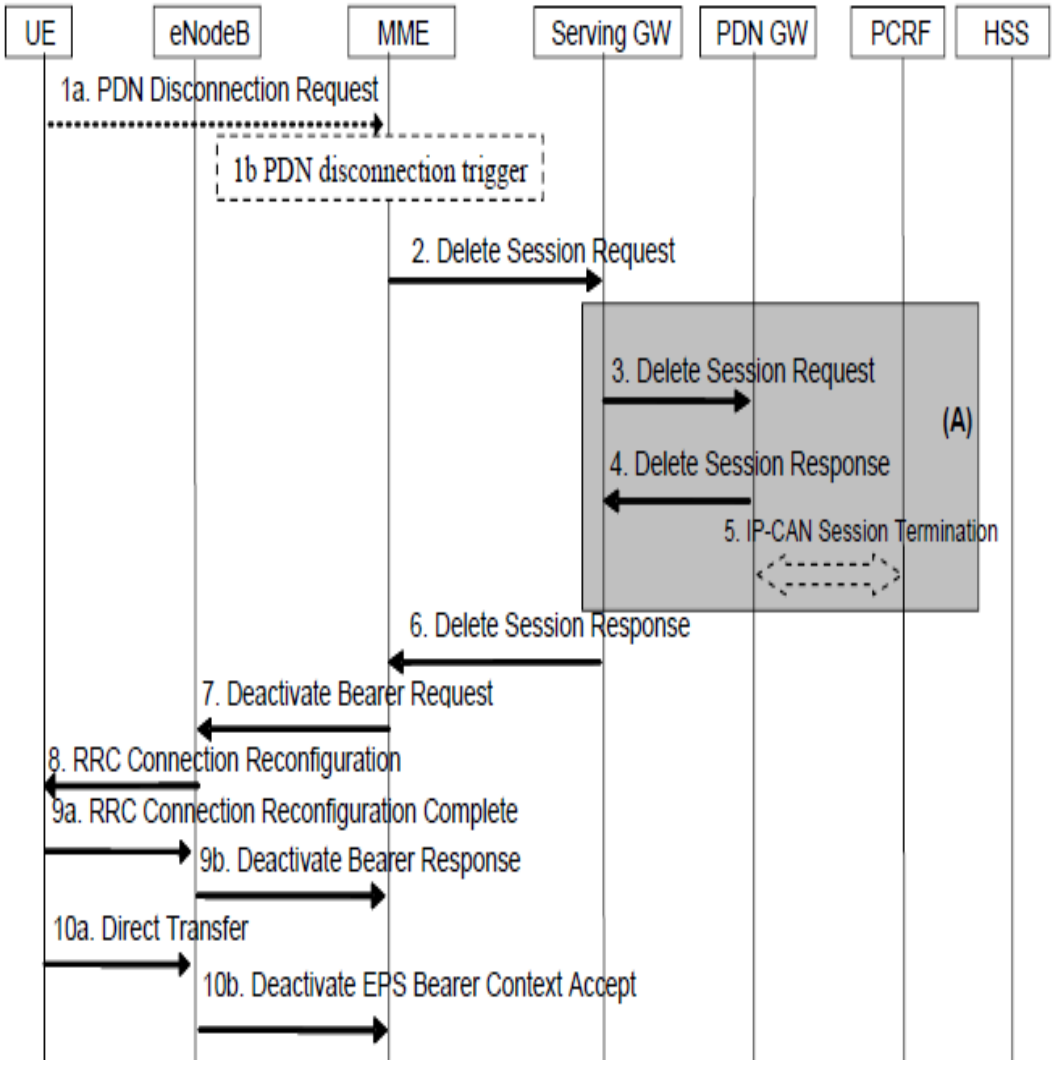


FIGURE A.16: UE or MME requested PDN disconnection

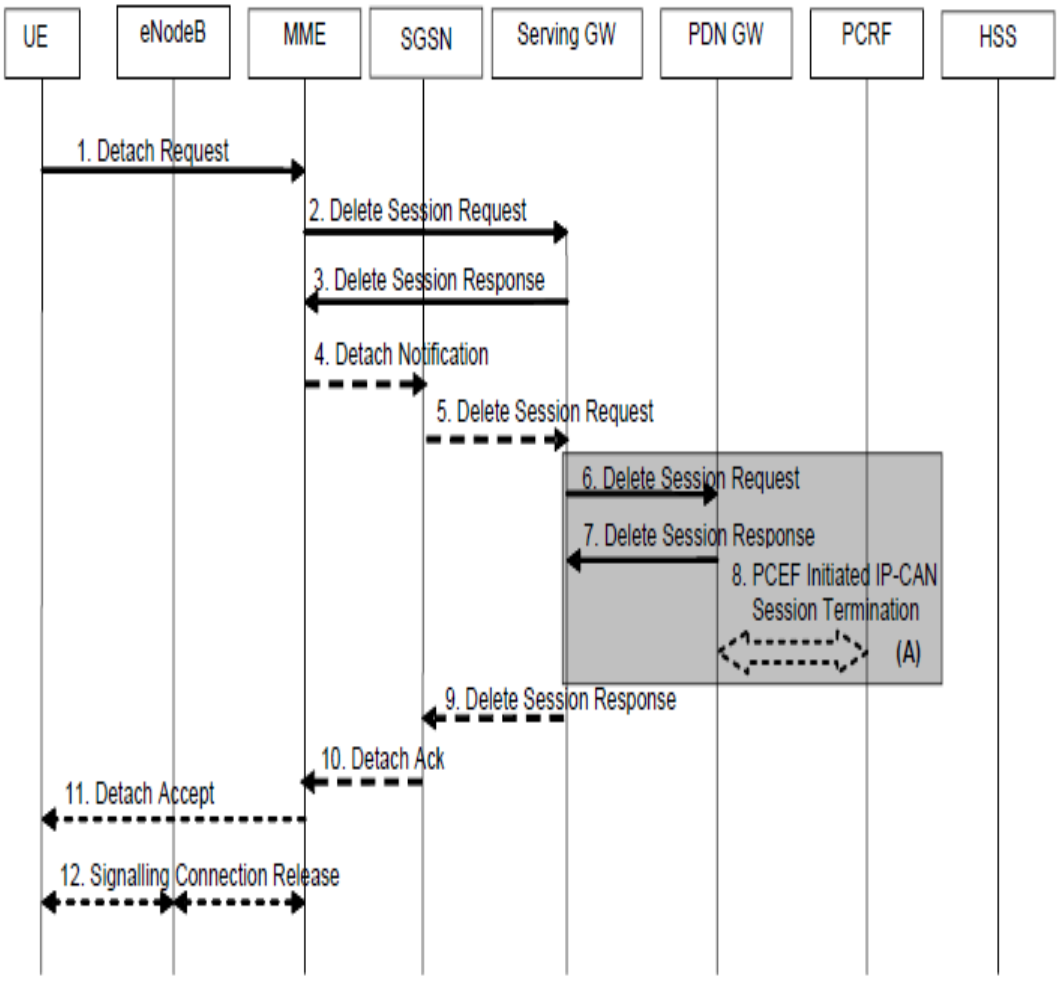


FIGURE A.17: UE-Initiated Detach Procedure - UE camping on E-UTRAN

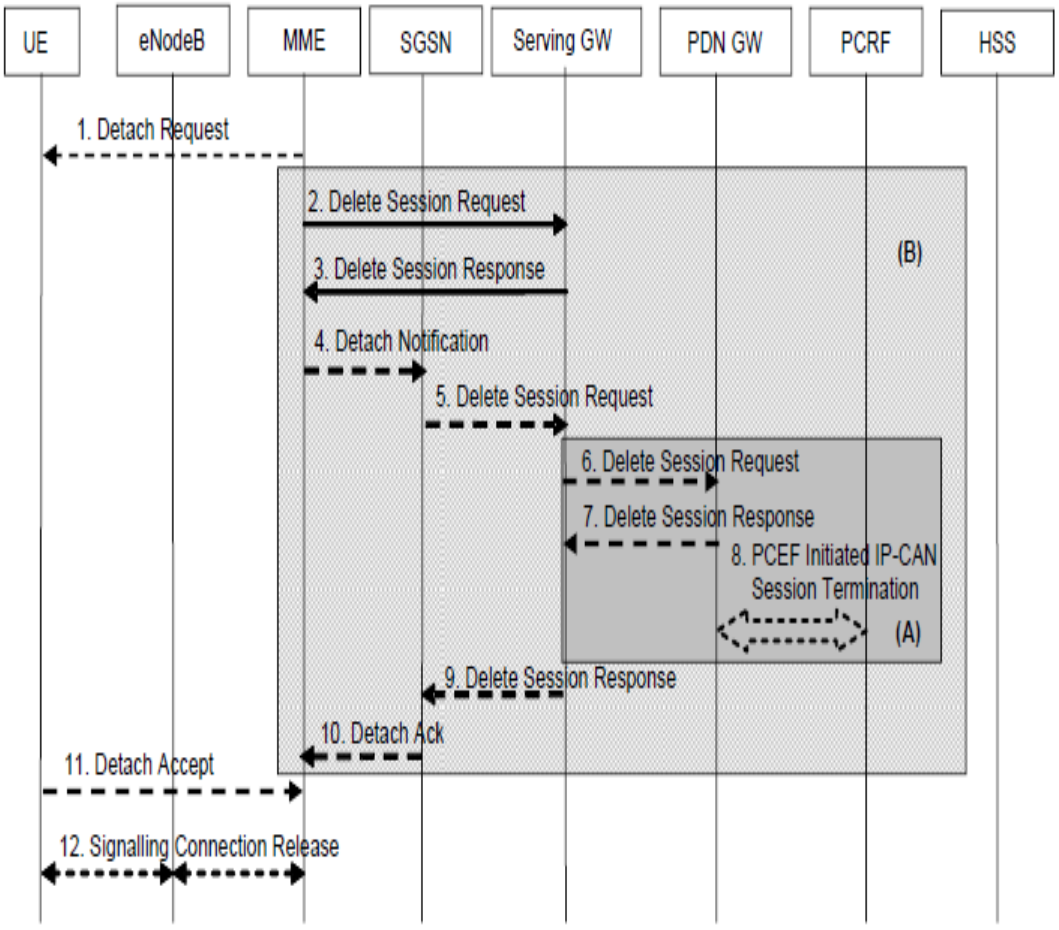


FIGURE A.18: MME-Initiated Detach Procedure

Appendix B

Towards the 5G Mobile Communications System

B.1 Overview of the overall mobile network generations: 1G to 5G

Mobile network's first generation (1G) was originally designed in the early 80's to support voice-only communication using the analogical mode of transmission. When digital mode of transmission evolved in the 1990s, the second generation (2G) of mobile networks, also known as the Global System for Mobile communications (GSM) was then deployed. The digital technology enabled new services such as improved voice through circuit switched data access as well as short text messaging. Compared with 1G, GSM allowed reduced costs and better network performances e.g., higher throughput. However, the 2G systems allowed no data transmission services.

The evolution of data services and the demand for mobile Internet access on the one hand and the saturation of the mobile market in the 2G on the other, had forced Telecom operators to evolve their network towards the third generation of mobile networks (3G), which enabled fast data services through packet switched data access and improved voice capacity. Finally, the recent 4G mobile communications system was deployed to provide mobile-users with high capacity and very high bit rate allowing the support of mobile multimedia services. At the time, 4G system represented a major shift for mobile networks as it changed the overall mobile infrastructure (see Section 2.2.1 of Chapter 2).

With billions of connected mobile users/devices, the mobile industry will process a massive transformation within the next few years, creating vast new capabilities allowing both humans as well as massive-objects communications.

While 5G mobile communications system is not expected by standards bodies (International Telecommunication Union (ITU) and 3GPP) before 2020, some operators have already started deploying a pre-standard 5G infrastructure within their network. As shown in Figure B.1, the early 5G deployments were driven mainly by the exploding demand for mobile broadband and the capacity limit of Radio Access networks.

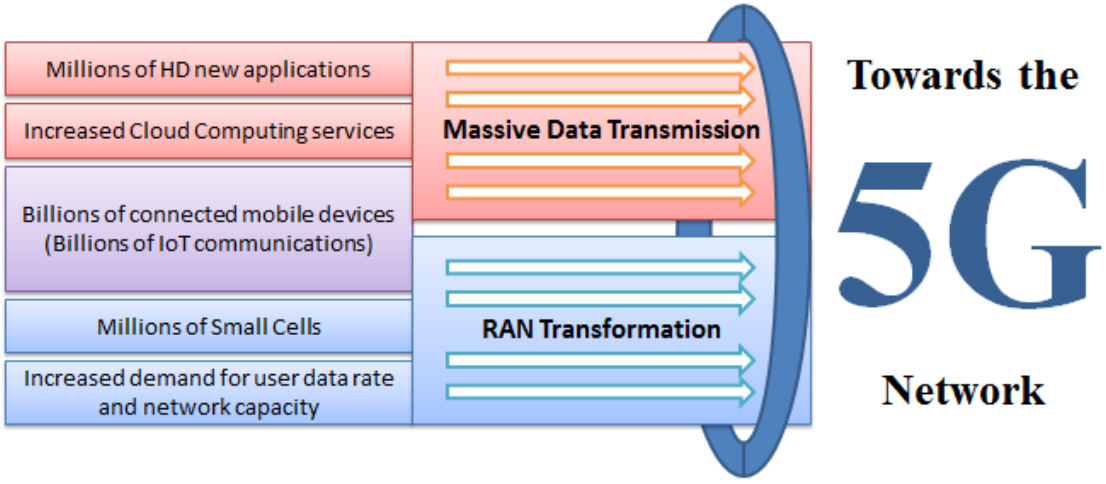


FIGURE B.1: Challenges leading to the deployment of 5G mobile communications

The factors leading to the exploding demand for mobile broadband are, notably:

- the increasing number of smart-phones and tablets that have row capacity (Memory card of up to 128 GB; multiple radio interfaces: 2G to 4G, Bluetooth and WiFi; high-definition screens; high-performance motion cameras; voice recognition, etc.), which allowed application developers to provide users with millions of new mobile Internet-based applications.
- the increasing demand for UHD services (4K, 8K, 3D video, etc.). For instance, in 2022 video traffic will reach 69 EB per month, which will account around 75% of global mobile data traffic [92].
- the introduction of Cloud computing, which allowed users to synchronize and store their data (photo, music, etc.) on the one side and watch videos on streaming thanks to cloud-based applications on the other. As a result, data consumption had increased on both uplink and downlink directions.
- the increasing number of connected devices (tens of billions expected by 2030) along with the evolving of wireless networks with high performance, have inspired

innovators to turn their attention to the IoT, which allows billions of connected wireless devices/machines, leading to a massive increase on mobile signalling and data traffic.

Whereas, the Radio network limitation are most particularly driven by:

- the increasing number of connected devices, which crowded the radio network frequencies (Currently, carriers can only squeeze so mini bits of data on the same radio frequency spectrum). Consequently, the more devices are connected to the network, the more we notice lower speeds, slower services and greater number of interruptions (i.e., more dropped connections).
- the increasing number of small cells (millions by 2030) have increased the risk of Radio-Frequency Interference (RFI) by creating large areas of overlap between cells, which can lead to an increased number in error rate or even a total loss of mobile data.

Typical 5G scenarios and services should include the support of high-traffic communications provided by users in residential or high-density public areas (e.g., Stadium, Subway, etc.), the support of high mobility (e.g., user in high-speed train), providing users with Mobile internet services (e.g., UHD video streaming, Augmented reality, Virtual reality, Cloud storage, etc.), IoT services (e.g., Smart home, Smart grid, Self-driving cars, etc.) as well as immediate sensing and control services e.g., through the realization of the tactile Internet-real-time.

B.2 5G standardization for tomorrow's mobile communications

The ITU, the standardization group for United Nations, has started the "IMT-2020" Project [16] for future 5G mobile communications system, whereas the following use cases were defined as the main usage scenarios to be respected in 5G deployment:

1. **enhanced mobile broadband services (eMBB)**
2. **massive Machine Type Communications (mMTC)**
3. **critical MTC (cMTC) communications**

3GPP group on the side, divided the potential 5G use cases in its "SMARTER" Project [89], into four main categories:

1. **enhanced mobile broadband services (eMBB)** [93]
2. **massive Internet of Things (mIoT)** [94]
3. **Critical communications** [95]
4. **Network Operation** [96]

Figure B.2 illustrates the triangle of performance improvements for the common identified 5G usage scenarios (eMBB, mMTC/mIoT, cMTC).

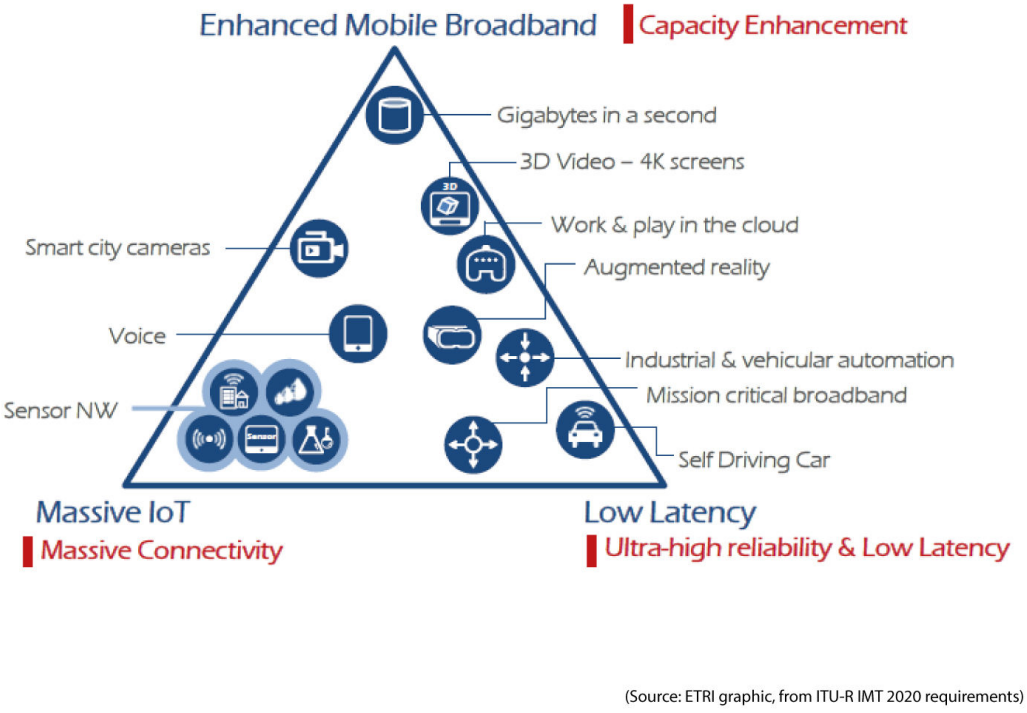


FIGURE B.2: Three Dimensions Usage Scenarios for IMT-2020 and Beyond [16]

The eMBB services represents the most obvious extension of LTE capability, aiming to provide users with very high speeds (data reception and transmission speeds of Gigabits per second)) for applications such as augmented reality, virtual reality, ultra-high-definition, video streaming/conferencing, high mobility support, etc.

The mMTC/mIOT aims to extend LTE Internet of Things capabilities, enabling billions of wireless connections by a wide range of use cases for the IoT including eHealth, wearables, eCity, etc.

The cMTC or critical communications use case focus on extending the number of possible wireless applications by enabling Ultra-reliable and low latency communications such as autonomous driving (self-driving vehicles), remote controlled machines, emergency

applications, etc. These types of ultra-real-time services require a radio latency of less than 1 ms,an end-to-end latency of only few ms.

The Network Operation use case addressed by 3GPP aims to provide users with the same services provided by LTE at better performances and reduced costs using new technologies such as Network Slicing, which allows a network operator to control the selection of the data path according to the user's subscription profile and the required network performances/services.

Moreover, Network virtualization would represent a key technology in future 5G mobile networks. For instance, SDN and NFV would allow operators to deploy software network functions instead of physical network elements, which reduces the CAPEX and OPEX costs and provides more flexibility and better network controlling.

Besides, 3GPP has also defined the Requirements for 5G radio access technology in [97] and [98], whereas a New Radio (NR) access is introduced to the network. The NR addresses very wide bandwidths at millimetre wave spectrum bands above 6 GHz as well as seamlessly combine licensed, shared licensed and unlicensed spectrum improve the radio capacity. The mmWaves can not travel well through buildings or other obstacles, and they attend to be absorbed by plants and rain. To avoid such problem, small cells could be distributed in dense and urban areas. 5G NR will also include advanced antenna techniques such as massive MIMO, which utilize massive arrays of antennas to increase the capacity of RAN, allowing critical IoT communications. However, the deployment of small cells along with massive MIMO could increase the risk of radio interference problem between cells. The solution for the interference issue is the use of adaptive beamforming and beam-tracking technologies within the NR. Beamforming is like a traffic signalling system for cellular signals. It allows base-stations to send a focused stream of data to a specific user instead of broadcasting the signal in every direction. Moreover, 5G will include new narrowband IoT technologies enabling battery operation lasting more than 10 years.

5G technical objectives include:

- Peak data rate of 20 Gbps for downlink, 10 Gbps for uplink
- User Experience data rate of 100 Mbps for downlink and 50 Mbps for uplink
- latency of less than 1 ms for URLLC, 4 ms for eMBB and few ms end-to-end for URLLC
- low energy consumption for long battery life (over 10 years for mMTC)

- Connection dencity of 1 million (10^6) devices per Km^2 for mMTC in urban environment
- Mobility support reaching up to 500 Km/h (high speed train)
- Bandwidth up to 1 GHz
- Reliability of $1 - 10^{-5}$ for URLLC

By the end of 2022, around 550 million 5G subscriptions is predicted by Ericsson in [92].

Bibliography

[1] Cisco Visual Networking Index Cisco. The zettabyte era—trends and analysis, 2015–2020. *white paper, July*, 2016.

[2] Rima Qureshi. Ericsson mobility report. *Ericsson, November*, 2015.

[3] Cisco Visual Networking Index Cisco. Forecast and methodology, 2015–2020. *white paper*, 2016.

[4] 3GPP, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," Technical specification, Release 13, TS 23.401, 2014.

[5] Souheir Eido and Annie Gravey. How much LTE traffic can be offloaded? In *Advances in Communication Networking*, pages 48–58. Springer, 2014.

[6] 3GPP, "Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO)," Technical report, Release 10, TR 23.829, 2011.

[7] 3GPP, "Local IP Access (LIPA) mobility and Selected IP Traffic Overload (SIPTO) at the local network," Technical report, Release 12, TR 23.859, 2013.

[8] "Juniper Networks Fixed Mobile Convergence Solution - Smooth Transitioning to Next-Generation FMC-Based Networks," Juniper Networks, 2009.

[9] Stéphane Gosselin, Anna Pizzinat, Xavier Grall, Dirk Breuer, Eckard Bogenfeld, Sandro Krauß, Jose Alfonso Torrijos Gijón, Ali Hamidian, Neiva Fonseca, and Björn Skubic. Fixed and mobile convergence: Which role for optical networks? *Journal of Optical Communications and Networking*, 7(11):1075–1083, 2015.

[10] J. Son Dr. Harrison and Dr. Michelle M. Do. 5g network as envisioned by kt - analysis of kt's 5g network architecture, 2015.

[11] European FP7 Project COMBO: COnvergence of fixed and Mobile BrOadband access/aggregation networks. http://www.ict-combo.eu.

[12] "Deliverable D3.2: Analysis of horizontal targets for functional convergence," FP7 COMBO Project, Grant Agreement N: 317762, April, 2015.

[13] "Deliverable D3.5: Assessment of candidate architectures for functional convergence," FP7 COMBO Project, Grant Agreement N: 317762, June, 2016.

[14] "Deliverable D3.1: Analysis of key functions, equipment and infrastructures of FMC networks," FP7 COMBO Project, Grant Agreement N: 317762, November, 2013.

[15] Stéphane Gosselin, Tahar Mamouni, Philippe Bertin, Jose Torrijos, Dirk Breuer, Erik Weis, and Jean-Charles Point. Converged fixed and mobile broadband networks based on next generation point of presence. In *Future Network and Mobile Summit (FutureNetworkSummit), 2013*, pages 1–9. IEEE, 2013.

[16] The technical specifications associated with 5G by ITU-T group. http://www.itu.int/fr/ITU-T/focusgroups/imt-2020/Pages/default.aspx.

[17] Alcatel-Lucent. Bell labs metro network traffic growth: An architecture impact study. *Strategic white paper, December*, 2013.

[18] Alan Ford, Costin Raiciu, Mark Handley, and Olivier Bonaventure. Rfc 6824, tcp extensions for multipath operation with multiple addresses, 2013.

[19] Souheir Eido, Pratibha Mitharwal, Annie Gravey, and Christophe Lohr. Mptcp solution for seamless local sipto mobility. In *High Performance Switching and Routing (HPSR), 2015 IEEE 16th International Conference on*, pages 1–6. IEEE, 2015.

[20] Moufida Feknous, Bertrand Le Guyader, and Annie Gravey. Revisiting access and aggregation network architecture. *Journal of advances in computer networks*, 2(3): 163–168, 2014.

[21] Rec ITU-T. G. 987.1: 10-gigabit-capable passive optical networks (xgpon): General requirements. *January*, 2010.

[22] Rec ITU-T. G. 984.1: Gigabit-capable passive optical networks (gpon): General characteristics. *March*, 2008.

[23] Rec CCITT. I. 361 "b-isdn atm layer specification,". *CCITT Document COM XVIII-R*, 34, 1992.

[24] Rec ITU-T. G. 983.3: A broadband optical access system with increased service capability by wavelength allocation: General characteristics. *March*, 2001.

[25] Howard Frazier. The 802.3 z gigabit ethernet standard. *IEEE Network*, 12(3):6–7, 1998.

[26] Gravey Annie Morvan Michel Sadeghioon Lida Gravey, Philippe and Bogdan Uscumlic. Status of timeslotted optical packet-switching and its application to flexible metropolitan networks. In *2ème colloque réseau large bande et internet rapide*. 2011.

[27] Chunming Qiao and Myungsik Yoo. Optical burst switching (obs)–a new paradigm for an optical internetˆ{1}. *Journal of high speed networks*, 8(1):69–84, 1999.

[28] John Y Wei and Ray I McFarland. Just-in-time signaling for wdm optical burst switching networks. *Journal of lightwave technology*, 18(12):2019, 2000.

[29] Myungsik Yoo, Myoungki Jeong, and Chunming Qiao. High-speed protocol for bursty traffic in optical networks. In *Voice, Video, and Data Communications*, pages 79–90. International Society for Optics and Photonics, 1997.

[30] Thomas Legrand, Hisao Nakajima, Paulette Gavignet, Benoit Charbonnier, and Bernard Cousin. Etude numérique de la résolution spectro-temporelle de contention de burst et réalisation d'un noeud obs. In *Journées Nationales d'Optique Guidée*, 2008.

[31] Thomas Legrand, Hisao Nakajima, Paulette Gavignet, and Bernard Cousin. Comparaison de l'obs conventionnel et de l'obs à label. In *Journées Nationales d'Optique Guidée*, 2008.

[32] Dominique Chiaroni. Optical packet add/drop multiplexers for packet ring networks. In *2008 34th European Conference on Optical Communication*, 2008.

[33] Dominique Chiaroni, Gema Buforn, Christian Simonneau, S Etienne, and J-C Antona. Optical packet add/drop systems. In *Optical Fiber Communication Conference*, page OThN3. Optical Society of America, 2010.

[34] Christian Cadéré, Nora Izri, Dominique Barth, Jean-Michel Fourneau, Dana Marinca, and Sandrine Vial. Virtual circuit allocation with qos guarantees in the ecoframe optical ring. In *Optical Network Design and Modeling (ONDM), 2010 14th Conference on*, pages 1–6. IEEE, 2010.

[35] Thaere Eido, Ferhan Pekergin, and Tulin Atmaca. Performance analysis of an enhanced distributed access mechanism in a novel multiservice ops architecture. In *Next Generation Internet Networks, 2009. NGI'09*, pages 1–7. IEEE, 2009.

[36] Indra Widjaja, Iraj Saniee, Randy Giles, and Debasis Mitra. Light core and intelligent edge for a flexible, thin-layered, and cost-effective optical transport network. *IEEE Communications Magazine*, 41(5):S30–S36, 2003.

[37] Iraj Saniee and Indra Widjaja. A new optical network architecture that exploits joint time and wavelength interleaving. In *Optical Fiber Communication Conference*, page TuH4. Optical Society of America, 2004.

[38] Glen Kramer, Marilet De Andrade, Rajesh Roy, and Pulak Chowdhury. Evolution of optical access networks: Architectures and capacity upgrades. *Proceedings of the IEEE*, 100(5):1188–1196, 2012.

[39] Harny Frederique. "principes généraux d'architecture physique du rbci," technical report, internal orange labs document. pages 30–31, 2012.

[40] "Deliverable D2.1: Framework Reference for Fixed and Mobile Convergence," FP7 COMBO Project, Grant Agreement N: 317762, 2014.

[41] Jérôme PONS. Réseaux cellulaires – evolution du système umts vers le système eps. *Techniques de l'ingénieur*, 2013.

[42] NSN JUNIPER. Lte: The trigger for next-gen backhaul. *white paper, August*, 2013.

[43] 3GPP, "Service requirements for the Evolved Packet System (EPS)," Technical Specification, Release 8, TS 22.278, 2009.

[44] 3GPP, "Architecture enhancements for non-3GPP accesses," Technical Specification, Release 14, TS 23.402, 2016.

[45] 3GPP, "Study on co-ordinated Packet data network GateWay (PGW) Change for Selected IP Traffic Offload (C-SIPTO)," Technical report, Release 13, TR 22.828, 2014.

[46] Douglas N Knisely, Takahito Yoshizawa, and Frank Favichia. Standardization of femtocells in 3gpp. *Communications Magazine, IEEE*, 47(9):68–75, 2009.

[47] Cisco Public Information Cisco. Cisco 3g femtocell. *document*, 2010.

[48] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN)," Technical report, Stage 2, Release 12, TS 36.300, 2014.

[49] Cisco Visual Networking Index Cisco. Forecast and methodology, 2012–2017. *white paper*, 2013.

[50] Vikram Chandrasekhar, Jeffrey G Andrews, and Alan Gatherer. Femtocell networks: a survey. *Communications Magazine, IEEE*, 46(9):59–67, 2008.

[51] Yong-hwan Kim, Youn-Hee Han, Min Kim, Yong Seok Park, Sang Jun Moon, Jin Ho Lee, and Dae Kyu Choi. Distributed pdn gateway support for scalable lte/epc networks. In *Consumer Communications and Networking Conference (CCNC), 2014 IEEE 11th*, pages 139–144. IEEE, 2014.

[52] Tarik Taleb and Adlen Ksentini. Follow me cloud: interworking federated clouds and distributed mobile networks. *Network, IEEE*, 27(5):12–19, 2013.

[53] Arijit Banerjee, Xu Chen, Jeffrey Erman, Vijay Gopalakrishnan, Seungjoon Lee, and Jacobus Van Der Merwe. Moca: a lightweight mobile cloud offloading architecture. In *Proceedings of the eighth ACM international workshop on Mobility in the evolving internet architecture*, pages 11–16. ACM, 2013.

[54] "Deliverable D1.5: Updated scenarios, requirements and KPIs for 5G mobile and wireless system with recommendations for future investigations," FP7 METIS Project, Grant Agreement N: 317669, 2015.

[55] Ngmn alliance, ngmn 5g white paper. Technical report, 2015.

[56] Jon Postel. Rfc 792, internet control message protocol. 1981.

[57] Alan Ford, Costin Raiciu, Mark Handley, Sebastien Barre, and Janardhan Iyengar. Architectural guidelines for multipath tcp development. Technical report, 2011.

[58] Nasif Ekiz, Preethi Natarajan, Martin Becke, Michael Tuexen, Thomas Dreibholz, Paul Amer, Randall Stewart, and Jana Iyengar. Load sharing for the stream control transmission protocol (sctp). 2015.

[59] Janardhan R Iyengar, Paul D Amer, and Randall Stewart. Concurrent multipath transfer using sctp multihoming over independent end-to-end paths. *IEEE/ACM Transactions on networking*, 14(5):951–964, 2006.

[60] Christoph Paasch, Gregory Detal, Fabien Duchene, Costin Raiciu, and Olivier Bonaventure. Exploring mobile/wifi handover with multipath tcp. In *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, pages 31–36. ACM, 2012.

[61] C. Paasch, S. Barre, et al. Multipath TCP in the Linux Kernel. http://www.multipath-tcp.org.

[62] Lingli, Deng and Dapeng, Liu and Tao, Sun and Mohamed, Boucadair and Gregory, Cauchie. Use-cases and requirements for mptcp proxy in isp networks, 2014.

[63] Lyndon Ong. Rfc 3286, an introduction to the stream control transmission protocol (sctp). 2002.

[64] Martin Becke, Hakim Adhari, Erwin P Rathgeb, Fu Fa, Xiong Yang, and Xing Zhou. Comparison of multipath tcp and cmt-sctp based on intercontinental measurements. In *Global Communications Conference (GLOBECOM), 2013 IEEE*, pages 1360–1366. IEEE, 2013.

[65] Markus Hofmann and Leland R Beaumont. *Content networking: architecture, protocols, and practice*. Elsevier, 2005.

[66] Novella Bartolini, Emiliano Casalicchio, and Salvatore Tucci. A walk through content delivery networks. In *Performance Tools and Applications to Networked Systems*, pages 1–25. Springer, 2004.

[67] Athena Vakali and George Pallis. Content delivery networks: Status and trends. *IEEE Internet Computing*, 7(6):68–74, 2003.

[68] Fred Douglis and M Frans Kaashoek. Guest editors' introduction: Scalable internet services. *IEEE Internet Computing*, 5(4):36–37, 2001.

[69] Bin Wu and Ajay D Kshemkalyani. Objective-optimal algorithms for long-term web prefetching. *Computers, IEEE Transactions on*, 55(1):2–17, 2006.

[70] Yan Chen, Lili Qiu, Weiyu Chen, Luan Nguyen, and Randy H Katz. Efficient and adaptive web replication using content clustering. *Selected Areas in Communications, IEEE Journal on*, 21(6):979–994, 2003.

[71] F. Guillemin A. Gravey and S. Moteau. Last mile caching of video content by an isp. In *ETS 2013: 2nd European Teletraffic Seminar*, 2013.

[72] T. Wright J. Young and N. Temple. "orange and akamai form content delivery strategic alliance,", 2012.

[73] Z. Li, M. K. Sbai, Y. Hadjadj-Aoul, D. Alliez, G. Simon, K. D. Singh, G. Madec, J. Garnier, and A. Gravey. Network Friendly Video Distribution. In *NoF 2012 : 3rd International Conference on the Network of the Future*, 2012.

[74] 3GPP, "Vocabulary for 3GPP Specifications," Technical report, Release 13, TR 21.905, 2016.

[75] J. Son Harrison. Korea communication review, q4 2015. Technical report, Technical Review, Netmanias Consulting, 2015.

[76] Georg Hampel, Anil Rana, and Thierry Klein. Seamless tcp mobility using lightweight mptcp proxy. In *Proceedings of the 11th ACM international symposium on Mobility management and wireless access*, pages 139–146. ACM, 2013.

[77] Z Savic. "LTE Design and Deployement Strategies," Cisco, 2011.

[78] R. Gupta and N. Rastogi. "LTE Advanced – LIPA and SIPTO," White Paper, 2012.

[79] 3GPP, "Feasibility study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN)," Technical report, Release 13, TR 25.912, 2015.

[80] 3GPP, "Evolved General Packet Radio Service (GPRS) Tunnelling Protocol for Control plane (GTPv2-C); Stage 3," Technical Specification, Release 14, TS 29.274, 2016.

[81] 3GPP, "Mobile radio interface signalling layer 3; General aspects," Technical Specification, Release 14, TS 24.007, 2017.

[82] 3GPP, "Numbering, addressing and identification," Technical Specification, Release 14, TS 23.003, 2017.

[83] 3GPP, "Mobile radio interface Layer 3 specification; Core network protocols; Stage 3," Technical Specification, Release 14, TS 24.008, 2017.

[84] 3GPP, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 Application Protocol (S1AP)," Technical Specification, Release 14, TS 36.413, 2017.

[85] 3GPP, "Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3," Technical Specification, Release 14, TS 24.301, 2017.

[86] UMTS ASN1 RANAP description. http://niviuk.free.fr/ranap.html.

[87] J. Son Harrison and M. Do Michelle. 5G network as envisioned by KT - Analysis of KT's 5G network architecture. http://www.netmanias.com/en/?m=view&id=blog&no=8256&xtag=5g-c-ran-fronthaul-kt-korea&xref=5g-network-as-envisioned-by-kt-analysis-of-kt-s-5g-network-architecture.

[88] Orange Fiber: already one million customers strong in France. http://www.orange.com/en/Press-and-medias/press-releases-2016/Orange-Fiber-already-one-million-customers-strong-in-France.

[89] 3GPP, "Study on New Services and Markets Technology Enablers," Technical Report, Release 14, TR 22.891, 2016.

[90] Nokia. 5g use cases and requirements, white paper. Technical report, 2016.

[91] Cisco Visual Networking Index Cisco. Global mobile data traffic forecast update, 2015–2020. *white paper*, 2016.

[92] Patrik Cerwall. Ericsson mobility report. *Ericsson, November*, 2016.

[93] 3GPP, "Feasibility study on new services and markets technology enablers for enhanced mobile broadband; Stage 1," Technical Report, Release 14, TR 22.863, 2016.

[94] 3GPP, "FS_SMARTER - massive Internet of Things," Technical Report, Release 14, TR 22.861, 2016.

[95] 3GPP, "Feasibility study on new services and markets technology enablers for critical communications; Stage 1," Technical Report, Release 14, TR 22.862, 2016.

[96] 3GPP, "Feasibility study on new services and markets technology enablers for network operation; Stage 1," Technical Report, Release 14, TR 22.864, 2016.

[97] 3GPP, "Study on channel model for frequency spectrum above 6 GHz," Technical Report, Release 14, TR 38.900, 2017.

[98] 3GPP, "Study on new radio access technology," Technical Report, Release 14, TR 38.912, 2017.

Résumé

Les réseaux fixes et mobiles font face à une croissance dramatique du trafic de données, qui est principalement due à la distribution de contenus vidéo. Les opérateurs Télécoms envisagent donc de décentraliser la distribution de contenus dans les futures architectures convergées fixe-mobile (FMC). Cette décentralisation, conjointement au déploiement d'un cœur de réseau mobile distribué, sera un élément majeur des futurs réseaux 5G. L'approche SIPTO définie par 3GPP permet déjà le délestage sur le réseau fixe du trafic mobile, et pourra donc être utilisée en 5G. SIPTO s'appuie sur la distribution des passerelles de données (PGW) qui permet ainsi de décharger le cœur du réseau mobile actuel. Cependant, dans certains cas de mobilité des usagers, SIPTO ne supporte pas la continuité de session, quand il est nécessaire de changer de PGW, donc de modifier l'adresse IP du terminal.

Cette thèse commence par quantifier le gain apporté par le délestage du trafic mobile en termes de capacité requise pour différentes portions du réseau. Un état de l'art des différentes solutions de délestage du trafic de données mobiles est fourni, démontrant qu'aucune des solutions existantes ne supporte la continuité de service pour les sessions de longue durée. C'est pourquoi, cette thèse propose des solutions pour supporter une mobilité transparente ; ces solutions s'appuient à la fois sur SIPTO et sur le protocole MultiPath TCP (MPTCP). Les protocoles du 3GPP sont inchangés car la continuité est maintenue par les extrémités. Enfin, ces solutions sont appliquées aux différentes implémentations d'architectures FMC envisagées à ce jour.

Mots clés : Réseaux mobiles 5G, Réseau LTE, FMC, Distribution de données, Délestage du trafic mobile, SIPTO, MPTCP

Abstract

Fixed and mobile networks are currently experiencing a dramatic growth in terms of data traffic, mainly driven by video content distribution. Telecoms operators are thus considering de-centralizing content distribution architecture for future Fixed and Mobile Converged (FMC) network architectures. This decentralization, together with a distributed mobile EPC, would be used for future 5G networks. Mobile data offloading, in particular SIPTO approaches, already represent a good implementation model for 5G network as it allows the use of distributed IP edges to offload Selected IP traffic off the currently centralized mobile core network. However, in some cases, SIPTO does not support session continuity during users' mobility. This is due to the fact that user's mobility may imply packet gateway (PGW) relocation and thus a modification of the UE's IP address.

This PhD thesis first quantifies the gain, in terms of bandwidth demands on various network portions, brought by the generalized use of mobile traffic offloading. A state of art of existing mobile data offloading solutions is presented, showing that none of the existing solutions solve the problem of session continuity for long-lived sessions. This is why, in the context of future FMC mobile network architectures, the PhD thesis proposes solutions to provide seamless mobility for users relying on SIPTO with the help of Multipath TCP (MPTCP). 3GPP standards are not modified, as session continuity is ensured by end-points. Lastly, the proposed solutions are mapped on different architecture options considered for future FMC networks.

Keywords: 5G Network, LTE, FMC, Content distribution, Mobile data offloading, SIPTO, MPTCP

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

N° d'ordre : 2017IMTA0025

IMT Atlantique Bretagne-Pays de la Loire - www.imt-atlantique.fr

Campus de Brest
Technopôle Brest-Iroise
CS 83818
29238 Brest Cedex 03
T +33 (0)2 29 00 11 11
F +33 (0)2 29 00 10 00

Campus de Nantes
4, rue Alfred Kastler - La Chantrerie
CS 20722
44307 Nantes Cedex 3
T +33 (0)2 51 85 81 00
F +33 (0)2 51 85 81 99

Campus de Rennes
2, rue de la Châtaigneraie
CS 17607
35576 Cesson Sévigné Cedex
T +33 (0)2 99 12 70 00
F +33 (0)2 99 12 70 08