Deliverable D3.5: Assessment of candidate architectures for functional convergence

Grant Agreement number: 317762

Project acronym: COMBO

Project title: COnvergence of fixed and Mobile BrOadband access/aggregation networks

Funding Scheme: Collaborative Project – Integrated Project

Date of latest version of the Deliverable: July 7, 2016

Delivery Date: May 31, 2016

Leader of the Deliverable: D3.5

File Name: D3.5: Assessment of candidate architectures for functional convergence

Version: V1.0

Authorisation code: PU = *Public*

Project coordinator name, title and organisation: Jean-Charles Point, JCP-Connect

Tel: + 33 2 23 27 12 46

E-mail: pointjc@jcp-connect.com

Project website address: www.ict-combo.eu

Executive Summary of the Deliverable

Description of the Deliverable content and purpose

The purpose of Work Package 3 “Fixed Mobile Convergent Architectures” is to propose, define and technically assess candidate architectures for future Fixed-Mobile Convergent (FMC) networks, both in terms of data plane (DP) and control plane (CP). This deliverable D3.5 technically assesses candidate architectures for functional convergence. It leverages on the previous work of WP3, in particular on the development by deliverable D3.2 of key functional blocks for FMC, the universal subscriber and user Authentication (uAUT) and the universal Data Path Management (uDPM).

Based on uAUT and uDPM functional blocks, this deliverable D3.5 describes, analyses and compares alternative architectures for functional convergence in 5G networks, based on the Next Generation Point-of-Presence (NG-POP) concept, which was introduced in deliverable D3.1. The NG-POP is a location in the network, where the operator could implement multiple functions, including a common subscriber IP edge for all network types (fixed, Wi-Fi, mobile). The two main FMC architectures identified and analysed by the COMBO project are respectively the Distributed NG-POP Architecture and the Centralised NG-POP Architecture, depending on the selected location for the common subscriber IP edge. The Centralised NG-POP architecture is characterized by a small number of NG-POP locations, at the sites of Core Central Offices (COs), which are the edges between the current fixed aggregation network and the core network. Conversely, the Distributed NG-POP architecture relies on a larger number of NG-POP locations, located in the Main COs, leading to an extension of the IP backbone towards the access network.

To enable the actual implementation of functional convergence in the two NG-POP architectures, the COMBO project introduces a functional entity called Universal Access Gateway (UAG), the DP of which is located at NG-POPs and has the role of common subscriber IP edge for fixed, Wi-Fi and mobile access. Specifying the UAG CP and DP, their respective locations and how they can be implemented and deployed in the two proposed NG-POP architectures is a main achievement of this deliverable D3.5. This analysis is enriched by a study of network sharing capabilities of UAG-enabled networks and a qualitative assessment of Distributed and Centralised NG-POP architectures for implementing network services and for realising FMC use cases defined in WP2.

This deliverable D3.5 presents the final achievements made within task T3.2 (Functional Convergence) of WP3. Final key recommendations regarding both structural and functional convergence shall be reported in deliverable D3.6 planned in September 2016.

Key achievements and findings of the Deliverable

Having a common subscriber IP edge (within the UAG DP) for all individual traffic flows of fixed, mobile and Wi-Fi users of a given area, allows co-locating this subscriber IP edge with other entities such as content distribution servers. This makes the UAG the natural gateway for all services distributed over the IP layer by integrated operators. Virtual network operators can also enable their subscribers to access broadband services from multiple access technology options; this is allowed by the uDPM functions of the UAGs provided by an integrated operator or infrastructure provider. A unified authentication, such as uAUT, also opens new opportunities for collaborations between network operators and Over-The-Top (OTT) players.

The UAG definition separates the UAG DP from the UAG CP, which brings various benefits including scalability, implementation and deployment flexibility. As the subscriber IP edge for all network types, the UAG DP is playing the role of Serving / Packet data network Gateway (S/P-GW) for 4G and of Broadband Network Gateway (BNG) for fixed networks. In addition to legacy user traffic processing functions and DP level monitoring functions, the UAG DP includes the Session Mapping Execution functional block of uDPM, which realises the mapping and distribution of traffic between the multiple data paths of a given user session. The UAG CP includes user access and session control functions in FMC context, namely through uAUT and uDPM: a uAUT agent (proxy or client) facing the uAUT server, and network parts of uDPM Decision Engine, Data Path Creation/Destruction, Path Coordination and Control, and direct control of User Equipment (UE). The UAG CP also includes legacy mobility control functions of Mobility Management Entity (MME), and service control functions, e.g. resources and policy control and charging control.

Two implementation models are considered for the UAG: the standalone model where no interface is specified between CP and DP and the split model, which relies on an explicit interface defined between CP and DP. With the promising split model, the UAG CP can be co-located with the UAG DP, but can also be located remotely to allow a better centralisation of control functions. This will be largely enabled by Software Defined Networking (SDN) techniques. Network Function Virtualization (NFV) is also relevant, particularly for the UAG CP, with related functions hosted on commodity servers.

The Distributed NG-POP architecture (UAG DP at Main COs) allows higher scalability and reliability and lower latency for network services and user applications compared to the Centralised NG-POP architecture (UAG DP at Core COs), which in turn allows easier deployment and migration with less changes at IP level. uAUT implementation is not influenced by the chosen architecture, as most of the authentication procedure relies on a centralised system, but Distributed NG-POP is a better option when considering uDPM implementation.

Specifically, the Distributed NG-POP architecture with split UAG and remote CP (i.e. UAG DP at Main COs and UAG CP at Core COs) appears as the best option allowing advanced FMC features together with key legacy functionalities. It will significantly reduce latency and improve performance of content delivery, due to interactions between cache controllers and the uDPM decision engine and the possibility to have cache nodes co-located with the UAG DP at Main COs. The

Distributed NG-POP architecture is also better for handling advanced uDPM features such as vertical handover between access technologies for real-time communication services. Distributed NG-POP architecture is also preferred when considering strict DP latency requirements of 20 ms or below, as for cloud-gaming and virtual / augmented reality applications. In the perspective of network slicing, locating the UAG DP at Main COs will enable creating and composing network slice infrastructures with differentiated service requirements. Having a remote CP at Core COs brings a centralised and complete view of all resources leading to handle finer resource granularity for partitioning and composing network slices.

The qualitative analysis of COMBO architectures for realising FMC use cases shows that the Centralised NG-POP architecture is a good trade-off, especially when the UAG CP is co-located with the UAG DP at core COs. This centralised NG-POP architecture already represents a significant improvement compared to the current architectures in terms of gateway distribution, i.e. scalability, reliability and latency, as the core COs are closer to the subscriber than the EPC gateways of current mobile networks. Nevertheless, the Distributed NG-POP with remote CP, i.e. with the UAG DP at main COs and the UAG CP at core COs, appears even more appealing for most of the FMC use cases.

Also, splitting the UAG CP functions between main COs and core COs, i.e. having some control functions in main COs and some other control functions in core COs can bring additional benefits to the Distributed NG-POP architecture with a split UAG model. This tuning of the location of UAG CP functions will be enabled by SDN and NFV techniques and allow optimisation of the network for a large variety of use cases and services.

Final recommendations for fixed mobile network integration in 5G context will be drawn in deliverable D3.6, taking into account the main outcomes of COMBO architecture comparisons of this deliverable D3.5 and 5G transport considerations of deliverable D3.4, together with the main experimental achievements of deliverable D6.3.

List of authors

Full Name – E-mail	Company – Country Code
Yaning Liu – yaning.liu@jcp-connect.com	JCP – FR
Markus Amend – markus.amend@telekom.de Eckard Bogenfeld – eckard.bogenfeld@telekom.de Dirk Breuer – D.Breuer@telekom.de Thomas Monath – Thomas.Monath@telekom.de	DTAG – DE
Jose Torrijos Gijón – jgijon@tid.es	TID – ES
Daniel Abgrall – daniel.abgrall@orange.com Gregory Akpolijohnson gregory.akpolijohnson@orange.com Stéphane Gosselin – stephane.gosselin@orange.com Xavier Grall – xavier.grall@orange.com	Orange – FR
Souheir Eido – souheir.eido@telecom-bretagne.eu **Annie Gravey – annie.gravey@telecom-bretagne.eu (editor)** Younes Khadraoui – younes.khadraoui@telecom-bretagne.eu Xavier Lagrange – xavier.lagrange@telecom-bretagne.eu Pratibha Mittarwal – pratibha.mittarwal@telecom-bretagne.eu	IMT-TB – FR
Stefan Höst – stefan.host@eit.lth.se	ULUND – SW
Ricardo Martinez – ricardo.martinez@cttc.es	CTTC – ES
Ali Hmaity – ali.hmaity@polimi.it Francesco Musumeci – francesco.musumeci@polimi.it Massimo Tornatore – massimo.tornatore@polimi.it	PoliMi – IT
Tibor Cinkler – cinkler@tmit.bme.hu	BME – HU
Alberto Pineda – alberto.pineda@fon.com	FON – ES
Selami Çiftçi – celami.ciftci@argela.com.tr Onur Eker – Onur.Eker@argela.com.tr	ARGELA – TR

List of reviewers

Full Name – E-mail	Company – Country Code
Achille Pattavina achille.pattavina@polimi.it	PoliMi – IT
Peter Olaszi polaszi@aitia.ai	AITIA - HU
Zere Ghebretensae zere.ghebretensae@ericsson.com	Ericsson AB - SW
Philippe Bertin philippe.bertin@orange.com	Orange - FR

Approval

Approval	Full Name – E-mail	Company – Country Code	Date
Task Leader	Annie Gravey annie.gravey@telecom-bretagne.eu	IT-TB – FR	06.07.2016
WP Leader	Dirk Breuer d.breuer@telekom.de	DTAG – DE	07.07.2016
Project Coordinator	Jean-Charles Point pointjc@jcp-connect.com	JCP – FR	07.07.2016
Other (PMC, SC, etc.)			

Document History

Edition	Date	Modifications / Comments	Author
0.0	14/09/2015	First proposal for TOC	Annie Gravey IT-TB
0.10	15/06/2016	Internal review of D3.5	Peter Olaszi, Achille Pattavina, Zere Ghebretensae
0.13	24/06/2016	External review of D3.5	Philippe Bertin
1.0	06/07/2016	Ready for publication	Annie Gravey IT-TB

Distribution List

Full Name or Group	Company	Date
PMC		07.07.2016
PSC		07.07.2016
Other - EC		07.07.2016

Table of Content

List of Tables

List of Figures

Glossary

2G	2nd Generation (mobile service)
3G	3rd Generation (mobile service)
3GPP	3rd Generation Partnership Project
5GS	5G Services
AAA	Authentication, Authorization and Accounting
ACS	Auto Configuration Server
AN	Access Node
ANDSF	Access Network Discovery and Selection Function
AP	Access Point
API	Application Programming Interface
BBF	Broadband Forum
BBU	Base Band Unit
BNG	Broadband Network Gateway
BSS	Business Support System
CAPEX	Capital Expenditures
CC	Cache Controller
CDN	Content Delivery Network
CN	Core Network
CO	Central Office
COTS	Commercial-Off-The-Shelf
CP	Control Plane
CPE	Customer Premises Equipment
C-RAN	Centralised, Co-operative, Cloud or Clean RAN
CWMP	CPE WAN Management Protocol
D-CPI	Data-Controller Plane Interface
DC	Data Centre
DE	Decision Engine
DMM	Distributed Mobility Management
DP	Data Plane
DPCF	Data Plane Control Function
DSL	Digital Subscriber Line
DSLAM	Digital Subscriber Line Access Multiplexer
DSMIP	Dual Stack MIP
EAP	Extensible Authentication Protocol

EMS	Element Management System
eNB	Evolved Node B (base station)
EPC	Evolved Packet core
ePDG	Evolved Packet Data Gateway
ETSI	European Telecommunications Standards Institute
FMC	Fixed Mobile Convergence
FMCO	Fixed Mobile Convergence Operator
FMNI	Fixed Mobile Network Integration
FNO	Fixed Network Operator
FPS	Frames Per Second
FPSG	First Player Shooter Game
FTTH	Fibre To The Home
GPON	Gigabit-capable Passive Optical Network
GS	Game Server
GT	Game Theory
GTP	Generic Tunnelling Protocol
GW	Gateway
HAG	Hybrid Access Gateway
HeNB	Home eNode B
HGW	Home Gateway
HLR	Home Location register
HSS	Home Subscriber Server
HT	Horizontal Target
HTTP	Hypertext Transfer Protocol
IBSG	Internet Business Solutions Group
IDPS	Intrusion Detection Prevention System
IETF	Internet Engineering Task Force
IFOM	IP Flow Mobility
IKE	Internet Key Exchange
IoT	Internet of Things
IP	Internet Protocol
IPTV	Internet Protocol Television
KPI	Key Performance Indicator
LGW	Local GateWay
LTE	Long Term Evolution

M2M	Machine to Machine
MAC	Media Access Control
MANO	Management and Orchestration
MASG	Mobile Aggregation Site Gateway
MEC	Mobile Edge Computing
MIH	Media Independent Handover
MIP	Mobile IP
MME	Mobility Management Entity
MMORPG	Multi-Player Online Role Playing Games
mMTC	Massive MTC
MNH	Multi-domain Network Hypervisor
MNO	Mobile Network Operator
MOCN	Multi-Operator core Network
MORAN	Multi-Operator Radio Access Network
MPE	Multipath Entity
MPLS	Multiprotocol Label Switching
MPTCP	Multi Path TCP
MSO	Multi-Domain SDN Orchestrator
MTC	Machine Type Cellular
MVNO	Mobile Virtual Network Operator
NAT	Network Address Translation
NBI	Northbound Interface
NERG	Network Enhanced Residential Gateway
NETCONF	Network Configuration Protocol
NFV	Network Function Virtualisation
NFVI	Network Function Virtualization Infrastructure
NFVO	NFV Orchestrator
NG-POP	Next Generation Point of Presence
OCS	Online Charging System
OFCS	Offline Charging System
OLT	Optical Line Termination
ONF	Open Networking Foundation
OPEX	Operational Expenditures
OSS	Operations Support System
OTT	Over The Top
OVSDB	Open VSwitch Database Protocol
PCEF	Policy and Charging Enforcement Function
PCEP	Path Computation Element Protocol

PCRF	Policy and Charging Rule Function
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PDP	Packet Data Protocol
PDV	Packet Delay Variation
PGW	Packet GateWay
PIN	Personal Identification Number
PLMN	Public Land Mobile Network
POP	Point of Presence
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAN	Radio Access Network
RFID	Radio Frequency Identification
RGW	Residential Gateway
RPC	Remote Procedure Call
RTT	Round Trip Time
SBI	Southbound Interface
SC	Service Chain
SDN	Software Defined Networking
SGW	Serving GateWay
SIPTO	Selected IP Traffic Offload
SLA	Service Level Agreement
SOAP	Simple Object Access Protocol
STB	Set-Top Box
TCP	Transmission Control Protocol
TM	Traffic Monitor
TPSG	Third Player Shooter Game
TWAG	Trusted WLAN Access Gateway
UAG	Universal Access Gateway
uAUT	Universal Authentication
UDC	User Data Convergence
uDPM	Universal Data Path Management
UDR	Universal Data Repository
UE	User Equipment
uMTC	Critical MTC
URI	Uniform Resource Identifier
UUID	Universally Unique Identifiers
VC	Video Conferencing
vEPC	Virtual Evolved Packet core
v(H)GW	Virtual Home Gateway
VIM	Virtualized Infrastructure Manager
VM	Virtual Machine

vMNO	Virtual Mobile Network Operator
VNF	Virtual Network Function
VNFM	VNF Manager
VNO	Virtual Network Operator
VOC	Virtual Optimisation Controller

VoD	Video on Demand
VR	Virtual Reality
WDM	Wavelength Division Multiplexing
Wi-Fi	IEEE 802.11 Wireless LAN
WOC	WAN Optimiser Controller

1 Introduction

The purpose of Work Package 3 “Fixed Mobile Convergent Architectures” is to propose, define and technically assess candidate architectures for future Fixed-Mobile Convergent (FMC) networks, both in terms of data plane (DP) and control plane (CP). This deliverable D3.5 technically assesses candidate architectures for functional convergence. The target of Task 3.2 is to define candidate architectures and scenarios for future networks, allowing the convergence of fixed and mobile network functions (Fixed Mobile Convergence, FMC).

Deliverable D3.1 [1] analysed the current architectures for fixed, mobile and Wi-Fi networks. It identified some functional groups to be developed to achieve functional convergence: Forwarding, Automatic Configuration and Management, Policy & Charging, Subscriber Data and Session Management and Mobility. These required developments were then summarized in two key intermediate goals or “Horizontal Targets” (HT) to achieve FMC: “converged subscriber and session management” (HT1), which addresses the convergence of authentication and subscriber data management, and “advanced interface selection and route control” (HT2), which aims at dynamically controlling the traffic on multiple data paths while maintaining session continuity.

Two main enablers were also identified in deliverable D3.1 [1] to foster functional convergence: Network Function Virtualisation (NFV), e.g. in order to facilitate the creation of common functional blocks for fixed and mobile networks and Software Defined Networking (SDN) e.g. to implement fixed, mobile and Wi-Fi DP functions on hardware with a generic control interface using a unique controlling protocol (e.g. OpenFlow).

Deliverable D3.2 [3] studied the horizontal targets; it also proposed Universal Subscriber and User Authentication (uAUT), as the technical solution for HT1, and Universal Data Path Management (uDPM), to solve HT2. uAUT and uDPM enable the network functions needed for the realisation of the use cases defined in deliverable D2.1 [2]. While uAUT proposes a single target solution together with a realistic migration path from the current situation to a fully converged situation, uDPM is proposed as a model, which can be used to build several tools helping to realise HT2. Two major options corresponding to two network architectures have also been identified in deliverable D3.2 [3]. These network architectures focus on the organization of the FMC network and take benefit from the concept of Next Generation Point of Presence (NG-POP) introduced in deliverable D3.1 [1]. One alternative is a “centralised” approach in which NG-POPs are deployed in a small number of locations, typically at Core Central Os. Alternatively, a “distributed” approach consists in deploying NG-POPs in a larger number of locations, such as the Main COs.

The current document, deliverable D3.5, analyses and compares the identified network architectures for functional convergence based on the possible NG-POP locations. Deliverable D3.5 focuses on showing how uAUT and uDPM can be

deployed, in particular in the functional entity called the Universal Access Gateway (UAG), specifying both its CP and its DP.

In the rest of this section, different scenarios in which implementing FMC brings potential benefits are identified in order to answer the question “why is FMC needed?” Then, the UAG functional block is introduced, as the set of functions operated by the network operator in order to achieve uAUT and uDPM. It is shown that locating the DP of the UAG fits with the two COMBO network architectures introduced in deliverable D3.2 [3].

1.1 In which network scenarios is convergence needed?

Nowadays there are many types of operators and service providers that are part of the ecosystem of network services. Some operators try to integrate a wide range of services into their portfolio: Wi-Fi, mobile, and fixed access, Over The Top (OTT) services, etc. Other operators and service providers sell their products and services to other operators. In such complex scenarios, new concepts and paradigms are needed to harmonise the interaction among different operators and service providers.

Network convergence addresses the integration of different access types (fixed, mobile, Wi-Fi) in a global network architecture. This concept demands a broader network management that supports different access modes (fixed, Wi-Fi and mobile). The two main functions of network convergence are (1) the universal authentication function, which provides a unified user authentication independent of the type of access and (2) unified data path management, which provides improved load balancing, traffic shaping and access interface selection management exploiting its central location and global knowledge of the network state.

Functional convergence can be an objective in many network scenarios, either for a single integrated operator, or in the framework of cooperation among multiple network operators, or by facilitating the distribution of OTT services to the customers of network operators.

These three general scenarios are briefly described in the following sections. The first one shall be developed in Sections 2 and 3, while the others are elaborated in Section 4.

1.1.1 Convergence for an integrated (fixed, mobile, Wi-Fi) network operator

Integrated operators run multiple networks with different access types, including fixed, mobile and Wi-Fi. The operation of multiple networks has the highest potential for convergence, with the highest achievable benefits through integration. Integrated operators are able to perform a full unification of their own networks, requiring less efforts and time than what would be required from single network operators wishing to collaborate, which would need to interact among them through well-defined interfaces and Service Level Agreements (SLA).

Convergence inside an integrated operators’ network is even more interesting because unified logical functions can be implemented in close dependencies or even integrated inside the same network equipment. Potentially, this would decrease latency, imply fewer physical/logical interfaces and fewer network equipment, which

potentially means a reduction in Capital Expenditures (CAPEX) and Operational Expenditures (OPEX).

Functional convergence will reduce the investments of integrated operators in multiple networks as a result of improved utilisation of the existing equipment and infrastructure. That can be achieved through an efficient use of fixed, mobile and Wi-Fi networks, for example, using the available capacity in one network when the capacity of other networks is exhausted or limited, thus enhancing the coverage and increasing the access capacity of the integrated network. This relies on the new functions uAUT and uDPM.

Integrated operators are in a better position to offer a higher quality of experience to the end user, through a seamless connectivity for all types of access networks, without depending on third parties.

Dealing with data traffic increase is also paramount for a network operator, and integrated operators may have more network capacity at hand that can be reorganised more efficiently to achieve a sustainable operating cost.

Integrated operators have new business opportunities, by supporting enhanced services to their end users. Indeed, implementing storage and processing resources in NG-POPs could be an enabler for storage and content delivery services operated by the convergent operator; these services would be available for mobile and fixed users. Additionally, the integrated operator can also support the service offers of other network operators willing to offer not only integrated services but also a carrier-grade and seamless connectivity, which they are unable to achieve without this outside support.

1.1.2 Convergence in the context of network sharing

Today's fixed and mobile networks can still be highly differentiated from each other, each network type having its own limitations, and therefore being operated differently. On the one hand, fixed networks are constrained by resource usage, cost and energy efficiency. On the other hand, mobile networks are dealing with the rising data demand, increasing number of subscribers and diversification of services [11].

Network sharing is believed to enable lower cost and more energy efficient use of network resources through distribution of traffic loads among available access networks, as well as through simplification of network management.

Converged network enables Fixed Network Operators (FNO) and Mobile Network Operators (MNO) to cooperate in order to extend their reach. While sharing a converged aggregation network segment and core networks functions, subscribers benefit from broadband services available through multiple access technology options. This also holds for subscribers of virtual mobile operators (vMNO). Previously, vMNOs proposed only mobile coverage, but now they can extend their coverage also to fixed access regions by taking part in network sharing agreements.

Several virtual networks can be created and maintained on a single infrastructure due to network slicing techniques. By enhancing network slicing in FMC networks, an infrastructure operator can share and maintain its own infrastructure network and fulfil explicit agreements with several virtual Network Operators. When the Virtual Network

Operator (VNO) operates a virtual network designed by network slicing, he can have the full control of all DP and CP functions (e.g. session and mobility control) independently from the infrastructure operator.

In a multi-operator environment, network operators can collaborate in order to optimise usage of network resources and energy consumption. One option is to offload one's excess traffic onto the network of another operator when bandwidth is available, and another is to offer roaming between different access networks.

1.1.3 Convergence opportunities in the context of OTT service delivery

Customers are used to interact in the service ecosystems offered by OTT players such as Apple, Facebook, Google, etc.; this creates traffic growth and thus requires increased network capacity. Currently, OTTs' video offerings account for an always-increasing share of total data volume [41].

FMC is a way to focus operators' and OTT services and the quality of the offer to the customer benefits, regardless of access type. Historically different subscriber management technologies emerged in each access type, yielding different access control policies. However, OTTs have shown that it is worth for a service to be linked only to the customer (or user) rather than to the access type. The uAUT unified authentication framework proposed by COMBO natively embeds user and subscriber authentication in the FMC environment, linking specific users to subscribers, and thus fosters collaboration between network operators and OTT service providers.

As operators have closer access to the subscribers, they can help the OTT to provide a better service through e.g. information on access type and quality, geo-location, strong authentication at access level, etc.

The uAUT functions allow subscribers to authenticate only once, and have seamless access to all network and OTT services within the limits of their contractual agreement [3]. For pay-as-you-go options (e.g. for renting a video content or buying a computer game), a PIN code may be asked, but the billing could be applied seamlessly in the monthly post-paid bill, without having to provide banking credentials (as it is today for most of OTT services).

Additionally, implementing storage and processing resources in NG-POPs could be an enabler for OTT service providers operating content delivery services offered to both mobile and fixed users. Using open Application Programming Interfaces (APIs), the integrated operator could indeed allow the OTT service providers to access these resources.

FMC is thus a means to improve the performance of OTTs' service delivery and to strengthen collaborations between network operators and OTTs.

1.2 How and where should convergence functions be implemented?

COMBO has previously identified in deliverable D3.2 [3] that two main functional blocks are key for fixed–mobile network integration: uAUT allowing a unified user authentication and session management regardless of the access type used and uDPM coordinating the use of all available paths through available networks.

Both uAUT and uDPM rely on the UE collaborating with various network level entities. Both also necessitate that the network operator is aware of the users' identity; it thus makes sense to group them in a functional group.

COMBO presents the Universal Access Gateway (UAG) as one of the major functional blocks, operated by the network operator, and used to realise uAUT and uDPM. COMBO claims that the UAG concept can simplify the network by reducing the technology variety, improving latency and reliability at the same time.

The UAG is a functional entity whose DP component is defined as the common subscriber IP edge for fixed, mobile and Wi-Fi access networks. This means that the data flows of a given user can be accessed individually within the UAG DP – they are not aggregated with the data flows of other users in a single tunnel. IP level DP functions (such as filtering, scheduling, forwarding, etc.) can be applied within the DP of the UAG. IP level control functions (such as users' IP session control) can also partly be hosted within the UAG.

As shown in Figure 1, the UAG DP could be located at the current COs, or at the Main CO (special CO with higher aggregation level than standard COs and not connected directly to the network's core), or even higher in the network at Core CO level. The Core CO is the current location of the Broadband Network Gateway (BNG), which is the IP edge for fixed and Wi-Fi traffic, whereas the IP edge for mobile traffic is located at the PGW, which could be located even higher in the network than the Core CO.

Network operators are currently attempting to reduce the number of COs, i.e. to move the CO up to the Main CO. This is due to the deployment of fibre based access networks, which makes possible increasing the distance between the customer premises and the CO without degrading fixed access performance. This is why in the present document we only consider locating the UAG DP either at the Main CO or at the Core CO.

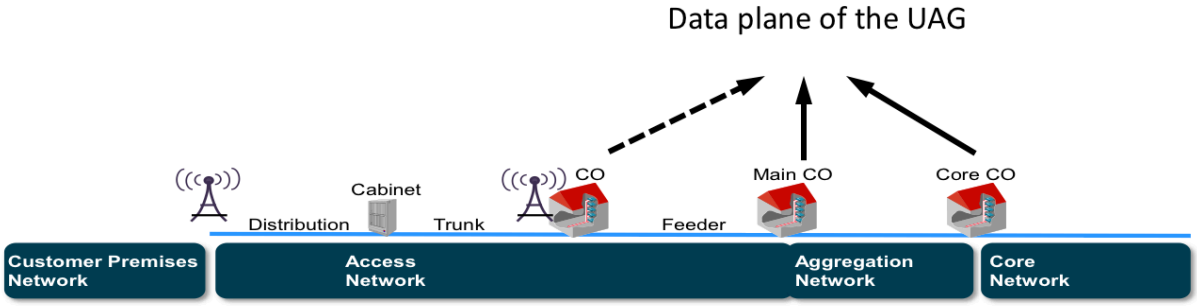


Figure 1: Reference locations for locating the data plane of the UAG

The UAG can encompass functions for mobile aggregation routers and data plane Evolved Packet Core (EPC) gateways, BNG and security gateways. The Baseband Unit (BBU) for the radio processing units could also be part of it, depending on the location of the considered CO.

As the IP layer of all traffic flows can be accessed within the DP of the UAG, it makes sense to co-locate this functional block with other entities, which require access to the IP layer of traffic flows. Typical examples of such entities are cloud storage

servers, web servers, content distribution servers, etc. Indeed, this location is potentially appropriate for efficiently distributing cloud-based services to the users of fixed, mobile and Wi-Fi networks. This specific location is named NG-POP, as it is the natural location to all services distributed over the IP layer through the network operator [1].

However, uAUT and uDPM cannot be integrally supported by the UAG, as they involve the User Equipment (UE) collaborating with several network operated functional entities: Access Nodes (ANs), UAG and centralised functional entities implementing subscriber management features. This is represented in Figure 2.

The UAG includes a part of uAUT, which relies on improvements to the 3GPP's User Data Convergence (UDC) concept (as defined in 3GPP TS 23.235). In UDC, a single global Universal Data Repository (UDR) hosts and organizes user and subscriber data providing a unified view of users and subscriptions to the operator with the idea that a subscriber using any authorized network access type should be able to be seamlessly authenticated on the other network access types.

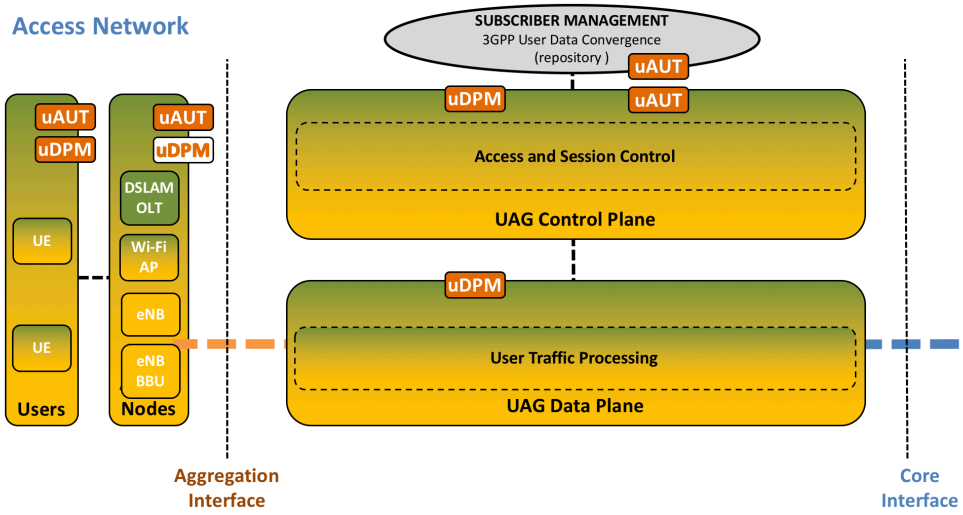


Figure 2: Splitting uAUT and uDPM in various locations

The uDPM, within the UAG, provides the tools to map a given IP session on one or several data paths ensuring session continuity. Note that data plane uDPM functions within the access node are operated below the IP layer. The uDPM allows the operator to control the path(s) that is(are) used to route users' data. This entity is triggered by a session event, which represents any singular event relative to the activity of a particular UE. When one of these triggers occurs, the uDPM decides the path or paths regarding the policies associated to the subscriber. The major part of the uDPM network controlled functions (i.e. those that provide the intelligence to select interfaces) is supported by the UAG.

1.3 Outline of D3.5

The rest of the deliverable is structured in four technical parts:

- Chapter 2 develops the UAG functional entity, explaining how the CP and DP functions involved in the uAUT and uDPM can be mapped on the UAG. It also

identifies possible implementations, the relationship between UAG and the NG-POP and the resulting NG-POP architectures. It also addresses how a generalized mobility concept can be implemented using functions provided by uAUT and uDPM.

- Chapter 3 analyses the role of the UAG in a relevant set of services such as content delivery services, real time communication services, cloud based services and Internet of Things (IoT), describing the impact of the UAG on the service delivery.
- Chapter 4 focuses on network sharing in an FMC context, studying how the UAG can be implemented in legacy and new approaches (e.g. network slicing) with multiple types of network operators.
- Chapter 5 provides a qualitative analysis about how the uses cases defined in WP2 can be mapped on the implementation options of COMBO architectures.

2 Specification and Implementation of a Universal Access Gateway

A UAG is a functional entity and thus allows the separation of the CP and the DP (or *user plane* in 3GPP terminology, 3GPP TR 21.905).

Data and control functions may be implemented in separate functional entities, as UAG DP and UAG CP, providing various benefits:

- Scalability of planes can be managed independently of each other;
- Deployment flexibility is enhanced by locating each plane at different topological or geographical places of the network;
- Different implementations (e.g., CP in commodity servers and DP in specific hardware equipment) and providers (e.g., CP from a software vendor and DP from a network equipment manufacturer) can be considered;
- The implementation of innovative network control functions is facilitated by developing new software for the UAG CP (typically by open source communities).

Regarding the CP, network applications and service creation can be moved into the cloud, i.e. in virtual machines (VMs) hosted in data centres (DC) and in centrally operated network servers, where SDN/NFV concepts could be applied. Mobile and fixed network functions (e.g. authentication, policing, charging, deep packet inspection, network address translation, etc.) can thus be abstracted, merged and implemented as common functions for both networks.

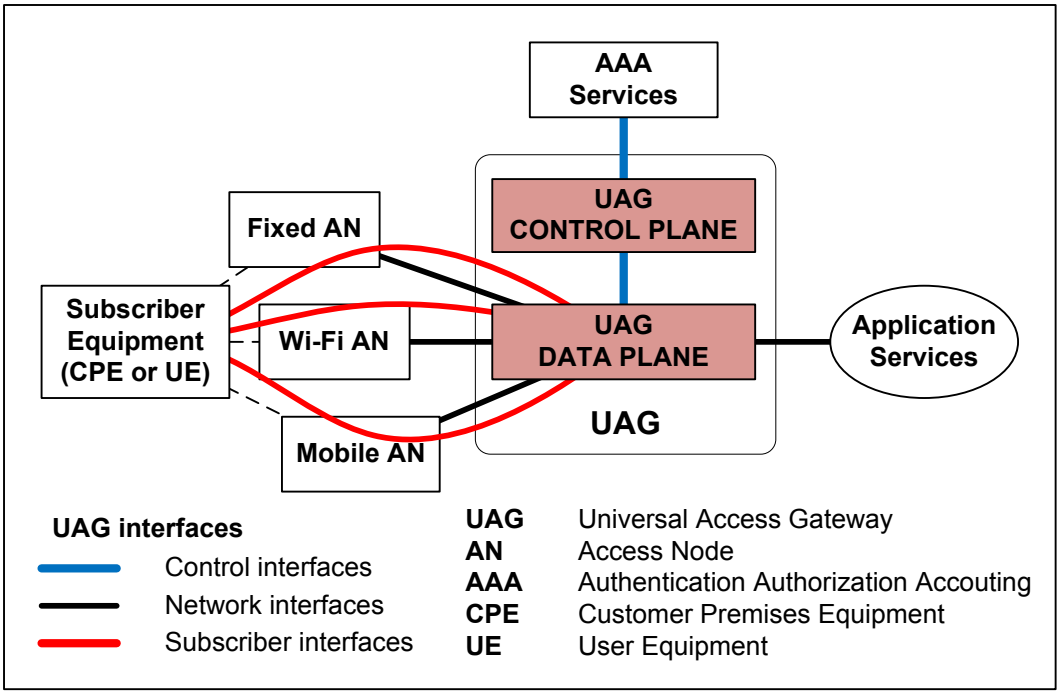


Figure 3: High level functional view of the Universal Access Gateway

Figure 3 depicts a high level functional view of the UAG showing the separation of DP and CP and its main interfaces:

- the control interface towards AAA services (authentication, policing, charging etc., as provided for instance by 3GPP-defined HLR/HSS, PCRF, OCS and OFCS functional entities,
- the control interface between the CP and DP,
- the network interfaces towards:
  - the ANs, which may be connected either directly (e.g. when eNB-related BBU function is located in the same site) or through the aggregation network,
  - the application services, terminating L4 connection, which may be also connected either directly (e.g., for local services such as CDN server) or through the core network (e.g., for internet services),
- the subscriber interfaces towards each subscriber equipment allowing the UAG to control the user traffic on a per-subscriber basis.

The network and subscriber interfaces mainly support user traffic, but also some control traffic exchanged between the UAG and respectively with ANs and subscriber equipment.

The UAG DP and CP, together with their roles in realising essential networking functional blocks, are represented in Figure 4.

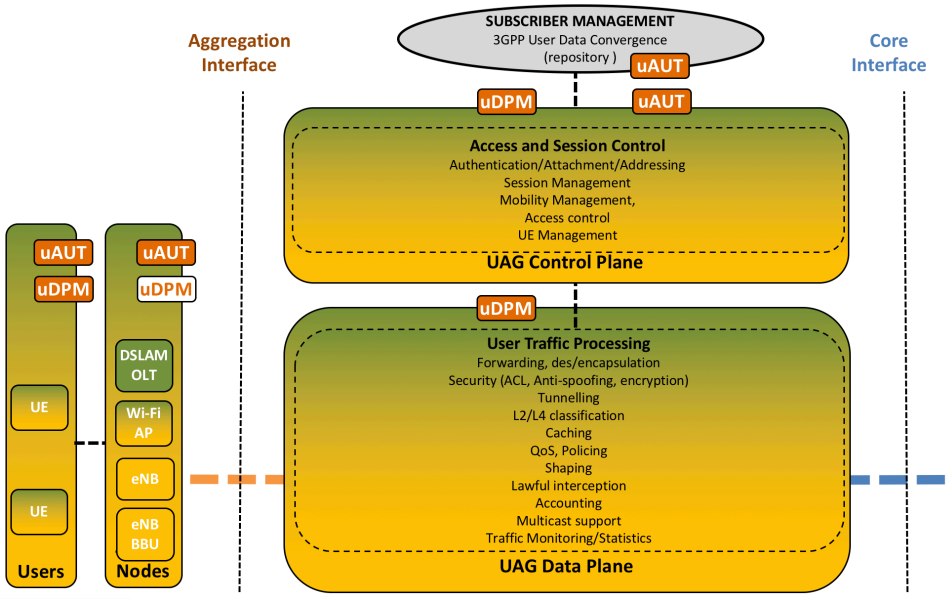


Figure 4: Mapping networking functional in the UAG Data and Control Planes

In the following, it is first assessed how uAUT and uDPM sets of functions[1] are taken into account within the UAG. The actual implementation of the UAG is then

[1] Note that data plane uDPM functions within the access node are operated below the IP layer.

addressed, depending on whether the CP is split from the DP or not. The two COMBO architectures (distributed versus centralised) are then discussed, physically locating the DP of the UAG within the network's architecture. Lastly, the role of the UAG in relationship with mobility management is addressed at the end of the chapter.

2.1 Mapping uAUT and uDPM on the UAG

This section describes how uAUT and uDPM are mapped in the UAG. Moreover, it places them in the overall architecture and describes their features.

2.1.1 Data Plane of the UAG

"Session Mapping Execution", which is part of uDPM [3], is a DP function specific to functional convergence within the UAG when implemented in the IP layer.

Session Mapping Execution is performed within a Multipath Entity (MPE) where multiple paths from the user are merged into a single logical path in the upstream direction, and where the packets from a session are forwarded over one or more multiple data paths that link the UAG to the user in the downstream direction. The MPE could e.g. be implemented as an MPTCP proxy (see section 2.1.6.2), or by directly controlling layer 3 forwarding as reported in deliverable D6.3 [6], Section 2.4.2. In case of mobility, the UAG could implement DP function related to handover.

It may also be necessary to implement DP level monitoring functions in the UAG in order to help supervising the characteristics of the different data paths between user and the UAG. This supervision is required to perform resource allocation and to interact with the uDPM functional block.

2.1.2 Control Plane of the UAG

Most of the functions identified in deliverable D3.2 [3] as being part of uAUT and uDPM are CP functions. Therefore, the CP of the UAG has to play a significant role in their implementation.

2.1.3 uAUT related functions within the UAG Control Plane

The Universal Authentication (uAUT) is in charge of the authentications of the users in all the networks associated with the UAG. It is also in charge of the accounting of the users' traffic. Lastly, it participates to the Authorisation process.

Figure 5 shows a proposed architecture for the uAUT, which has already been presented, in a previous deliverable D3.2 [3]. It acts as a proxy authentication server and redirects the authentication message to the corresponding provider's AAA service.

The uAUT is the base of the user/subscriber paradigm implementation. Figure 5 shows a proposed architecture for the uAUT. It shows the subscriber database as part of the uAUT. This database establishes the relationship between users and subscribers. In this way, when a user's access request arrives to the uAUT, it queries the subscriber database in order to know who is the subscriber for this user. The uAUT authenticates the user in the AAA server of the operator and it is able to apply different permissions to different users of the same subscription.

So the uAUT unifies all the authentication functions for the different networks that the UAG serves. The uAUT allows operators to identify the users with their sets of credentials applicable to all access networks.

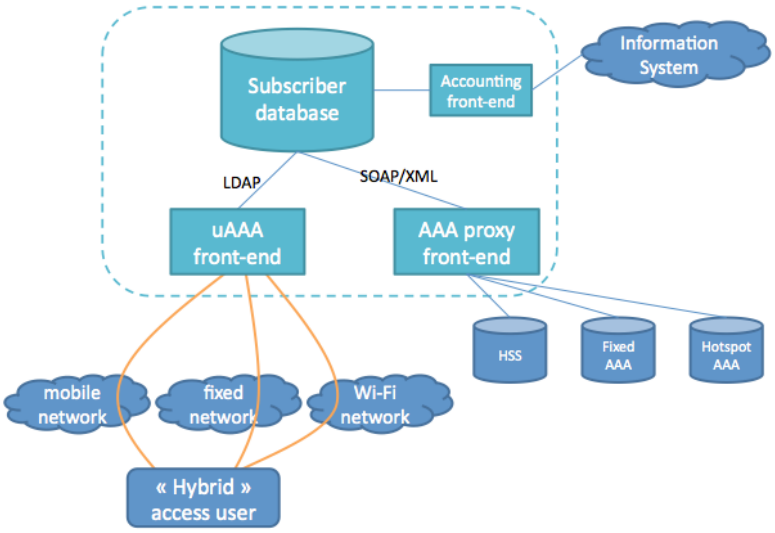


Figure 5: The uAUT architecture [3]

The CP functions that could be implemented include a uAUT client, i.e. a AAA agent (proxy or client) facing the uAUT server, as a front-end of the AAA services in order to allow authenticating to multiple access networks on a single logical network. That uAUT server is the unique contact point of the UAG for all subscriber data and authentication related functions, whatever access type used. If the subscriber profiles are unified in a common (logical) database in the uAUT server, then the operator will be able to recognize a user behind any access type and thus will be able to provide him with convergent services, such as unified accounting, seamless authentication to service platforms (IPTV, VoD) and OTT partners, consistent traffic policing (rate limiting, prioritization...) and access to the best performing network available to the subscriber.

As a first step in introducing the uAUT concept, COMBO proposes to implement the uAUT as a proxy server. This entity can be deployed in a centralised or in a decentralised scenario. The uAUT can authenticate the users acting as an AAA server. This is a good option for small areas or ad-hoc deployments.

2.1.4 uDPM related functions within the UAG control plane

The UAG also implements CP functions that realise uDPM over the available access networks. This includes [3]:

- Network part of a decision engine for data path management;
- Network part of data path creation/destruction;
- Path coordination and control;
- Direct control of UE.

These uDPM functions are related to the current control functions included in the different network subscriber-oriented gateways, such as BNG and EPC gateways.

The traffic split between multiple interfaces will need to be adapted in a flexible and fast way to changing parameters (e.g. throughput, latency, packet delay variation, packet loss) of access networks.

The uDPM could allow the network operator to configure, control and prioritize traffic flows according to its preference (e.g. “cheapest pipe first” or “limitation of data rate to maximum upper limit”). The final selection of the data path(s) used by the session would depend on a “negotiation” between the UE and the UAG.

Indeed, in order to be able to send the traffic through a selected interface, one of the most important parts that should be controlled is the UE. Because of that, the uDPM could include the direct control of mobile UE functionality. This functionality would change the behaviour of the UE (e.g. can enable or disable interfaces). With this function, the UAG could ask the UE to connect through one or another access network. As a result of this control, the network could e.g. require a mobile UE to enable the Wi-Fi interface if the decision engine has decided to redirect some traffic through the Wi-Fi network.

Moreover, the UAG can implement the Policy and Charging Enforcement Function (PCEF, as defined in 3GPP TS 23.203), which is responsible for providing controller functions in traffic handling, QoS at the UAG, and service data flow detection.

In order to improve the QoS of some services such as content distribution services or Massive Multi-Player Online Role Playing Games (MMORPG), servers delivering these services can be co-located within the NG-POP, so they locally interface with the UAG. For example, a cache controller would interface with the Decision Engine of the uDPM so that contents will be distributed optimally (see Section 3.1).

2.1.5 How uAUT and uDPM facilitate functional convergence

COMBO T3.2 started its work on FMC by identifying commonalities and differences between how different networks realised typical functional groups. This is reported in section 3.2 of deliverable D3.1 [1] and summarised in Table 1. In the present section, these functions are revisited and it is shown how uAUT and uDPM can indeed reconcile the major differences.

2.1.5.1 Forwarding

Forwarding control

uDPM grants the ability to forward the traffic based on Layer 3 information. The router forwards the packet according to the Forwarding Information Base (FIB), instantiated in the DP, which is calculated by the forwarding control block in the uDPM.

Tunnel management and Node Selection

These critical functions enabled by uDPM allow UEs, in case of 3GPP access networks, to connect to other packet data networks (e.g., the Internet) on the path of the eNB, the Serving Gateway (SGW) and the PDN Gateway (PGW). In case of non-

3GPP access networks, tunnel control protocols (such as IKEv2, DSMIP and MIPv4) accommodate the UEs to connect via non-trusted access networks.

2.1.5.2 Automatic Configuration and Management

Automatic configuration and management is a functional group that is addressed in various ways in the networks, and many of the involved functions are part of the management plane, which is not directly addressed by the COMBO project.

Table 1: Key aspects of the functional analysis with its commonalities and differences [1]

	Commonalities	Differences
Forwarding	Basically the same mechanisms	Mobile traffic transport via tunnelling protocols
Automatic Configuration Management	-	Different, incompatible mechanisms and standards
Resilience	Protection in aggregation and core network	-
Security	Similar network security functions (DoS attacks, anti-spoofing)	No encryption for point to point fixed lines
OAM	Similar mechanism and protocols	-
Synchronisation	Same mechanism and protocols	No need for Wi-Fi
Policy & Charging	-	Different mechanisms and standards, activities between BBF and 3GPP to harmonize approaches
Subscriber Data and Session Management	-	Different, incompatible mechanisms and standards
Lawful Interception and Data Retention	Similar mechanism and protocols	Different locations in the network
Traffic analysis	Similar mechanism and protocols	-
Mobility	-	Only implemented in mobile networks

2.1.5.3 Policy & Charging

Policy and charging control functions, provided by uAUT and uDPM, allow enforcing policies (e.g. for admission control) on user's traffic by use of rules, e.g., to shape and charge it according to the user's subscription. The required interfaces will be provisioned in order to be able to locally enforce the policy required by the centralised policy control point.

2.1.5.4 Subscriber Data and Session Management

Authentication and authorization

The uAUT provides the front-end for functions that perform identification and authentication of the user equipment and the validation of the service request type to ensure that the user is authorized to use the particular network services. These

functions rely on an AAA infrastructure, which is in charge of provisioning initial policies at network attachment and must be located in a higher hierarchical place

Network attachment

Network attachment, provided by uAUT, includes IP configuration, corresponding to the negotiation and configuration of network protocol parameters, e.g. IP address or prefix allocation, by relaying DHCP request or even managing a local IP address pool depending on the number of subscribers managed by the NG-POP.

Authentication is provided for network attachment, and this module interfaces with the external AAA management plane, e.g. by implementing the authenticator or receiving authentication and authorization policies from the AAA infrastructure.

Security context configuration function, e.g. security key exchange, is also provided.

Subscriber session management

Subscriber session management, also provided by uAUT, covers session/context creation and termination, session detection and identification, accounting, monitoring and session database handling.

The implementation of the subscriber session management function depends on the nature of the subscriber session and may interwork with the AAA infrastructure by triggering a dedicated AAA client interfaced with the external policy infrastructure. It also interfaces with network attachment functions, with QoS enforcement point and module and policy-based routing, in order to ensure the mapping to specific subscriber policies.

2.1.5.5 Lawful Interception

The architecture for lawful interception proposed by uAUT, relies on a centralised mediation platform, which is connected to the network elements. A simplified representation of the architecture for lawful Interception is shown in Figure 6.

The architecture applies to both legacy voice services (wireless or wireline) and to packet-based services.

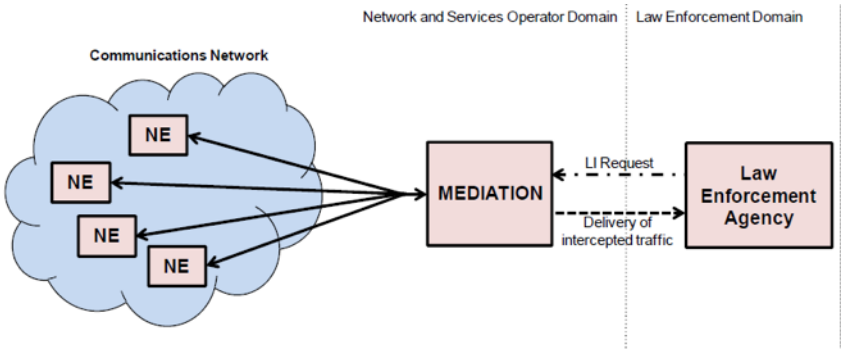


Figure 6: Simplified view of lawful interception architecture

2.1.5.6 Mobility

Mobility management procedures enabled by uDPM allow tracking of the location of the UE in order to deliver incoming calls to the UE, perform handover when the UE is already connected and include location update procedures and paging.

When a UE moves from one location to another, the UE reports its new location to the network through the location update procedure. When an incoming call to the UE arrives from the network, it identifies the location of the UE via the paging procedure. Thus mobility management procedures must be deployed with a relevant level of centralisation to ensure session continuity in case of transit between areas managed by different UAGs. This important function is more fully described in Section 0.

2.1.6 Some examples of the UAG in action

In the following, two specific applications on the role of the UAG in delivering advanced functionalities that a network operator can implement are described.

2.1.6.1 Dynamic configuration and control of Home Gateway

Control functions are distributed between the customer premises and the operator's network. The management of a Home Gateway (HGW) function includes the configuration, performance monitoring, troubleshooting, and fault management activities associated with HGW function within the context of a higher layer Service function (e.g., Activation, Diagnostics) implemented by Operations Support Systems and Business Support Systems (OSS/BSSs).

As such, the management of an FMC HGW requires a management functional architecture. The management functional architecture may incorporate an SDN orchestration management function that interfaces between the following systems:

- OSS/BSS, for management of services;
- Auto Configuration Server (ACS), Element Management System (EMS), AAA for management of network functions;
- Virtual Network Function (VNF) Manager (part of the SDN Orchestrator) for management of the underlying host environment;
- SDN Controller for management of the forwarding CP.

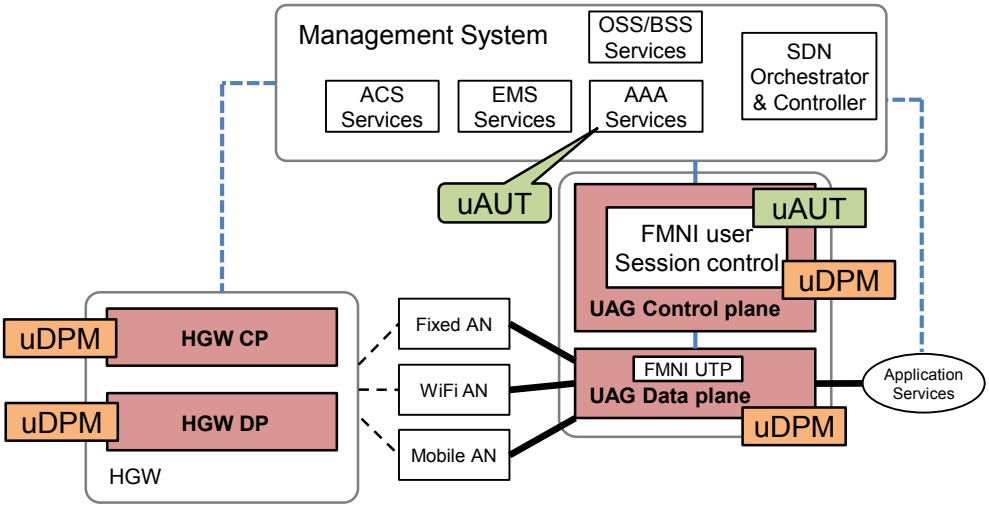


Figure 7: UAG as a converged IP edge in combination with HGW

Figure 7 shows the management functional architecture building blocks applied within the UAG context as a converged IP edge in combination with classical HGW.

The HGW management includes auto-configuration and dynamic service provisioning, software/firmware image management, software module management, performance monitoring, and diagnostics. Today the reference is a widespread deployment of the CPE WAN Management Protocol (CWMP) TR-069 [57], and associated data models [TR-181 [58], TR-098 [59]]. CWMP provides Remote Procedure Call (RPC) based communication between the Auto Configuration Server (ACS) and HGW using HTTP, or HTTPS if security is required and supporting Simple Object Access Protocol (SOAP) messaging.

The key specifications that are required for HGW management (dynamic configuration and control) are TR-069 Amendment-5 [57] and TR-181 Issue 2 Amendment 10 (TR-181-2-10-0) [58].

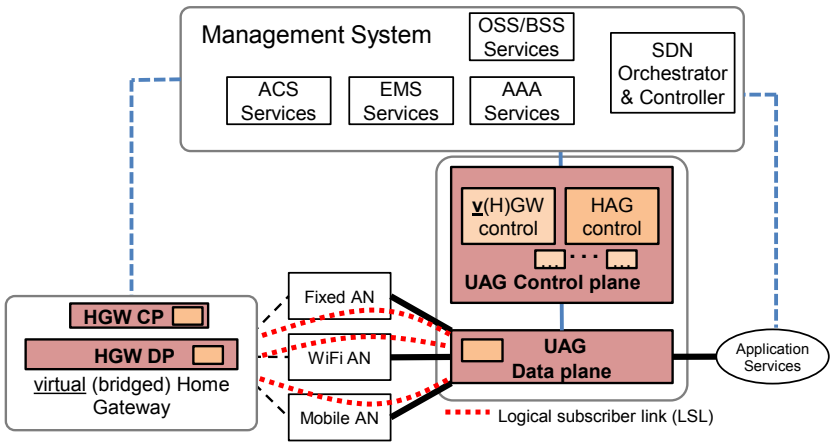


Figure 8: UAG as a converged IP edge in combination with Functional Distribution of Network Enhanced Residential Gateway Capabilities (NERG)

The Broadband Forum (BBF) currently specifies within WT.317 [61] the Network Enhanced Residential Gateway (NERG) architecture. NERG consists in shifting some of the functionalities of a residential gateway (RG), as defined in TR-124 [63], to the operator's network. If the MPE concept were applied within this context, the virtual Home Gateway (v(H)GW) as shown in Figure 8 would indeed be part of the UAG.

2.1.6.2 Implementing Priority Scheduling based on MPTCP

We show in the following a possible implementation of priority scheduling relying on the following functions:

- Identification the state of physical connections regarding number, bandwidth, delay, etc.
- Implementation of scheduling prioritization that can be turned on and off and can be changed during operation possibly based on predefined rules
- Activation/deactivation of one or more interfaces (only in case of traffic demand “overflow”)

Figure 9 depicts the CP in action. The CP consists of a path monitor and a scheduling policy API (Prio API), which handles priority rules and steers data distribution.

In the example reported here, the Prio API accepts absolute numbers in combination with a flow or interface identifier from an external repository. A lower number means higher priority. In the upper part of Figure 9, it is shown that DSL should be used in priority, then Wi-Fi and finally LTE. The Prio API interacts directly with the scheduling engine in real-time. A tie between equal priority values could e.g. result in a round robin usage of the relevant data paths or would be resolved by preferring the path with the lowest RTT. Other solutions are also possible.

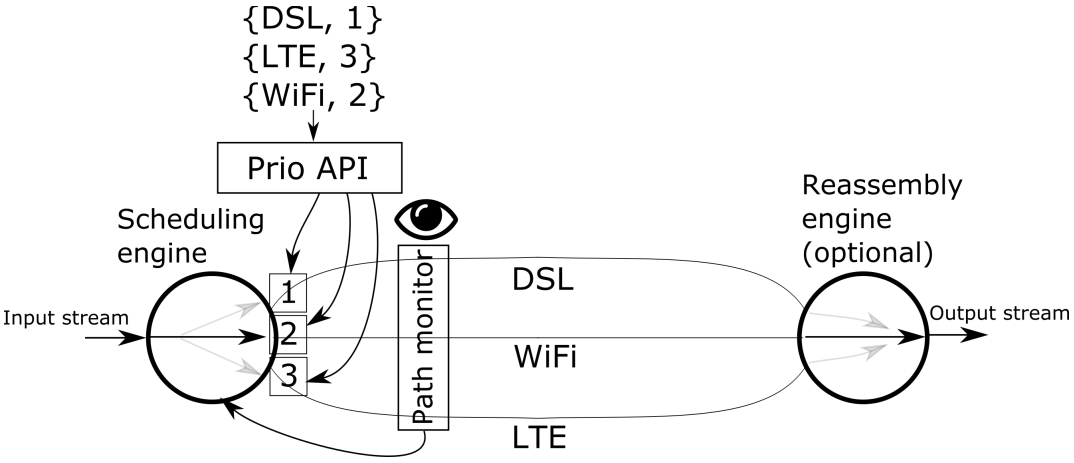


Figure 9: Control Plane for Priority Scheduling based on MPTCP

The Prio API can work on both bundling endpoints and can be different in the two directions. Session based and packet based scheduling are both possible, as well as data handover/failover. For real-time decisions, which might be necessary when using volatile access mediums, it makes sense to have low coupled (more or less independent) local CPs on RGW and UAG.

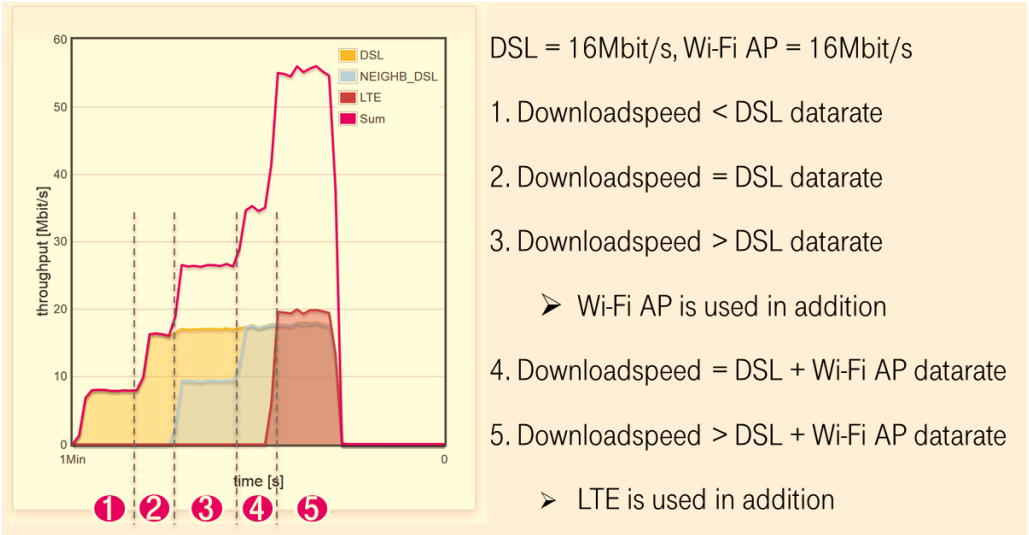


Figure 10: Resulting bandwidth (red line) for an example implementation of priority scheduling based on MPTCP

One possible result of the application of the above scheduler is shown in Figure 10. Different phases are identified, depending on the magnitude of the downstream throughput demand (represented on the x-axis) versus the capacity of the access

technologies (DSL, Wi-Fi and LTE). The total RGW bandwidth demand starts within the DSL rate in phase one and increased up to the DSL maximum. Within the third phase the demand further increases and a Wi-Fi Access Point (AP) is used. In the fifth phase the bandwidth demand increases even more and the LTE connection is activated.

2.1.6.3 Implementing uAUT using Hotspot 2.0 for a mobile UE

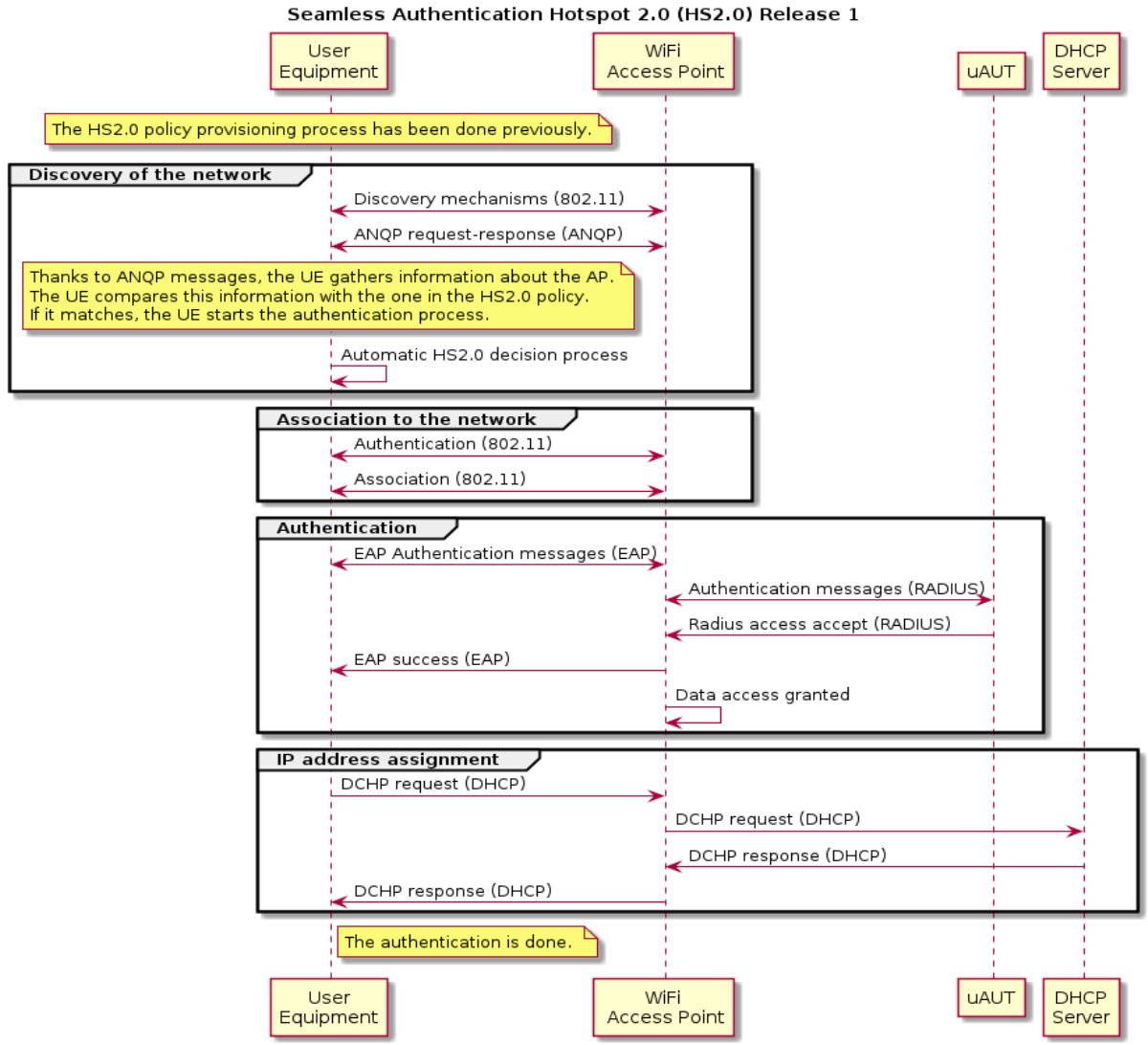


Figure 11: Seamless Authentication with Hotspot 2.0

Regarding the behaviour of the mobile UE in the Wi-Fi networks, COMBO takes advantage of a technology called Hotspot 2.0. This technology was launched by the Wi-Fi Alliance and was introduced in deliverable D3.2 [3]. Summarizing, this technology is based on a policy that is provisioned in the UE and that indicates some parameters that an AP must have in order to be eligible to connect to. Changing the policy and the parameter included on it, the provider can indicate to the UE which are the access points to which the UE should connect. The call flow shown in Figure 11 describes the operation of Hotspot 2.0.

During the final demonstration of the project, an experimental validation of the uAUT has been accomplished. The uAUT has been deployed as a proxy of the authentication server. The deployment is detailed in section 4.3 of deliverable D6.3 [6].

2.2 Implementation options

Two deployment models must be considered for the UAG: the “standalone model” where no external interface is specified between CP and DP and the “split model”, which relies on an explicit interface defined between CP and DP. In the standalone model, both CP and DP functions have to be located in the same equipment, whereas in the split model, CP and DP functions can be implemented in different equipment, and possibly in different locations.

The UAG DP being the topological IP edge node is located at the border of the access/aggregation and IP core network in both models. The split model allows keeping the DP distributed at the edge of the IP core network while implementing the CP distant from the DP, remotely within the IP core network. Figure 12 shows both the standalone and split models.

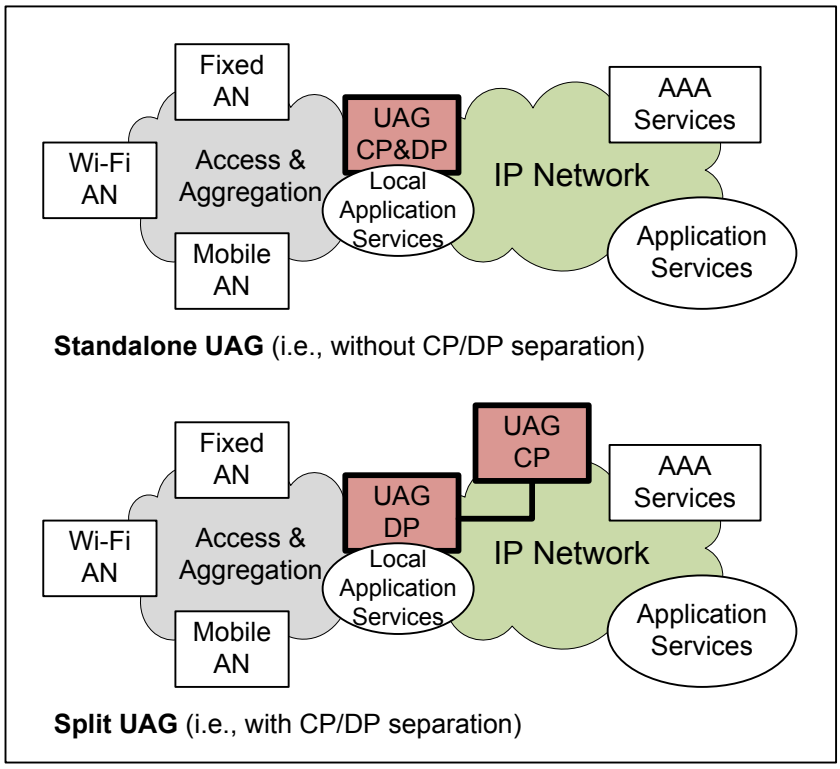


Figure 12: UAG deployment models

In the split UAG model, the level of DP/CP separation needs to be addressed, since a piece of CP must remain in the same location as the UAG DP entity (at least, to manage the interface with the UAG CP, and for the latency sensitive control functions). The objective of the split model is to keep a minimum amount of control functions at the IP edge, together with the UAG DP entity.

The direct consequence of separation between DP and CP is that the DP/CP interface (a.k.a. southbound interface, SBI, in the SDN paradigm) must also provide the ability to control all the DP functions (e.g. for tunnelling or encryption control).

The application of NFV technology may raise scalability and latency issues that must be addressed (e.g. by considering software and hardware acceleration techniques). This is mostly relevant for the UAG DP virtualisation, but may also be for the UAG CP, depending on the requirements of the control functions.

2.2.1 Implementation requirements

The UAG should support all the required functions to enable IP traffic control and processing on a per-user-basis: attachment (authentication, addressing, etc.), data path management (forwarding, tunnelling, multi-path connections, mobility, etc.), policy enforcement (QoS, legal interception, routing, filtering, etc.), data path monitoring (for data path selection, optimal content delivery, etc.).

These functions should be provided by the aforementioned uDPM and uAUT functional blocks, and by additional functions such as PCEF, so that the UAG constitutes a complete and optimal converged IP edge.

In order to manage the migration from the currently specified 3GPP and BBF functional entities, the UAG must include subscriber-related functions of the entities controlling subscriber IP sessions:

- BNG (BBF TR-101, TR-146, TR-178 and TR-291);
- Wi-Fi in Multi-Service Broadband Networks (BBF WT-321);
- Hybrid Access Gateway (HAG, BBF WT-348);
- Packet Data Network Gateway (PGW, 3GPP TS 23.401 and TS 23.402).

Since the UAG interfaces with different ANs, it should also incorporate the following 3GPP-specified entities, which manage subscriber contexts (but not necessarily at the IP layer):

- Serving Gateway (SGW, 3GPP TS 23.401);
- Mobility Management Entity (MME, 3GPP TS 23.401);
- Evolved Packet Data Gateway (ePDG, 3GPP TS 23.402);
- Trusted WLAN Gateway (TWAG, 3GPP TS 23.402).

The UAG also needs to interface with the following entities (which are part of the AAA services as described previously):

- BBF AAA server and PDP;
- 3GPP HSS, PCRF, OCS, OFCS, AAA server and other UDR front-ends.

Some functions remain out of the scope of the UAG as a functional entity. These include aggregation functions that may be necessary for backhauling fixed and mobile traffic (such as Wi-Fi AC, MASG, HeNB GW and Security GW), as well as different access functions (e.g. OLT, eNB, Wi-Fi AP). However, these functions may be co-located with the UAG in the same node (see Section 2.3).

2.2.2 Enablers

SDN is an important enabler, because it allows for the separation of DP and CP (e.g., through using OpenFlow), making it possible to consider the split model.

The ETSI NFV specifications [20] are also relevant, particularly for the UAG CP entity, since they enable hosting related functions on commodity servers.

The effort on the IP anchor distribution – currently provided by the IETF DMM working group (IETF RFC 7333) – should also be useful for addressing the questions raised by the distributed variant of the UAG DP.

FMC specifications deriving from the common work done by 3GPP and BBF must also be considered – particularly BBF TR300 and 3GPP TS 23.203 Annex P –, because they specify the same control interfaces for both fixed and mobile access towards the AAA services for policing and charging functions (namely Gx, Gy and Gz).

2.2.3 Standalone UAG

A first possible approach for the UAG implementation is the integration of the current BBF and 3GPP functional entities into a single node, so that the UAG can be regarded as a kind of structural fixed–mobile converged subscriber IP edge.

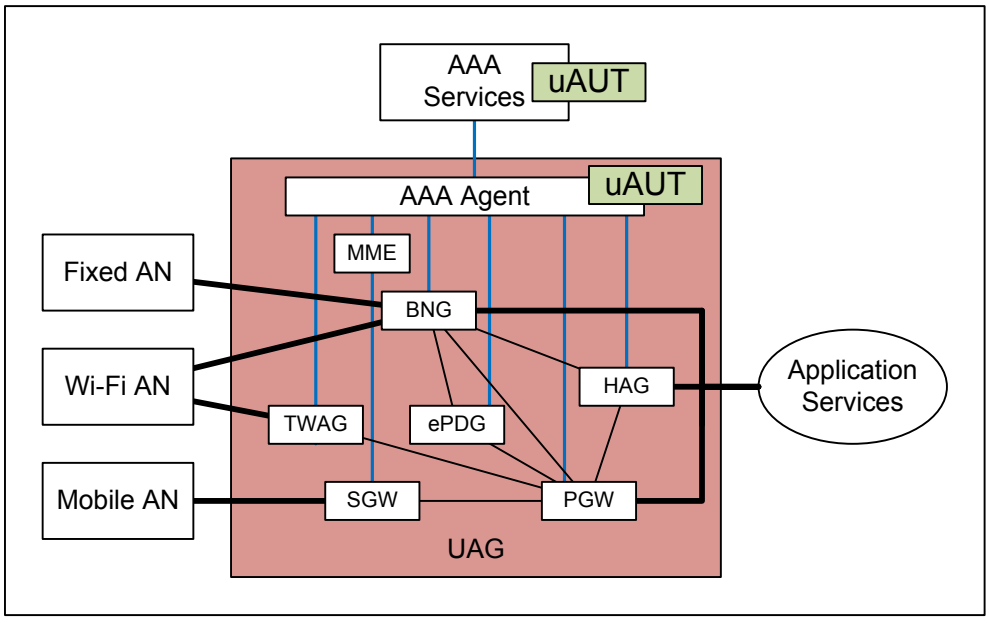


Figure 13: Standalone UAG (incremental implementation)

Thus, the UAG integrates CP and DP functions as depicted on Figure 13. That implementation option could be a first step for the UAG achievement, with even so a functionally converged AAA interface through the uAUT functional block.

A uAUT-related AAA agent (proxy or client) is included in the standalone UAG to provide a unified interface towards the AAA services.

2.2.4 Split UAG

In the split UAG approach there is an explicit interface defined between CP and DP, which can be implemented in different equipment, and possibly in different locations. The UAG CP is in charge of the user session control functions encompassing control mechanisms (routing, QoS, etc.) while the UAG DP performs user traffic processing (i.e., switching, forwarding, tunnelling, encapsulation, etc.).

The DP of the fixed and mobile entities can be functionally unified within the UAG for processing user traffic in an integrated way. It implements generic Fixed Mobile Network Integration (FMNI) functions, to be used by all access types. By implementing those FMNI functions, the UAG can be considered a true functionally converged subscriber IP edge entity.

The CP of the UAG has to implement the control functions listed in section 2.2.1.

This can be done by implementing control functions specific to each access type and can be considered as an incremental approach from the current implementations.

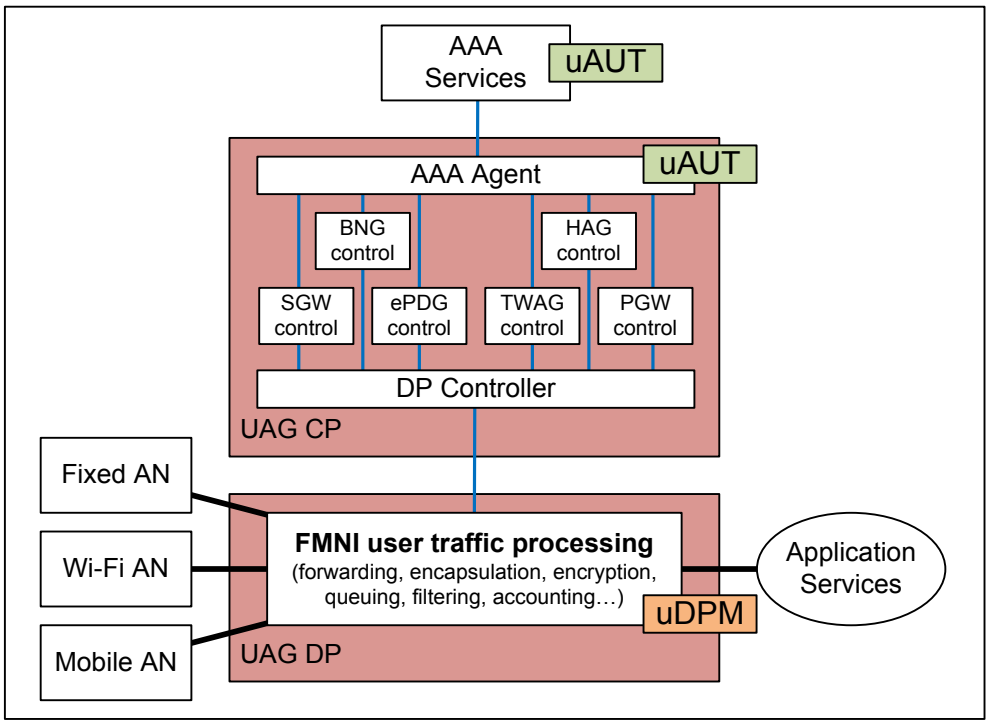


Figure 14: Split UAG as an IP edge with a fully converged DP and coordination of legacy CP functions (intermediate implementation)

Figure 14 represents this implementation with an AAA agent included in the UAG CP. A DP controller is then required in the UAG CP for unifying the control of the UAG DP. In other words, the aim of the so-called DP-controller is to provide the set of controlling functions that allow the configuration of the underlying DP element. In particular, the DP controller would be in charge of the CP components of the uDPM.

In a fully converged network, the uDPM and uAUT functional blocks can be fully integrated. The UAG CP then implements generic Fixed Mobile Network Integration (FMNI) functions, to be used by all access types. This is shown in Figure 15.

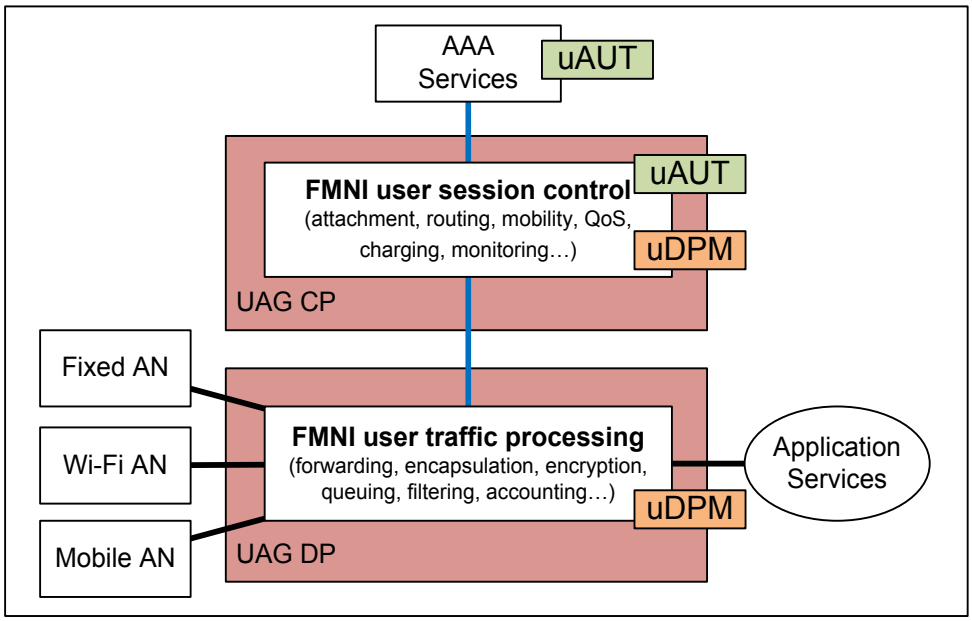


Figure 15: Split UAG with fully converged DP and CP (disruptive implementation)

2.3 Locating the UAG within the functional network architecture

Deploying UAGs instead of separated IP edges for each network type allows the network to control all user sessions, taking advantage of the resources available within each access network. This should simplify the connection of subscriber equipment with multiple interfaces (e.g. smartphones, home gateways with both DSL and cellular interface). Moreover, it would allow sharing service level resources (such as storage facilities, content distribution servers, game logics, M2M gateways) among users, regardless of the access network that is used.

Regarding the deployment scenarios for UAG, different placement options could be proposed. For a user accessing services through the mobile network, the UAG hosts functions of both SGW and PGW; for a user accessing services through the fixed network, the UAG hosts the functions related to the BNG. If UAGs are located at the Main COs (or even closer to the UEs at the COs), this implies a large number of locations. As the functions associated with the BNG, SGW and PGW are complex, increasing the number of locations where those functions are located would potentially increase the cost of the network, and add complexity to its deployment. On the other hand, due to reduced latency and improved scalability potential, the equipment supporting these functions could be simpler.

2.3.1 Definition of NG-POP

The NG-POP concept was introduced in deliverable D3.1 [1]. The NG-POP is a location in the network where the FMC operator can implement multiple functions, including the UAG, and thus a common IP edge to all access network types. As shown in Figure 1, the possible locations for the NG-POP are the CO, the Main CO or the Core CO. Keep in mind that in a non-fully converged network, the IP edges for fixed or Wi-Fi traffic on the one hand and for mobile traffic on the other are not located in a single location which makes it hard to implement FMC.

Section 1.2 has made the point that it makes sense to co-locate the UAG with other entities, which require access to the IP layer of traffic flows. Typical examples of such entities are content distribution servers, content servers, or even VMs hosted in DCs. Indeed, this location seems to be appropriate for efficiently distributing cloud-based services to the users of fixed, mobile and Wi-Fi networks.

The COMBO project has identified two architectures, called respectively “Distributed COMBO Architecture” and “Centralised COMBO Architecture” depending on the selected location for the common IP edge: the distributed COMBO Architecture corresponds to the IP edge located at the Main CO, whereas the centralised COMBO Architecture corresponds to the IP edge located at the Core CO. These two architectures are further detailed below.

The UAG is a functional block, and can therefore be deployed in several manners. In particular, using NFV and SDN, it is possible to split the DP from the CP (see Section 2.2.4). Although the NG-POP is the location of the IP edge for all access types, and thus of the DP of the UAG, functions from the UAG CP could be localized in other locations. This allows for example to implement the UAG CP functions in a central location within the IP core, while the UAG DP functions are implemented at the IP edge.

The three UAG implementation options identified in section 2.2 (i.e., standalone UAG, split UAG with co-located DP and CP, split UAG with CP distant from DP) are feasible solutions to be deployed. Nevertheless, according to the current networking trends, decoupling CP from DP (leveraging the SDN fundamentals) is gaining momentum and the split UAG cases are expected to be widely deployed.

In the following we identify two COMBO architectures, which differ on how the network operator distributes the UAG DP within its network.

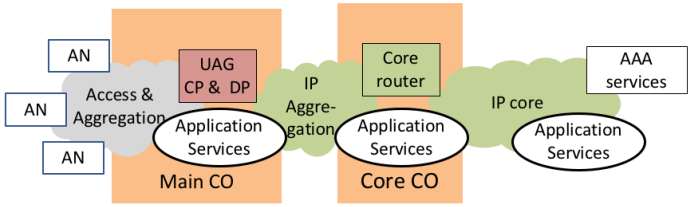
2.3.2 Distributed COMBO architecture

In the distributed COMBO architecture, the UAG DP is located at the Main CO.

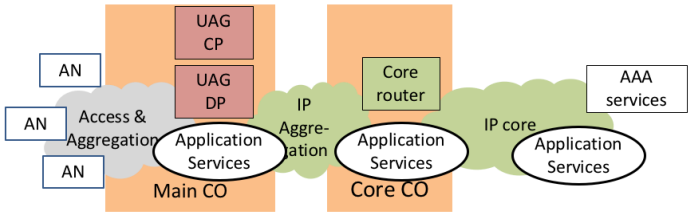
Figure 16 illustrates different deployment scenarios for the distributed COMBO architecture. It indicates the possible separation of CP and DP, allowing the network operator to locate and scale each plane independently.

The UAG DP distribution at the Main CO implies mobile IP edge distribution, so that the architecture needs to allow such distributed mobility anchoring, and thus include inter-UAG interfaces at the CP and/or DP.

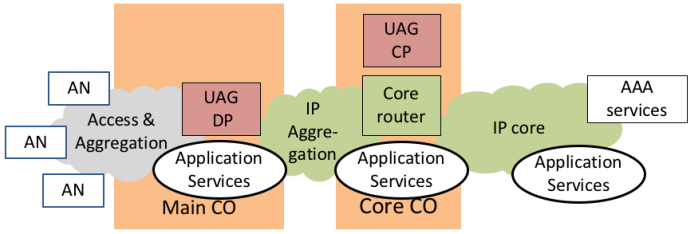
The UAG DP distribution also implies the extension of the IP network to the Main CO since the current edge reaches to the Core CO only. Such an extension (which may be considered as an IP aggregation network) can strongly impact routing management, and thus require reviewing the IP architecture.



Standalone UAG in COMBO distributed architecture



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



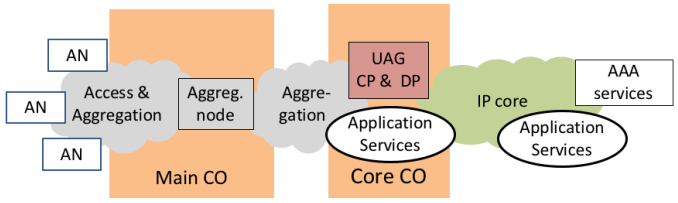
Split UAG with CP distant from DP in COMBO distributed architecture

Figure 16: UAG deployment with UAG DP at Main CO (distributed COMBO architecture)

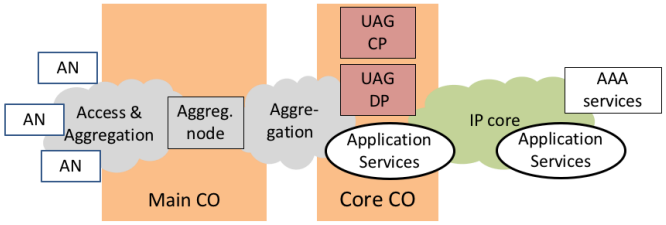
2.3.3 Centralised COMBO architecture

In the centralised COMBO architecture, the UAG DP is located at the Core CO.

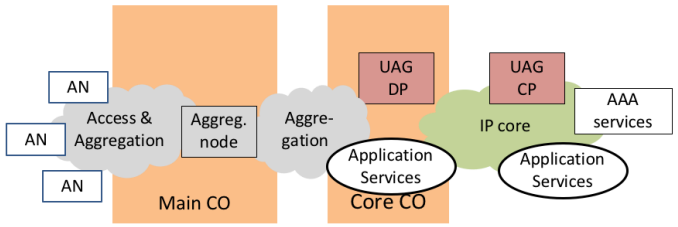
Figure 17 illustrates different deployment scenarios for the centralised COMBO architecture. It indicates the possible separation of CP and DP, allowing the network operator to locate and scale each plane independently.



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 17: UAG deployment with UAG DP at Core CO (centralised COMBO architecture)

2.3.4 Assessing UAG implementations in distributed and centralised COMBO architectures

Distributed and centralised architectures are valid approaches for NG-POP. However, they have different benefits and constraints. We can also compare them with the currently deployed fixed and mobile networks, which present distinct IP edges located at different levels within the network hierarchy.

One of the main differences between distributed and centralised architectures is that a centralised NG-POP will minimise the number of locations where it is deployed, enabling an easier operation and deployment and increasing the utilisation of network equipment. On the other hand, a decentralised NG-POP is more scalable (which is especially relevant for IoT), more reliable, and requires simpler network equipment.

Traffic volume is also another important consideration, as a distributed NG-POP reduces latency, enabling low latency applications without any further network change. Also, the amount of traffic can be reduced in the aggregation network, requiring fewer network interfaces and/or lower bitrates, and thus delaying investments by the network operator. The security gateway location is different in the two architectures and that is important because of the overhead of IPsec encryption (IPsec increases the traffic by 14% [21]) and the mobile traffic path for eNB

interconnection. Unicast video distribution in a distributed architecture will decrease the size of the aggregation network and increase the scalability of video services.

An NG-POP at the Main CO can be beneficial from the traffic aggregation point of view. In that case, the NG-POP can concentrate all the traffic coming from the access networks to the Main CO, thus reducing the number of aggregation links towards the Core CO or the aggregation network equipment inside the Main CO. The UAG can accommodate network interfaces to replace dedicated L2 aggregation nodes (used in current network architectures and still required in a centralised architecture).

The impact at IP level is higher in distributed NG-POP, which implies a more complex migration of current L2 network aggregation models to an L3 IP network between the Main CO and the Core CO. This migration may require an upgrade of network elements and a complex routing management as commented previously. The distribution of the fixed IP edge is a transformation trend in some fixed network operators to unify and simplify aggregation and transport network and to deal with traffic increase (among others); that distribution is aligned with the distributed NG-POP architecture. The impact on the mobile IP core will also be higher, although perhaps more at structural level than at functional level. The distributed architecture requires more logical instances (with the same network functionalities) running in more physical network nodes inside multiple NG-POP locations.

A distributed architecture will potentially enable new synergies that are not possible with the centralised architecture, such as integrating in the same location of the distributed NG-POP of other localized network services (e.g. caching, cloud computing, advanced monitoring, etc.), applications services (content on demand, gaming, etc.), C-RAN infrastructure, radio coordination controllers or even access network nodes.

Table 2: Main key differences of distributed and centralised COMBO architectures

Distributed	Centralised
• Higher scalability and reliability • Lower latency for network services and user applications • Simpler network equipment • Reduced traffic volume in the aggregation network • Improved integration with localized network services	• Potentially reduced CapEX and OPEX • Better utilisation of deployed equipment and better energy efficiency • Easier deployment • Easier migration with lower changes at IP level • Easier routing management

Table 2 summarizes the main differences between distributed and centralised NG-POP architectures, identifying the key aspects where each architecture potentially performs better than the other.

2.4 Relying on SDN/NFV to implement the UAG

The separation of both CP and DP is one of the key elements of the widely accepted and targeted SDN architecture. SDN aims at fully decoupling the currently coupled CP and DP within a network element, relying on the utilisation of well-defined and open control interfaces to allow the communication between both planes. This in turn

brings a number of appealing advantages to the network operator executing applications running on top of the SDN controller to configure the DP network nodes (e.g., to implement QoS policies, an efficient use of network resources, an effective multi-layer connection provisioning, virtualisation of network infrastructure, etc.).

Specifically, a logically centralised CP (i.e., SDN controller) is responsible for the configuration of the forwarding tables (i.e., adding flow rules in terms of matches and actions) of a number of DP elements (even from different vendors) via the same control interface. The SDN architectural aspects are currently being defined in the context of the ONF (Open Networking Foundation) where the interface between the CP and DP elements is referred to as the Data-Controller Plane Interface (D-CPI) or southbound interface (SBI) [15].

D-CPI can be developed according to different solutions such as (non-exhaustive list) OpenFlow [16], Path Computation Element Protocol (PCEP) [17], Network Configuration Protocol (NETCONF) [18], and Open VSwitch Database Protocol (OVSDB) [19].

The endpoints of the D-CPI at the CP (SDN controller) and the network node (in the COMBO context the UAG DP) are referred to as SDN control logic (which includes the so called Data Plane Control Function, DPCF) and the DP agent, respectively. The goal is that the decisions taken by the SDN controller are handled by its control logic, which creates the required protocol messages sent towards the DP agent. This agent interprets the received protocol messages, which are finally reflected on the (re)configuration of the UAG traffic processing (e.g., forwarding table, tunnelling, filtering, etc.). For instance, a mobile flow arriving from an eNB (via the S1-U interface) to the UAG needs to be terminated at the IP layer. This means the removal of the S1-U GTP encapsulation and the forwarding of the resulting packet flow from the UAG towards Internet.

An important aspect of the interface between the SDN controller and the DP is the information model. In general, the information exposed by the DP towards the SDN controller (CP) is abstracted. This means that physical DP resource characteristics (e.g., input/output ports, traffic forwarding and processing) are summarized to the SDN controller using abstractions. The SDN controller then operates with such an abstracted view of the DP and the DP agent is responsible for translating the operations to be done (and steered via the D-CPI interface) into the lower layer resources (including virtual switches) that actually are handled at the physical level.

2.4.1 Integrating the UAG CP within a SDN controller

Figure 18 illustrates how the UAG CP functions are implemented thanks to a SDN controller that performs the UAG DP configuration, but also configures the network elements within the access and aggregation network segments.

This means that all the network equipment (e.g., packet and optical switches) providing the connectivity between the ANs and the UAG are controlled by such an SDN controller.

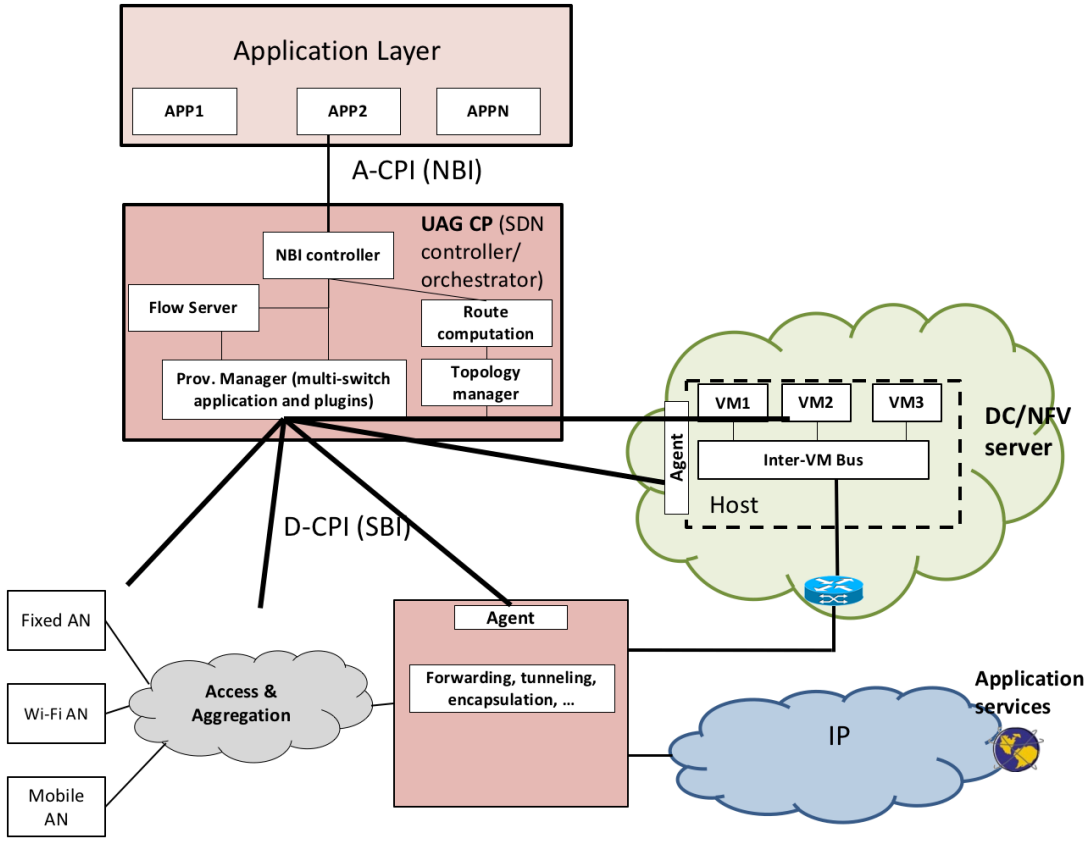


Figure 18: Example of centralised SDN-controller for the split-UAG; UAG DP is distributed combining both physical and virtualised network elements

The basic functional elements constituting the UAG CP (or SDN orchestrator when controlling other network elements than the UAG) are shown in Figure 18:

- NorthBound Interface (NBI) Controller: used to attend incoming requests (e.g., mobile services) from the application layer. This triggers the creation, modification or release of the resources from the underlying network elements (including the UAG DP). For the sake of completeness, observe that specific CP functions that are virtualised and are running in DC's VMs (e.g., vEPC MME) may communicate with the SDN controller as applications through the A-CPI. By doing so, such CP functions are allowed to instruct the SDN controller to configure the underlying DP infrastructure (i.e., switch configuration). This is illustrated in Figure 20 and Figure 21.
- Route Computation: computes the route and the networking resources of the complete DP infrastructure (including the UAG) satisfying the requirements of the requested service.
- Topology Manager: gathers details of the whole connectivity and (abstracted) view of the information such as the availability of the underlying network resources (e.g., link bandwidth, etc.).
- Provisioning manager: coordinates the configuration of the underlying network elements according to the selected decision (e.g., route computation). It is worth noting that this element may support multi-switch application and plugins

for the D-CPI enabling the communication with the network element agents (e.g., OpenFlow, OVSDB, PCEP, etc.).

- Flow Server: keeps track of the existing flows within the network in terms of the used route, allocated network resources, etc.

For the UAG DP, the main requirement derived from the virtualisation of the UAG CP is the implementation and provisioning of the D-CPI interface and its corresponding protocol(s) to allow the effective configuration of the underlying network infrastructure constituting the UAG DP. The UAG DP can be implemented on a single physical network element or virtualised as VNFs running in VMs either into common off-the-shelf hardware or hosted in DCs.

Considering a virtualised UAG as described above, Figure 19 shows (from a high level point of view) the complete workflow for establishing a new (fixed, mobile Wi-Fi) service involving the UAG DP. Without losing generality, it is assumed that all the operations of any of the above service are triggered from the application layer. In other words, there is an application dedicated for instance to handle the mobile services (e.g., with tight communication with the MME of the vEPC), another application for the Wi-Fi services, and the same applied for the fixed services.

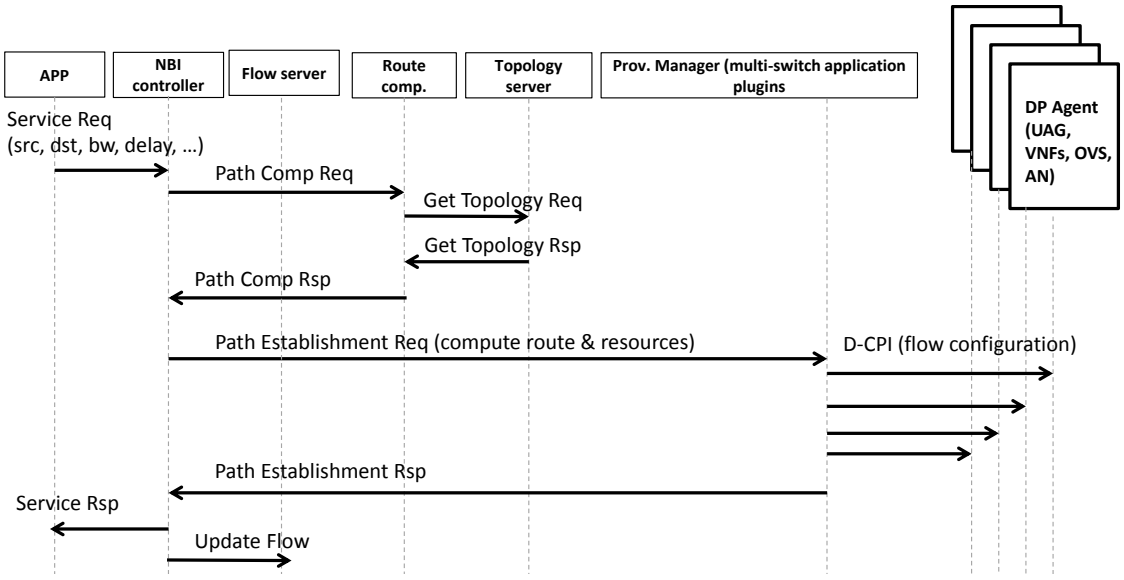


Figure 19: Generic SDN controller workflow for establishing a new (fixed, mobile, Wi-Fi) service in a virtualised UAG

When a new service has to be established, the request is sent (e.g., via REST interface) between the corresponding application and the SDN controller through the A-CPI interface. This request specifies required information such as the endpoints (e.g., eNB and S/P-GW), the requested bandwidth and other QoS parameters such as the latency.

The request is processed by the NBI controller, which coordinates all the actions to complete the requested service. This entails, first requesting the route computation manager (using the Path Comp Req in Figure 19) to find a feasible path satisfying the bandwidth and the QoS requirements. To this end, the route computation retrieves updated topology and network resource information from the topology

server (Get Topology Req/Rsp messages). The computed route is then sent to the NBI controller (Path Comp Rsp) and afterwards forwarded to the Provisioning Manager (using the Path Establishment Req).

The Provisioning Manager is able to segment the computed path into different DP components. This is done since it is very likely that the configuration of the whole DP network elements may entail the utilisation of different D-CPI protocols (i.e., OpenFlow, PCEP, OVSDB, etc.). As a result, the Provisioning Manager requires supporting multiple plugins to enable the configuration of the underlying network elements (e.g., physical and virtual UAG DP, packet and optical switches of the access and aggregation infrastructure, etc.).

2.4.2 Virtualising the UAG

The virtualisation of specifics parts of the UAG takes advantage of the NFV concept defined into the ETSI [20].

The concept of virtualisation considered by COMBO is aligned with the definition made by ONF [15] where virtualisation is an abstraction whose selection criterion is a dedication of resources to a particular client or application. An example of virtualisation is provided in Section 4.2. Leveraging on the abstraction of the underlying network infrastructure, selected resources are isolated and offered as virtual networks to client and applications running on top of the SDN controller. The latter is then responsible for composing/creating such virtual topologies and for ensuring the isolation between virtual infrastructures.

We assume that physical networking equipment (e.g., packet and optical switches) are combined with VNFs executed for instance in DCs, to provide the UAG CP and possibly CP functions.

In both implementations, the UAG CP functions required by access technologies such as the uAUT, some specific uDPM functionalities or the control part of the mobile EPC (e.g., MME) are performed by VNFs running on VMs hosted by a DC.

The difference between the proposed implementation approaches stems from the fact whether the UAG DP functions are processed or not within the cloud (DC) as VNFs.

Hereafter in this section, we focus on the split UAG flavours (i.e., with co-located DP and CP, or with CP distant from DP). Indeed, the aim of the following is to describe a pair of implementations applicable to the two split UAG solutions. It is important to state that those potential implementations must be understood as candidate strategies for the split UAG, as other variants and approaches compliant with the split UAG solutions could also be devised.

2.4.2.1 Partially virtualised UAG

In a partially virtualised UAG, the UAG DP functions such as forwarding, tunnelling, encapsulation, filtering, etc. are implemented in an equipment standing between the aggregation and the IP networks. This is illustrated in Figure 20.

Using the mobile access example, we observe that the control messages (S1-MME) between the eNBs and the MME (VNF) are forwarded through the UAG DP towards

the DC. On the other hand, the processing of the mobile user data traffics (i.e., S1-U and SGi) are directly handled at the physical UAG DP. The configuration of the actions and rules to be applied for each (control and user traffic) flow is performed by the SDN controller.

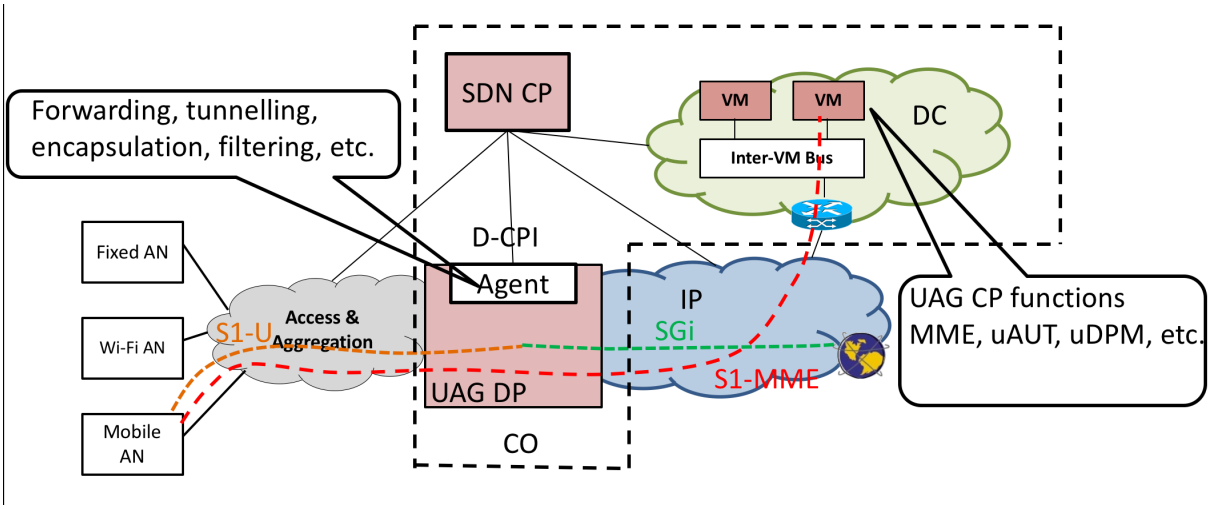


Figure 20: Virtualised UAG CP controlling a non-virtualised UAG DP

The DC can be either co-located with the IP edge in the NG-POP, or it can be implemented in a remote DC, distant from the NG-POP.

In the context of WP6, an implementation of the UAG CP configuring both physical and virtualised DP component is described in deliverable D6.3 [6].

2.4.2.2 Fully virtualised UAG

In a fully virtualised UAG, the functions to be performed within the equipment standing between the aggregation and IP core networks are considerably simplified. This is illustrated in Figure 21.

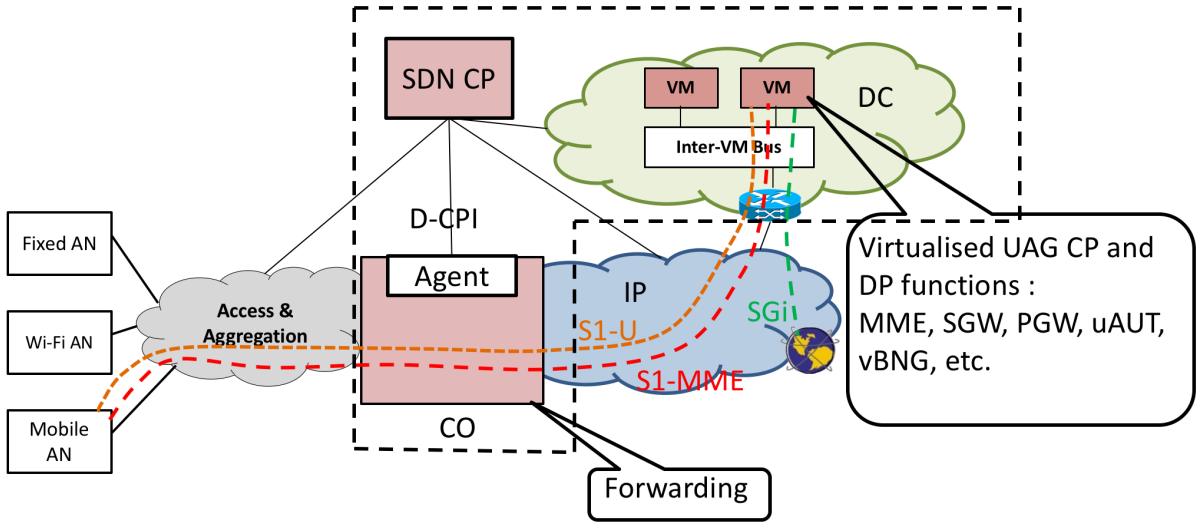


Figure 21: Fully virtualised UAG

The UAG DP functions are fully implemented within VMs accessed through a physical bare metal switch (e.g., L2/L3 switch). Only forwarding is required from the bare metal equipment standing between the aggregation and IP networks whilst the rest of the data traffic processing functions (e.g., tunnelling, encapsulation, etc.) are performed by VNFs. The VNFs are executed into VMs running in either a DC or in a single NFV server. The instantiated VNFs perform DP functions used to terminate data flows, which are originated in a number of access technologies (i.e., fixed, mobile and Wi-Fi).

All traffic coming from the user, whether data or control is forwarded to the agent in the DC that routes it towards the appropriate VMs. For instance, the regular control and user traffic flows handled in the LTE/EPC mobile communications are forwarded by the L2/L3 switch to the DC where it is routed to the appropriate VM.

Continuing with the mobile service example, in the fully virtualised UAG model, the EPC (MME, SGW and PGW) is considered to be entirely virtualised. This means that both control and user data traffic flows (S1-MME, S1-U and SGi) are processed within VNFs forming the UAG DP. Besides the virtualisation of the EPC, it is also feasible to virtualise functions for BNG, Wi-Fi gateways, etc.

Although Figure 21 shows the case when all VMs are hosted in a single DC, it is not necessarily the case. Indeed, it seems reasonable to co-locate the DC hosting the VMs implementing the DP functions with the NG-POP, whereas the DC hosting the VMs implementing CP functions can be either co-located with the IP edge in the NG-POP, or it can be implemented in a remote DC, distant from the NG-POP.

An experimental validation of a centralised UAG implementation considering the full instantiation of control and user plane functions of the mobile core EPC where the

aggregation network infrastructure is controlled by a centralised SDN orchestrator is detailed in section 2.3 of deliverable D6.3 [6].

2.5 Mobility with COMBO architectures

Mobility management is a key functionality traditionally implemented in cellular (mobile) networks. With growing availability of Wi-Fi access, Wi-Fi mobility has also to be considered. Especially in view of interoperation between fixed wireless and mobile networks (e.g. traffic offloading), mobility support within Wi-Fi and mobile access (horizontal handover), between Wi-Fi and mobile access (vertical handover between multiple access types) as well as among different operators will be required.

An FMC mobility management framework should allow mobility on demand and enable the operators to provide mobility support, which, if needed, includes session continuity for all types of devices that connect to the FMC network. It is thus expected that the FMC network will require different levels of mobility support (e.g. based on velocity of the UE or service continuity requirements), and the concept of on-demand mobility implies that the network may limit the level of mobility support for certain UEs, e.g., for stationary terminals like some sensors.

FMC mobility should also improve content distribution over all access types and facilitate load balancing between servers distributing content. Specifically, optimising the server choice for content distribution requires activating new data paths for a given user session, when necessary.

This section first discusses the role of the UAG in proving mobility features. Then, two COMBO proposals related to mobility use cases are described, and it is shown how the UAG concept applies to both cases.

2.5.1 Role of the UAG in providing mobility features

The uAUT and uDPM provide key enablers for FMC mobility features reminded above, and thus give a central role to the UAG for an FMC mobility management framework. As the mobility support should be adapted to the provided services and the user's situation, instead of the UE and selected access technology as is the case in legacy network, it can be envisaged in a more distributed and dynamic fashion.

One important step will be to provide flexible mobility anchoring functions that can be activated on demand and distributed as close as possible to the user's terminal, e.g. at the level of the UAG DP. The approaches developed in the framework of the Distributed Mobility Management (DMM) working group in the IETF can thus be applied to the UAG-based functional architecture defined by COMBO.

Another important step will be to adapt to the FMC environment all control functions required for mobility, as endorsed by the MME in current mobile networks.

UAG CP will include these FMC mobility control functions, as part of user session control. The UAG will thus have a key role in two aspects of mobility management:

- The UAG CP will include an "FMC compatible" evolution of the MME;
- The UAG DP will implement flexible mobility anchors for FMC traffic (it will play the role of both the SGW and the PGW).

This statement has important implications for the qualitative assessment of architectural options for UAG, namely in centralised versus distributed NG-POP scenarios:

- In the centralised NG-POP scenario, the UAG DP is located at Core COs, at the boundary between (L2) aggregation network and (IP) core network. Compared to the legacy mobile network (with currently a few SGWs and a very limited number of P-GWs that are highly centralised in the core network), having mobility anchors in the UAG DP at Core COs increases the flexibility and scalability of the DP, avoids "tromboning[2]" (and thus lessens potential congestion risks) by decreasing the data traffic load in the core network. Still, Core COs are rather centralised, at regional level, so that frequent changes of mobility anchors are avoided for a given user, thus avoiding too much signalling traffic. In this centralised NG-POP scenario, having a remote UAG CP deeply in the core IP network is not ideal, as this corresponds to the current location for MMEs, with the risk of scalability issues for 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). In the centralised NG-POP scenario, as far as mobility is concerned, it is thus recommended to locate UAG CP functions at Core COs, as well as UAG DP functions;
- In the distributed NG-POP scenario, the UAG DP is located at Main COs, closer to the users/devices. Compared to the centralised NG-POP scenario, having mobility anchors in the UAG DP at Main COs improves even further the flexibility and scalability of the DP by limiting even more possible tromboning effects and by decreasing the data traffic load not only in the core network but also in the aggregation network. Nevertheless, as mobility anchors will be more distributed at Main COs, this will result in more frequent changes of mobility anchors. If the CP is also located in the Main CO, this will generate a lot of signalling messages, in particular the transfer of contexts either between different Main COs or between a Main CO and a central entity like the HSS. This will not only load the communication links but also the processing units of the devices. For that reason, is it not recommended to have mobility control functions at Main COs, meaning that the UAG CP should be better placed at Core CO level. This represents a compromise between larger scalability of UAG mobility control functions and lower signalling traffic related to changes of mobility anchors.

In terms of mobility aspects, it is thus recommended to have the UAG CP at Core CO level, whatever the location of UAG DP is. Having the UAG DP at Main CO or at Core CO result in different pros and cons in terms of mobility management, which are however not critical to discriminate clearly centralised and distributed NG-POP scenarios.

[2] Basically, tromboning occurs when traffic originates at a certain point in the network, and follows a path out into the network and back to a destination close to where the original traffic originated. This is similar to the "shape" of a trombone.

2.5.2 Wi-Fi offloading of mobile traffic

When we think about Wi-Fi offloading of mobile traffic, the first thing that should be clear is whether the UE is able to connect to or not to the Wi-Fi network.

Regarding the authentication in the network, there are entities such as uAUT and technologies such as Hotspot 2.0 that are explained in this document and that solve this part (see section 2.1.6.3).

Another problem is the activation of the Wi-Fi interface. Currently, the activation and deactivation of the Wi-Fi interface in the UE is only controlled by the user. Even if the UE has a Hotspot 2.0 profile with the credentials in order to connect seamlessly to the Wi-Fi network selected by the operator, if the user has deactivated the Wi-Fi interface, the offloading is not going to happen.

COMBO proposes an entity that negotiates with the UE the activation of the Wi-Fi interface when the network decides that a Wi-Fi offloading is needed. This way, the offloading is not fully controlled by the user that has partial information, but also by the network that has all the information about the status of the different accesses (mobile and Wi-Fi).

This entity is part of the uDPM and is based on an extension of ANDSF. Moreover, this approach needs to extend, also, the UE's connection manager so that it can receive a number of new commands that are sent by the previously mentioned entity. These commands are real-time ones and they are covered in the following list:

- Switch the Wi-Fi interface ON.
- Disconnect the UE from the current Wi-Fi connection.
- Switch the Wi-Fi interface OFF.

On the other hand, the FMC network controls the status of the Wi-Fi interface of each device. In order to facilitate this control, the UE has to inform the network of its status:

- When the interface is ON.
- When it is OFF.
- When it is connected to a network (operator managed or non-managed).
- When it is not connected.

As explained in deliverable D3.2 [3], COMBO proposes an extension of the S14 interface allowing the UE to send messages to the network.

3GPP has proposed since Release 10 [22] some enhancement to the cellular network to allow offloading of mobile traffic through Wi-Fi. For instance, IP Flow Mobility (IFOM) allows to choose on which network an IP flow is transmitted. In all propositions, Wi-Fi APs are connected to the LTE EPC through specific access gateways.

COMBO proposed *very tight coupling between LTE and Wi-Fi* [23]. The idea is to connect the Wi-Fi APs with the eNBs that cover them, which is possible thanks to the convergence of fixed and cellular networks. It is thus possible to reuse LTE security provided by the PDCP layer for Wi-Fi transmissions. This avoids having to implement specific Wi-Fi security mechanisms and allows faster attachment to the Wi-Fi AP.

Thus, even mobile users who remain only for a few seconds within the coverage of a Wi-Fi AP can use Wi-Fi. Very tight coupling proposes an integration of the two access networks at a layer lower than IP, i.e. at PDCP. Thus, the UE manages only one IP address, which ensures session continuity and a seamless experience when moving from Wi-Fi to LTE and vice versa.

The concept of very tight coupling was developed within COMBO in 2013 and published in April 2014. Qualcomm developed a similar idea called LTE Wi-Fi aggregation and made a public announcement of it in February 2015 at the Mobile World Congress. LTE Wi-Fi aggregation was very recently integrated in Release 13 [81]. Though there are some slight differences between LTE Wi-Fi aggregation and very tight coupling, studies made by COMBO (see below) are valid for both solutions.

In very tight coupling, we propose a full dual connectivity between LTE and Wi-Fi by always keeping the LTE CP (i.e. Radio Resource Control connections) over LTE, while the user traffic can be transmitted either through one interface (e.g. Wi-Fi to offload the cellular network) or through two interfaces (Wi-Fi + LTE) to possibly increase the user bit rate. In that case, the multi-path entity acts below the IP layer and it does not have to be co-located with the NG-POP.

LTE and Wi-Fi have very different characteristics, in terms of bit rate, delay and packet loss. Using both in parallel may disturb higher layers such as TCP, which adapts its bit rate according to the delay of the overall connection it measures. On the other hand, very tight coupling allows traffic offloading for moving users, and thus requires very fast decisions and path establishment. This is possible by having the functions of the decision engine and path management, which are a part of the CP, as close as possible to the terminal, i.e. the Main CO. The multipath entity should always be co-located with the eNB. This makes UAG with BBU hostel in the Main CO ideally suited for very tight coupling. Thus, the most suitable COMBO structural architectures for very tight coupling implementation are standalone UAG at Main CO, and split UAG with DP and CP co-location. Other architectures can be used (UAG in the core CO) but DP should be split: the multipath entity, which is part of the DP of uDPM, should always be co-located with the eNB. Figure 22 shows the TCP throughput obtained on a real-time testbed when LTE and Wi-Fi are used simultaneously and when the two connections are used [24][25].

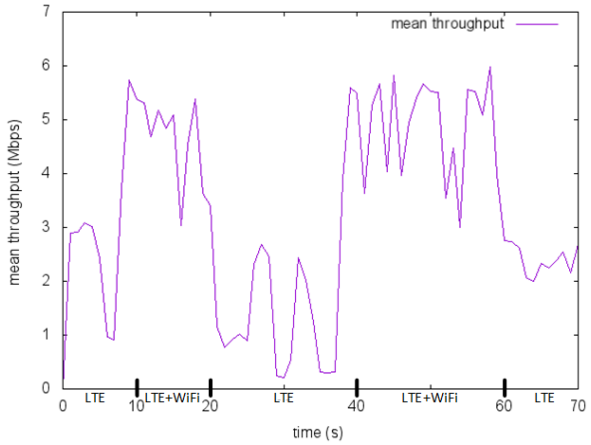


Figure 22: Mean Throughput on a test bed implementing Very Tight Coupling

In Table 3, different offloading policies are compared using two indicators: the mean global throughput computed over the entire experiment time and the mean throughput when Wi-Fi is activated, which is computed only over the periods during which Wi-Fi is used. In the first offloading policy, i.e. Full-LTE offload, the DP is transmitted over Wi-Fi only. In the case of LTE/Wi-Fi aggregation policy, Wi-Fi is used simultaneously with LTE. For this latter policy, a reordering function was added to PDCP and the experiments were performed when this function is activated and when it is not.

Table 3: Comparison of several offloading strategies

Indicator	Full-LTE offload	LTE/Wi-Fi aggregation with reordering	LTE/Wi-Fi aggregation with no reordering
Mean global throughput (Mpbs)	3.1 +- 0.1	2.0 +- 0.3	3.1 +- 0.2
Mean throughput when Wi-Fi is activated (Mpbs)	5.2 +- 0.1	2.7 +- 0.6	4.7 +- 0.4

The results first show that using Wi-Fi in parallel does not necessarily increase the throughput, compared to when Wi-Fi is used alone (first column vs. last column). This is due to the rate adaptation mechanisms used by TCP that are not really adapted when multiple paths with different delays are used simultaneously. The second thing to notice is the very bad impact that the introduction of a reordering function can have at a low layer such as PDCP in the case of multi-path transmissions (second column vs. last column). Indeed, such function disturbs the operation of higher layers like TCP by blocking received frames in the buffer even if only one is missing. If the difference of delays on the two links is too large, TCP considers these blocked frames as lost and decreases its bit rate accordingly. In addition, TCP unnecessarily retransmits these frames.

2.5.3 Fixed network offloading of mobile traffic

3GPP proposed the Selected IP Traffic Offload (SIPTO) approach in [52] and [54] 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 macrocells. One of the main objectives of using SIPTO is to ensure a better mobile connectivity service; i.e. a UE will always use the best available data path towards the external IP network. SIPTO re-assigns new Packet Data Network (PDN) gateways that are geographically closer to the current UEs locations, either co-located with the radio base station or represented as separate entities. Consequently, part of the mobile traffic is routed towards the EPC network while SIPTO traffic is offloaded within the access/metro segment of the network. According to [50], the use of SIPTO could save more than 30% of bandwidth used in backbone and up to 15% of bandwidth used in the metro/core segment of the network.

COMBO proposes to integrate the IP edges of fixed, mobile, and Wi-Fi gateway functionalities within a UAG. The integration of gateways will realise more efficient control functionalities by having a NG-POP where the global IP edge and servers are co-located.

3GPP has identified, in the framework of SIPTO in [52] and [53], some mobility use cases where session continuity is not supported due to the fact that mobility of UEs with PGW relocation implies changing the UEs IP address. Consequently, on-going SIPTO sessions requiring IP address preservation might be disrupted.

COMBO proposes a solution to provide seamless mobility during PGW relocation based on Multi-Path Transmission Control Protocol (MPTCP) [51]. MPTCP enables any host to use multiple available network interfaces simultaneously for a single TCP session. Each interface carries a sub-flow of a single TCP session presented to the application layer. COMBO's main idea is to associate the multiple addresses obtained through different PGWs to a given session in order to provide mobility support; this is one of the tools related to uDPM.

Two different cases are considered in the following.

- In the first one, the breakout point is within the UAG, which includes a co-located SGW/PGW (Section 2.5.3.1).
- In the second case, Local Gateways (LGWs), which are co-located either with Home gateways (HGWs) or with eNBs, are introduced as breakout points and allow mobile traffic to be off-loaded to the fixed access network close to mobile users (Section 2.5.3.2).

2.5.3.1 Mobility support for SIPTO above the RAN

In the present case, we assume that there is a centralised (default) PGW deep inside the core network, and that there are also multiple UAGs at the edge of the core network, which each present a co-located SGW/PGW. Part of the mobile traffic can be routed to the default PGW (e.g. communication services) while other services (e.g. content distribution services) can take advantage of the distributed SGW/PGW by co-locating servers with the UAG within the NG-POP.

As long as the UE does not change its attachment to a given SGW/PGW, its IP address is not modified and session continuity is maintained. On the other hand, for a mobile UE that changes its attachment (e.g. a UE in a moving vehicle), different IP addresses may be allocated to the UE during the lifetime of a session, and session continuity is not maintained [52], [55].

In order to provide a smooth handover, it is proposed to use MPTCP on the one hand and to slightly modify the behaviour of the MME in case of mobility on the other hand.

In order to realise a smooth SIPTO solution we first need to enhance the 3GPP SIPTO mechanism by setting up an MPTCP connection between the UE and the server. At the attachment to the network, the UE receives an IP address by the default PGW. Using this IP address, the UE connects to an appropriate, MPTCP capable server. The data path, built between the UE and the server, is a default data path that shall be used for all MPTCP signalling messages. The UE establishes an MPTCP connection over the default data path. It then requests the establishment of a SIPTO data path to the server, and thus obtains another IP address from the co-located SGW/PGW, called "local" IP address. This address is communicated to the server by the UE, for updating the server's list of addresses it uses to communicate with the UE. Using the *MP-JOIN* option of MPTCP, 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 sub-flow over the default data path as “backup path” and the sub-flow over the SIPTO data path as “regular path”. At the end, all downstream traffic from the server arrives at the UE through the regular (SIPTO) path. All MPTCP signalling messages shall be exchanged over the default (backup) data path, which is always available.

The modification of the MME behaviour is now described. Whenever the MME receives a handover required message from the source eNB, it selects the target SGW/PGW and an Indirect Forwarding Tunnel is then established between source and target SGWs. Note that the MME should have a global vision of all available SGW/PGW. If the MME is changed due to e.g. mobility, the UE’s context has to be forwarded by the initial MME to the final MME.

The traffic handover is performed similarly to the ”inter eNB/inter SGW” handover procedure defined in [55]. Then, we use the established MPTCP connection to set up a new MPTCP sub-flow to carry over the user’s traffic using the new SIPTO path. Before, the initial SIPTO traffic handover completion and after the new SIPTO path establishment, the UE will have three data paths at a time: the default data path towards the default PGW, the initial SIPTO data path towards the source co-located SGW/PGW and the new SIPTO data path towards the target co-located SGW/PGW. During this period of time, the UE may receive traffic over the initial SIPTO data path, using the Indirect Forwarding Tunnel, and over the new SIPTO path. After the completion of the handover procedure, and with the help of MPTCP, the MPTCP sub-flow dedicated to the initial SIPTO data path will be dis-joined and the IP address allocated by the source PGW for this UE will be deleted from the list of addresses stored within the server. Finally, a deactivation procedure for the initial SIPTO connection will be requested by the MME.

The distributed COMBO architecture is a better choice than the centralised COMBO architecture as it is possible to locate servers closer to the UEs in the former case.

For static UEs or for UEs with low velocity (a user who is walking or biking), session continuity is not an issue. Indeed, session continuity in case of mobility for SIPTO above the RAN is an issue only for UE which have a high velocity (e.g. a user riding a car, a bus or a train). It is also more an issue for the distributed COMBO architecture than for the centralised COMBO architecture as the distance between Core COs is higher than the distance between Main COs.

In both architectures, as pointed out in section 2.5.1, the UAG CP should be better placed at Core CO level. This represents a compromise between larger scalability of UAG mobility control functions and lower signalling traffic related to changes of mobility anchors.

2.5.3.2 Mobility support for SIPTO at Local Network

As in the previous section, we could assume that part of the mobile traffic is routed to the default PGW while other services (e.g. content distribution services) take advantage of the COMBO architecture by co-locating servers with the UAG within the NG-POP. We could also assume that the EPC is fully distributed and that mobile traffic is initially routed through the co-located SGW/PGW within the NG-POP.

For SIPTO at Local Network, the breakout point from the mobile architecture is quite close to the UE, which takes advantage of LGWs located e.g. within HGWs that host HeNBs. An MPTCP connection is set up between the UE and the server within the NG-POP, with the UE having two IP addresses, one provided for all mobile traffic and the other by the LGW. In case of mobility (for example, one can consider a student walking around a large university campus, and attaching its UE to different HeNBs), as the local address is changed every time the UE attaches to a new LGW, the session is broken.

MPTCP is used as described previously. However, as LGWs are not considered as co-located SGW/PGW, it is necessary to slightly modify the handover procedure described above by introducing a “Proxy-SGW” function within the LGW [51]. A Proxy-SGW is a purely internal function to the LGW 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 HeNB by the LGW and as a LGW by the HeNB.

When, due to mobility, the MME selects a new target LGW/Proxy-SGW, it establishes an indirect tunnel between the source and target Proxy-SGW/LGWs. This tunnel is then used to forward the SIPTO traffic uplink and downlink data traffic from and towards the UE during handover. A new SIPTO path is also established, and is set as a sub-flow to the existing MPTCP connection. Once the new SIPTO path to the target LGW is active, the initial SIPTO path can be disconnected by MPTCP.

The distributed COMBO architecture is a better choice than the centralised COMBO architecture as it is possible to locate servers closer to the UEs in the former case.

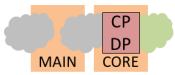
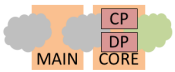
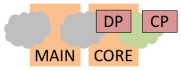
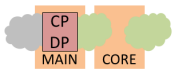
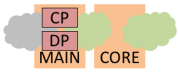
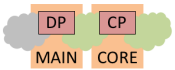
SIPTO at local network is interesting only for static UEs or for UEs with low velocity (a user who is walking or biking). Session continuity in that case is of course only an issue for moving UEs.

In both architectures, for a slowly moving UE, the MME would not be modified during the session. Implementing the MME (i.e. the UAG CP) in the Core CO seems a good choice, as in section 2.5.1.

2.6 Assessing the impact of UAG implementation options on network function realisation

In this section, we briefly recap the qualitative analysis made previously regarding how each implementation option of the UAG on the two COMBO architecture impacts the realisation of network functions. We focus on uAUT, uDPM, and on the support of mobility in the FMC network.

Table 4: Qualitative comparison of UAG implementations regarding functions

Centralised COMBO architecture			Distributed COMBO architecture		
Standalone UAG at Core CO (0) 	Split UAG with co-located DP and CP at Core CO (0) 	Split UAG with nonadjacent DP and CP and DP at Core CO 	Standalone UAG at Main CO (0) 	Split UAG with co-located DP and CP at Main CO (0) 	Split UAG with nonadjacent DP and CP and DP at Main CO 
Implementation of uAUT (1)					
IRR	IRR	IRR	IRR	IRR	IRR
Implementation of uDPM DP (3)					
+ (3)	+ (3)	+ (3)	++ (5)	++ (5)	++ (5)
Implementation of uDPM CP					
+ (2)	+ (2)	+/- (2)	++ (4)	++ (4)	+ or +/-(4)
Mobility management					
+ (6)	++ (6)	- (7)	-- (8)	- (8)	++ (6)
Implementation of very tight coupling					
- (12)	- (12)	-- (12)	++ (11)	++ (11)	+/- (11)
Smooth handover for SIPTO above the RAN (9)					
+/-	+/-	- (7)	+ (10)	+ (10)	++(10)
Smooth handover for SIPTO at local network (9)					
+/- (6)	+/- (6)	- (7)	+ (8)	+ (8)	++(6)

++ = a very good fit for the architecture
+ = a good fit
+/- = some positive and negative aspects
- = a bad fit
-- = a very bad fit
IRR: irrelevant to this particular implementation

Notes:

(0) Implementing DP and CP functions in a single equipment versus co-locating them in the same location only differs in terms of scalability performance. Co-location is made possible thanks to SDN, and allows to manage the scalability of DP and CP independently from one another

(1) Most of the authentication procedure relies on a centralised system, and the UAG presents only a client to the centralised server. See section 2.1.3

(2) Controlling how the available data paths are used at Core CO is an improvement from the current control architecture as the number of controlled UE per UAG CP is smaller than in the legacy

architecture. Using a CP remote from the DP may present scalability issues, as the number of controlled UE per remote CP instance may be significantly larger. See section 2.3.4

(3) Implementing a common IP edge at Core CO is an improvement from the current architecture as it makes it possible to efficiently utilise all data path between UE and servers, some of which can even be co-located with the IP edge. See section 2.3.4

(4) Controlling how the available data paths are used at Main CO is an improvement from controlling them at Core CO as it reduces even more the number of controlled UE per UAG CP. Using a CP remote from the DP negates this advantage. See section 2.3.4

(5) Implementing a common IP edge at Main CO is an improvement from implementing it at Core CO. Furthermore, it is easier to implement monitoring of the data paths attached to the UAG DP. See section 2.3.4

(6) Having MME functions in Core CO compared to Main CO is probably ideal, but standalone implementation with DP limits flexibility/scalability of CP functions. Split implementation co-located at Core CO could be thus the best solution for mobility aspects. See section 2.5.1.

(7) In this implementation MME functions stay very high in the core network. Even if scaling of MME is possible this would lead to unreasonable signalling load with the increase of the number of devices. See section 2.5.1.

(8) Having MME functions very low in the network (Main CO) is impractical to manage mobility. See section 2.5.1.

(9) Implementing the UAG DP and the SGW/PGW in the Main CO is better than in the Core CO, as in the former case, servers are distributed closer to the users. See section 2.5.3.1.

(10) If the MME co-located with the UAG DP at Main CO, mobility management when the UE moves from one UAG to another, has to include changing the MME, which is possibly more complex than relying on a MME located at Core CO. See section 2.5.3.1.

(11) BBU hostels should be located at the Main CO. Having DP management at the Main CO allows selecting the path (Wi-Fi or LTE) with a precise knowledge of the load of the cellular network. In very tight coupling, the CP for mobile traffic is still managed by the LTE network. Thus, the CP can be either co-located with the DP or remote from the DP. However, the multi-path entity should be co-located with the eNB. See section 2.5.2.

(12) The path selection is made by the eNB (i.e. in the BBUs at the Main CO). When data has to be transmitted through Wi-Fi, a tromboning effect appears (e.g. data from the internet goes through Core CO / BBU / Core CO / Wi-Fi AP) which may lead to congestion. See section 2.5.2.

We can derive some global conclusions from the above qualitative analysis:

- there is a significant advantage of locating the DP as close to the UE as possible, i.e. locating the common IP edge at the Main CO.
- on the other hand, there is a trade-off regarding the location of the CP: the closer it is from the UE, the finer is the control that can be applied to the multiple data paths available between the UE and the services; on the other hand, in case of mobility, locating the CP close to the user implies an increased load for signalling.

- Implementation costs may also present trade-offs between the complexity of CP logic controlling a large area and the simpler CP logic distributed in a larger number of locations, and controlling smaller areas.

3 Role of the UAG in delivering services

As the IP edge for all types of access networks is located in the UAG DP, this facilitates the delivery of services, which may be requested by the user on any available access network. In the present chapter, a relevant set of such services and the role of the UAG, interacting with the service delivery logic, are identified. In particular, we assess whether the selected COMBO architecture has an impact on service delivery.

3.1 Content delivery services

A CDN network typically distributes content to several points of presence which are either located outside the network operator's domain in case of OTT service, or within the operator's domain when the service is supported by the network operator

In legacy architectures, CDN edge servers are located outside the mobile network. When Internet content is requested by a mobile device, the content has to be fetched from the CDN servers. Although the CDN helps in reducing Internet bandwidth consumption and associated delay/jitter, the content still has to travel through the wireless carrier Core Network (CN) and Radio Access Network (RAN) before reaching the mobile device. Having to bring each requested content from the Internet can put a significant strain on the carrier's CN and RAN backhaul, leading to congestion, significant delay, and important requirements on the network's capacity to serve large number of concurrent video requests.

In COMBO, we focus on converged content delivery solutions in FMC networks.

The scenario proposed in the present section is as follows: the network operator builds, controls and manages the deployment, allocation and maintenance of the caching system, but may delegate to the OTT, or content provider, the task of managing content placement. Such an agreement between network operator and CDN could allow network operator to improve the QoS, reduce internal costs and offer new services.

Beyond the framework of agreements between FMC network operator and CDN, we also investigate the concept of "in-network caching" in FMC, as cache functionalities can be achieved by today's technologies and are available on network devices like routers, switches, etc. In our solution, the storage and caching functions are enabled in customers' Home Gateways (HGW) and within the NG-POP. By caching in access network and in the NG-POP, the content can be intelligently duplicated closer to the users, and efficiently delivered on all types of access networks.

We introduce a controlled content caching system including two components: Cache Node (CN) and Cache Controller (CC). The CN performs the caching and prefetching functions and is controlled by the CC. The CC offers "Caching-as-a-Service" to content service providers in order to improve users' QoS/QoE. It provides new added values for the network operator. The CN is deployed in HGW and in NG-POP, while the CC is running in the NG-POP. This is represented in Figure 23. The control of a CN managed by a CC has been demonstrated in WP6, and described in section 2.5.2 of [6].

The QoS of content delivery services for mobile users can be improved thanks to FMC with functional blocks including uAUT and uDPM. The four functional blocks in uDPM, namely Decision Engine (DE), data path creation and destruction, path coordination and control, and session mapping execution, have a high interaction with content delivery services. First, the DE can help in the selection of the optimal UE interface for traffic offloading and of an efficient caching decision, which in turn can help improving the QoS of content delivery services. Then through managing, controlling the data path and session mapping, uDPM provides a seamless handover and achieves optimised content delivery services.

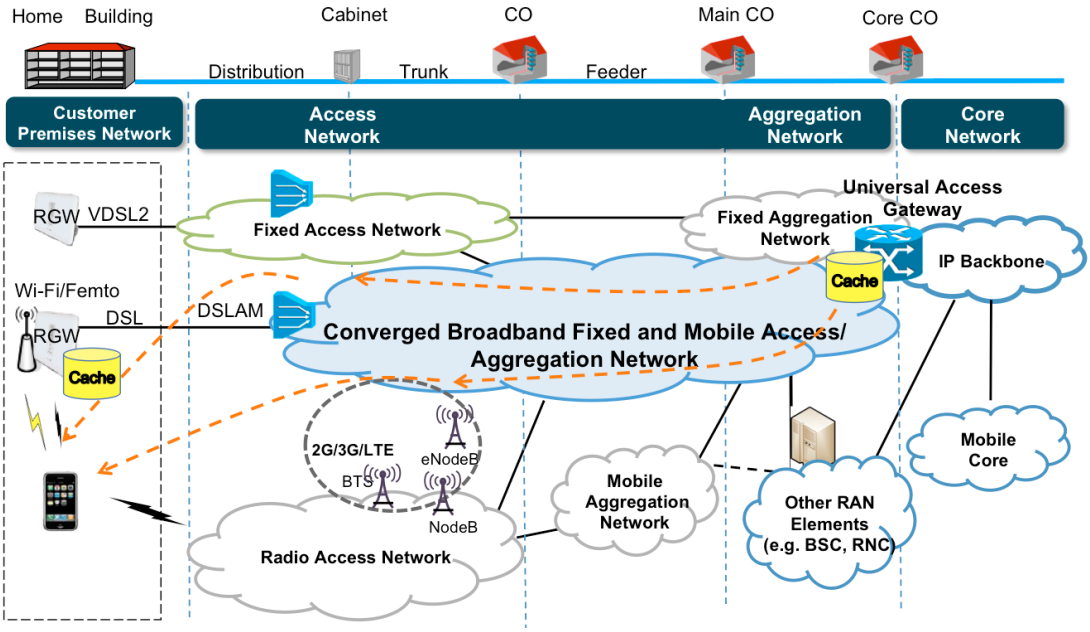


Figure 23: Architecture of converged content delivery solution

Content delivery services are impacted by the type of COMBO architecture (centralised versus distributed). In these two architectures, the path of traffic offloading and the location of content caching should be designed and deployed in different ways.

For example, in the distributed architecture, content is more localized and dynamic, so the cache hits could be less predictable than cache hits in a centralised architecture. It requires a managed caching solution such as prefetching content that will be most probably watched by users (popular content or socially estimated content) in the near future.

In the following, we assess the impact of the COMBO proposed functional architectures on content delivery services. We also explain the achieved caching benefit respectively in COMBO centralised and distributed architectures.

3.1.1 Efficiency of content distribution

There are several important metrics to evaluate the benefits of content distribution relying on a CDN.

The first metric is Cache Hit Ratio that is the probability that a request made by a user will be served by the cache. “Cacheability”, which is defined as the probability that the cacheable request or cacheable data volume can be cached in a given timeframe, is also often used to quantify the gains of caching. Cacheability is equivalent to the request hit-ratio or byte hit-ratio when there is no cache limit.

The second metric is Cache Benefit that is the amount of savings in bandwidth or improvement in user QoS. In order to maximise this benefit, the caches should be located as close to the users as possible. Hence there is a trade-off between two metrics, improving the cache hit ratio and deploying caches closer to end users.

CDN can improve the QoS of content delivery services because they deliver the content from a nearby location. Table 5 from [64] illustrates the impact of deploying CDN content servers in different locations in the network on throughput and network RTT. It is obvious that the RTT and throughput are improved if the content servers are deployed close to the users. The measurement study [65] shows that the mean download time in a CDN network is half of the mean download time from a server centralised in the core network, and is one fifth of the mean download time from a server in an international location.

Table 5: Qualitative effect of distance on throughput and download time [64].

Distance (Server to User)	**Network RTT**	**Typical Packet Loss**	**Throughput**	**4GB DVD Download Time**
Local: <100 mi.	1.6 ms	0.6%	44 Mbps (high quality HDTV)	12 min.
Regional: 500–1,000 mi.	16 ms	0.7%	4 Mbps (basic HDTV)	2.2 hrs.
Cross-continent: ~3,000 mi.	48 ms	1.0%	1 Mbps (SD TV)	8.2 hrs.
Multi-continent: ~6,000 mi.	96 ms	1.4%	0.4 Mbps (poor)	20 hrs

There are two different CDN architectures. One takes a distributed approach by deploying its servers at thousands of locations across the world (e.g. Akamai, Google), and the others (e.g. Level3, Limelight) take a more centralised approach by building a few DCs, each peering with many ISPs. With a global consideration of reliability, scalability, performance and Internet frequent failures, a distributed CDN solution has usually a better design than a centralised solution. Akamai in [64] reports that it deploys globally tens of thousands of CDN servers that run sophisticated algorithms to enable the delivery of highly scalable distributed applications. Akamai‘s approach generally has been to reach out to the true edge of the Internet, deploying not only in large Tier 1 and Tier 2 DCs, but also in ISPs with

large numbers of end users. It is worth noting that servers located in NG-POPs, as COMBO proposes, would be even closer to the end-users.

Caching effectiveness depends on the number of users connected to the cache server. Indeed, caching effectiveness is not that high if the network caches are used in the points where the user population is low. In general, caching effectiveness increases in direct proportion to the amount of the user population. The measurement for an access network in FTTH access [64] with 30 000 customers and in a xDSL network [67] shows that almost 50% of the requests are cacheable and that caching could reduce traffic by one third. The study in [68] shows that the average cacheability in LTE network that can be achieved is roughly 30% when serving 1 million users in a Core CO, but drops to 10% when serving 10 persons attached to a HGW. Furthermore, the evaluation in a mobile access network [69] has shown that caching at BS or HGW levels would not be efficient. However, even if passive caching is not effective, coupling caching facilities with a proactive policy of prefetching popular content could improve the caching benefit, and possibly justify the deployment of caches with small numbers of connected users.

3.1.2 Interfacing data and control planes for content delivery services

In this section, we first introduce in details our content delivery system, and then explain the relationship between the content delivery system, and the uAUT and uDPM functional blocks proposed by COMBO.

The basic function of the CC is to split the management and placement of content from the management of the content delivery infrastructure (e.g., forwarding and caching equipment). This allows content service providers to concentrate on improving their service quality while letting network operators take care of the configuration of the underlying network. Network operators can provide their content delivery infrastructures and the configuration interface as a service to content service providers.

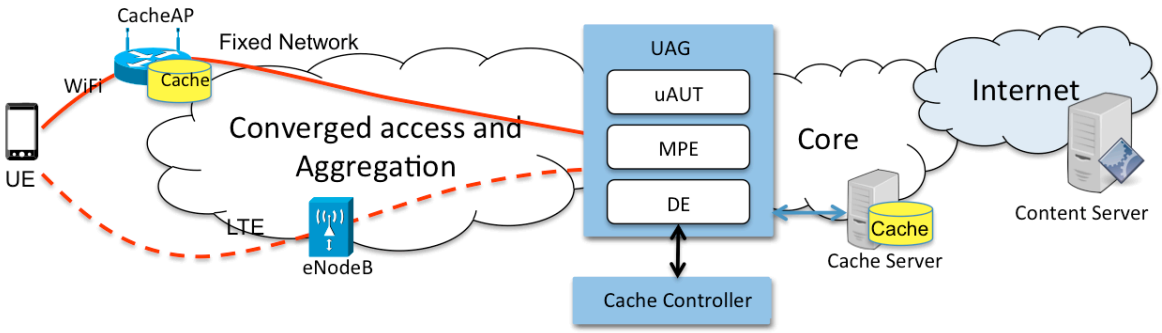


Figure 24: Converged content delivery solution interacting with uDPM and uAUT

The architecture of our proposed content delivery system is described in Figure 24. The major uDPM functions used in this architecture are the DE and the MPE. To offer the caching service, the CC has full control on the operations of the caches deployed in the HGWs or NG-POPs. Moreover, the DE should provide information about users' profiles and network performance to the CE that has to take optimised cache decisions.

The CC shall provide the information of caching decision to the decision engine, in terms of a preferred path for an end user. Then the DE can make an optimal path selection decision to get a better quality of content delivery service for this end user. We have implemented the communication between CC and DE with JSON-RPC (http://www.jsonrpc.org/specification). Three messages are defined:

- Resource_Information_Request: sent by the CC to the DE to get information regarding available resources
- Path_Preference: sent by the CC to the DE to announce preferred data path(s)
- Path_Switching: sent by the DE to the CC to announce that the user is using a new data path

The communication between DE and CC has been implemented and demonstrated in section 2.5.3 of [6].

3.1.2.1 Issues related to uAUT

uAUT acts in the first phase of the access to the network when the user is authenticated by the network operator. In that stage, the uAUT does not affect the content delivery services operation.

Nevertheless, having an entity that unifies all the authentications of the network is the first step towards unifying all the authentications required to deliver services, including OTT services, including content distribution services, as described in deliverable D3.2 [3].

3.1.2.2 Issues related to uDPM

By communicating with uDPM, the content delivery system allows operators to configure, control and prioritize delivered content according to its preference or to an agreement between OTT and content provider.

Thanks to the COMBO architecture, the content delivery system can interact with the DE in uDPM to achieve an optimal traffic offloading and content caching decision. Assuming that that the DE is aware of actual traffic load and network status, and can identify the traffic bottlenecks, it can communicate monitoring information to the CC which in turn can activate a caching strategy according to the information received from DE, in order to improve the quality of the content delivery service.

When a content provider receives many requests for the same content from many different users in the same area, it can also decide to activate content caching for these specific content. The DE provides information relative to user location and network performance to help the cache controller making an optimised caching decision as shown in Figure 25.

One potential issue related to caching is that the amount of HTTPS traffic used by some popular website such as Facebook and YouTube is currently increasing. Measurement shows that, approximately 35% of the total HTTP traffic is HTTPS [66]. Different approaches to overcome this are currently discussed in various standard bodies. For example, a proposal sent to ETSI proposes to insert a proxy between

users and servers. Specifically, when a UE requests a piece of content, it would send a request to the proxy. The proxy would then take over, fetching the content, and delivering it to the UE. This would break the https connection between the UE and the server, allowing smart content caching and prefetching of encrypted content in the network. In COMBO, we assume this issue can be resolved through agreements between network operators and OTT (or content provider), which would provide network operators the necessary control over content delivery.

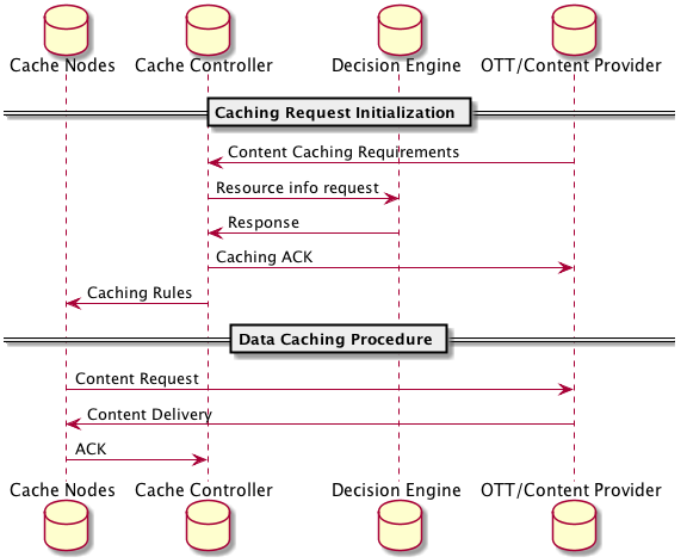


Figure 25: Caching interaction between uDPM and content delivery system

When the DE takes an interface switching decision for an end user, due e.g. to a handover, this is communicated to CC that will activate the necessary caches or pre-fetch some content. The DE provides the information of user location and network performance to help the cache controller make an optimised prefetching decision. The communication between different modules helps improving the QoS of the content delivery service. The scenario reported in section 2.5.3 of deliverable D6.3 [6] is represented in Figure 26.

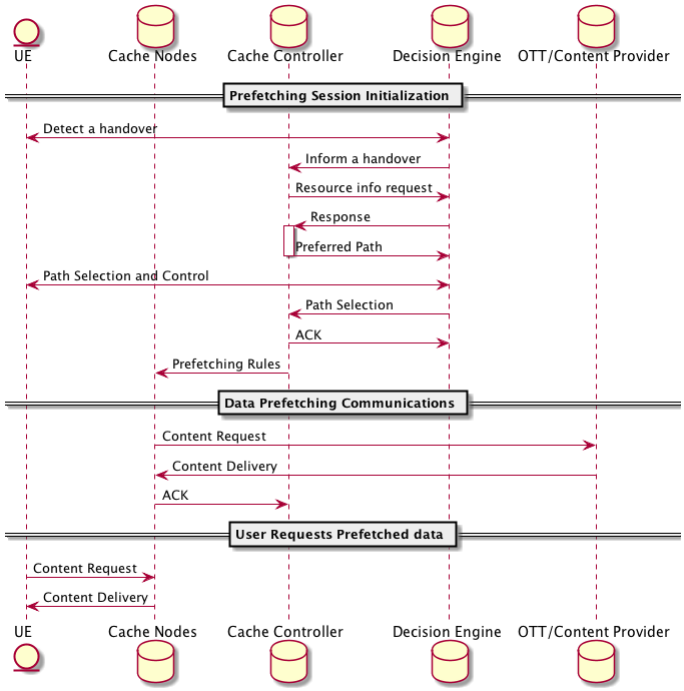


Figure 26: Prefetching interaction between uDPM and content delivery system

Using caches in the mobile network is currently not straightforward due to the tunnelling between UE and PGW. However, SIPTO allows UEs to access the IP layer using LGWs without need of traversing the EPC network (see section 2.5.3). This makes caching in the HGWs feasible. Section 2.5.3 presents a solution relying on MPTCP to provide seamless handover with SIPTO in case of mobility of the UE.

This scenario represented in Figure 27 can be described as follows:

- When a user sends a request to the content provider, the content provider sends a message to CC to initiate a session with a given QoS requirement.

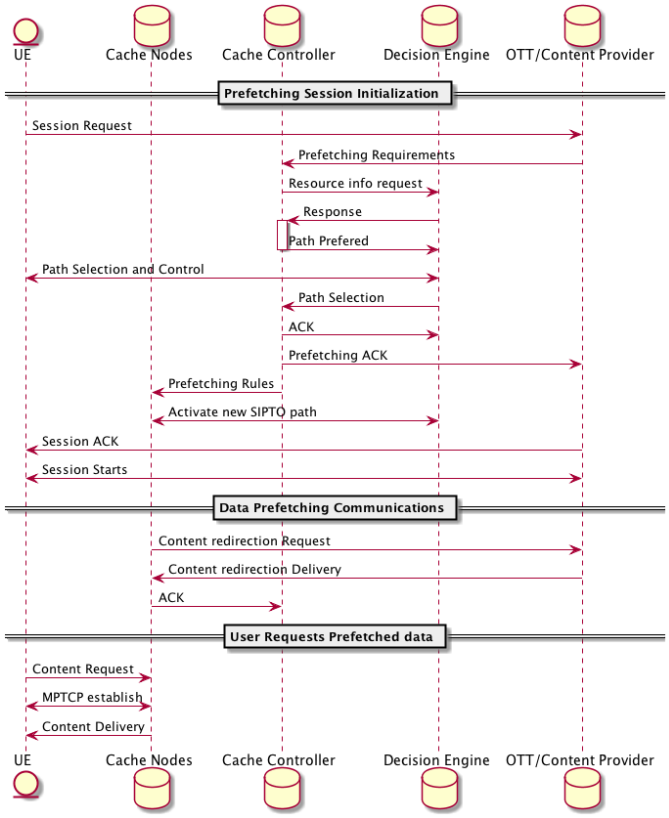


Figure 27: Prefetching interaction with MPTCP and SIPTO solutions

- In order to be able to make an optimised caching decision, the CC asks the DE information relative to user location and network performance
- When the CC receives this information from the DE, it may apply existing prefetching rules to store the requested content in some cache nodes.
- The DE selects an interface by taking into account the preferred path selection forwarded by the CC. This decision of where the end user will connect is sent back to CC that will acknowledge the session request to content provider and the end user.
- Whenever the DE detects some session events and has to take an interface switching decision for an end user, it should advise the CC that will activate the necessary caches, and pre-fetch content.
- The OTT provider initiates prefetching and the path to the caching server has been selected by exchanging information between CC and DE.
- Then uDPM activates a SIPTO path to the appropriate cache
- There is a MPTCP sub flow establishment when the UE is redirected to the local cache for the final content delivery.

3.1.3 Content delivery in the centralised COMBO architecture

A CDN provider typically owns its servers or rents spaces from multiple DCs, called Points of Presence (PoPs).

In the centralised COMBO architecture, NG-POPs are located typically at the Core COs where the content distribution servers controlled by the network operator can be deployed. Content can thus be cached in NG-POPs. This architecture can be seen as a centralised CDN approach. By provisioning enough network connectivity, power supply, and servers at a limited number of DCs, one can assemble a very large aggregate CDN capacity at a relatively small number of DCs.

The centralised COMBO content deliver architecture can already reduce the traffic pressure between Core COs. However, statistics relative to access network [66] [67] show that almost half of the requests are cacheable and that traffic can be reduced by one third if content servers are deployed in Main COs. Furthermore, according to the Table 5, if putting down the content servers in Main CO, the latency and throughput can be efficiently improved, and thus a better QoS of end users can be achieved.

3.1.4 Content delivery in the distributed COMBO architecture

In the distributed COMBO architecture, the IP edge is located at the Main CO. This requires extending the span of the IP network. Then it allows deploying content distribution servers at the Main CO.

Measurements [65] [70] have studied the performance of deploying CDN servers in different locations and with different numbers. The study in [71] shows that doubling the number of peering points roughly doubles the aggregated throughput over a wide range of values and network topologies. However, the operating costs could be increased in the distributed COMBO content delivery solution comparing to the centralised solution. According to the RTT distribution in Table 5, the latency gets improved in the distributed COMBO architecture.

In the distributed COMBO content delivery solution, a collaborative caching algorithm can further improve the benefit. According to the study [72] of interaction of telco-CDN (ISP regional CDN), additional 12% to 20% of global traffic can be reduced if we use collaborative CDN caches located in different NG-POPs. However, in order to efficiently use the distributed content/caching servers located in the Main COs, it is necessary to have a more centralised control, and thus it seems preferable to implement the CC and the DE in the Core CO. As the DE is part of the UAG CP, this advocates in favour of implementing the UAG CP in the core CO.

3.2 Real Time communication services

This section analyses the impact of the UAG on the delivery of real time communication services, such as voice, video conferencing, real time streaming services (e.g. live sport events), multiplayer gaming or specific machine-to-machine applications.

3.2.1 Interfacing data and control planes for communication services

There are two topics to be discussed here: questions related to Universal Authentication and those related to Universal Data Path management.

3.2.1.1 Relation with uAUT

uAUT deals mainly with the initial phase of the services where the user is authenticated by the network or service provider. Although this phase is important from the authentication point of view, it is independent from the service delivery. Legacy fixed and mobile AAA systems and uAUT servers operate above the IP network, therefore the introduction of the UAG with a uAUT agent inside is transparent at service level.

However, since a user authenticated with uAUT can transparently use an access network that differs from the one used during the original authentication, it is possible to envisage implementing vertical handover (e.g. moving from a mobile access to a Wi-Fi access) seamlessly.

3.2.1.2 Issues related to uDPM

The impact of uDPM on the service delivery is high since it manages the data paths whereas traffic forwarding and session continuity are very relevant in real time communications services. Critical parameters for real time services include:

- Latency values requirements: latency for voice and video conferencing below 150 ms, real time streaming services below 400 ms and multiplayer gaming below 100 ms.
- Packet Delay Variation (jitter) requirements: jitter below 20 ms for voice and video conferencing, 50 ms for real time streaming services and below 50 to 100 ms for multiplayer gaming, depending on the game.

In all cases, the underlying network should be able to fulfil these requirements to successfully support real time services. Therefore, the uDPM should work within the limits mentioned above.

Other important consideration are:

- Handover from fixed network to mobile network when the access to the fixed network is lost: it may take about 150 ms to “wake-up” a mobile UE from the “sleep mode”.
- In case of packet based scheduling: either the uDPM or higher protocol layers should be able to harmonise the different packet delay times and perform reordering when necessary. An option would be to avoid packet based scheduling issues by performing session based scheduling only. This is possible only if the bandwidth of one channel is sufficient for the application.

Additionally, scalability issues might become significant for the placement of the uDPM functions. Typically, in a less complex real time scenario where one only needs to take into account a few parameters of involved networks, a centralised approach is adequate (as e.g. in legacy mobile networks). However, for more complex targeted network scenarios such as e.g. handover without service interruption between Wi-Fi and LTE, more network parameters need to be monitored and taken into account in the decision process. This is in favour of a more de-centralised placement of uDPM. This will be further elaborated in section 5.3.

3.2.2 Communication services in COMBO architectures

Distributed and centralised architectures could affect the service delivery of real time communications services in terms of latency, as the distance between the UE and the Core CO is higher than the distance between the UE and the Main CO.

However, the latency difference between both COMBO architectures (centralised versus distributed) is in the order of a few ms, which is not significant for the latency requirements identified previously. Indeed, a large part of the offset is due to the propagation delay, which is 1 ms for a 200 km distance. A more precise computation has been made based on a delay analysis for the German network (fibre propagation and switch processing delay considered), providing a difference in latency smaller than 5 ms (see section 3.2.2 of [4]).

For ultra-low latency services, such as virtual reality, augmented reality, etc. which require a latency in the order of tens of ms, a distributed architecture may be more suitable if the application servers are co-located in the Main CO or close to this location.

Nevertheless, a centralised NG-POP is compatible with very low latency services as well if new approaches such as Mobile Edge Computing (MEC) are followed [46] (section 3.3 provides more information regarding cloud services). MEC can work in two ways: an end-point mode, where the MEC servers, typically located at the mobile sites, terminate the user connection and pass-through, where the MEC servers are in the middle of the DP, allowing the applications servers to be located higher in the network for services supporting a higher latency.

3.3 Supporting Cloud based services

Cloud computing are essential parts of the future network as described by both Cisco's Internet Business Solutions Group (IBSG) list of technology trends [39] and Bell lab's future view of the networks [40]. As the access networks become capable of higher data rates, using e.g. FTTH or G.fast, the networks will be able to accommodate interactive real-time content. When this technology shift becomes more pronounced it will affect the traffic patterns over the networks.

Another clear trend listed in both [39] and [40] is that the user becomes more mobile, both in the fixed network using e.g. Wi-Fi, and in the mobile networks. Currently, the fixed network traffic widely exceeds the mobile traffic, but this is about to change and the annual traffic growth is essentially higher for mobile networks than for fixed [41]. When the performance and price differentiator between fixed and wireless access diminish, end users will become less conscious of how they access the Internet.

To support novel services and technologies the edge cloud-computing paradigm has been proposed. The paradigm goes by many names such as *Fog Computing* [42], *Telco-Cloud* [43], *Mobile Cloud* [44] [45] or *Mobile Edge Computing* (MEC) [46]. Nevertheless, edge cloud computing complements the prevailing centralised infrastructure by distributing cloud-computing capacity through the core and access networks. For example, the neighbourhood's required compute capacity is aggregated in a shared DC, accessible at low latency with a smaller global traffic footprint.

3.3.1 Interfacing data and control planes for cloud based services

The network requirements from DP functions vary widely between different cloud services. Some emerging services, such as cloud gaming or augmented reality, require ultra-low latency in the data path, while others have more traditional service requirements. Consequently, the location of the VNF servers will also have different requirements, whereas the low latency services must be implanted close to the user and others can afford to have longer distances. However, the CP function implementation does, in most cases, not have tight delay requirements and can thus be implemented further from the user. This will allow for a more efficient network management when the traffic optimisation can be performed over a larger area. Hence, for cloud services the split UAG implementation can be advantageous for cloud based services.

3.3.1.1 Relation with uAUT

uAUT acts in the first phase of the access to the network when the user is authenticated by the network or by the network operator. As it has been mentioned before, the uAUT is transparent to a service as cloud computing.

However, since a user authenticated with uAUT can transparently use an access network that differs from the one used during the original authentication, it is possible to envisage implementing vertical handover (e.g. moving from a mobile access to a Wi-Fi access) seamlessly.

3.3.1.2 Issues related to uDPM

For cloud-based services with tight latency and jitter requirements such as cloud gaming or augmented reality, the cloud servers should be as close to the users as possible. Since in case of multipath connections, the cloud servers must be located beyond the path split, i.e. the MPE which is part of the Session Mapping Execution function, this implies that the uDPM DP should also be as close to the users as possible. This implies that the UAG DP should preferably be located at the main CO.

The role of the CP for uDPM in a cloud service involves the “Path coordination and control” and the “Decision engine”, and “Data path creation and destruction” for the case of user mobility. Especially for the latter, it is advantageous to work over a larger part of the network, within limits of scalability. Hence, for the CP it is preferred to locate it higher up in the network, e.g. co-located with the Core CO.

So for low latency cloud services it is advantageous to have a split architecture, with the DP functions at the Main CO, and the CP functions at, or beyond, the Core CO. For services that can cope with higher delays, in the order of 100s of ms, it is not necessary to split the UAG structure physically.

3.3.2 Delay requirements for delay-critical cloud applications

To investigate the positioning of the NFV servers in the network, the first step is to find reasonable requirements on the network, in terms of latency. The goal of the cloud trend is that all compute and storage resources move from the homes to the cloud. We can be rather certain the future eventually looks like this as IT and

telecommunications services are commoditised and the end users' time, interest and skill to acquire and maintain electronics fade away.

The ultimate goal of cloud computing is to move all computer power from the UE to the cloud infrastructure at the DC. The UE would thus be mainly a terminal with interfaces for the user, such as keyboard, mouse pointer and touch screens. Anything more advanced than moving the pointer on the screen would be performed within the DC. Consequently, all updates of the screen would be performed in the cloud and sent to the user equipment as video.

Real-time applications naturally impose strict latency requirements on the connectivity between the DC and the user. To provide the reader with a contrast, for example, cloud-based word processing services updates are fairly low rate and thus transmitting a new image every second is not crucial. However, when considering more time critical services such as gaming, the requirement goes way beyond what contemporary cloud infrastructures can deliver.

The notion of employing pervasive and ubiquitous computing for all our computing needs is coming to fruition. We have had real-time collaborative cloud applications such as word processing and spreadsheets for a while. Cloud gaming [47] is still in its infancy but there are examples of quite mature tests implemented, e.g. [48], that are pushing the boundaries of what is possible with contemporary infrastructure. In the next section the delay requirement for the most time critical services is estimated.

Different cloud applications have different network performance requirements. For example, in cloud based word processing the screen is essentially a still image and changes every time a key is pressed. This requires an average of at most a couple of updates per second. It is simply a matter of asynchronously replicating the keystrokes, requiring a quite low data rate, relatively insensitive to delays. At the other end of the scale are real time applications like cloud gaming. Gamers are notoriously sensitive to delays, which set higher demands on the networks. Since it is one of the most demanding services, to study the network requirements for cloud services, we use cloud gaming as our reference model for delay requirements.

For cloud gaming, it is first important to make a technical distinction between on-line gaming and cloud gaming.

In on-line gaming the gaming hardware is co-located with the gamer and all computation is done locally. Somewhere on the Internet there is a server that for example aggregates the player’s movements, actions, and scores. This data is asynchronously transmitted to all players, which is then fed back into the game dynamics on the local hardware to reflect the current state of the game. In cloud gaming the gaming hardware is in the cloud where all computation is done. In both cases, a large delay between either the on-line gaming server of the cloud gaming server will result in inaccurate local representation of the game state.

Typically, gamers are worried about two parameters, the ping time and FPS (frames per second). The ping time is the RTT to the server, and if this is too high the screen image is not corresponding to the view in the server, referred to as lag. The movements are relatively slow and normally it is not a problem with a ping time of 70-80 ms [49]. The FPS refers to how often the screen is updated, and can thus be viewed as the sampling frequency of the game. Depending on the type of game the

measure should be between 30 and 60 Hz. Slow games like simulation games and some strategy games can cope with the lower update rate, while more time critical games like First Player Shooter Game (FPSG), Third Player Shooter Game (TPSG) and racing often require rates of up to 50 Hz or 60 Hz. Normally role playing games, like Massive Multi-Player Online Role Playing Game (MMORPG), are in the middle requiring 40-50~Hz update frequency.

To formulate realistic delay requirement bounds we consider the FPS as the sampling frequency of the system, i.e. the game itself. A low sampling frequency and a high delay will have the same effect on the QoE; the game will not run smoothly and the reaction time will suffer. To study the sources of delay we turn our attention to the intermediate network. Figure 28 depicts the cloud gaming network architecture. The UE is attached to the RGW, which is connected to the Internet. In case of a mobile access, a specialised router typically replaces the RGW, but since games are often played in the home, we retain the term RGW for both fixed and mobile access. The game server (GS) is located somewhere in a DC, and constitutes the equivalence of either a gaming console or computer. The GS can then be connected to an on-line server for multi-player games, as usual.

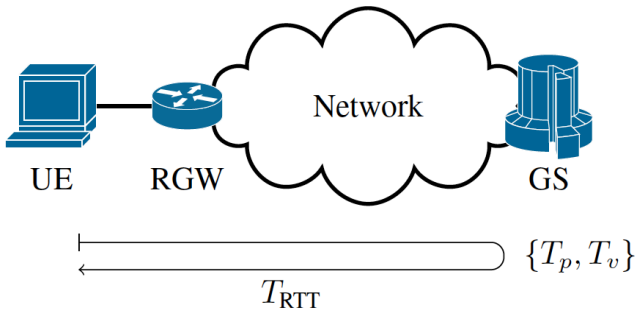


Figure 28 Cloud gaming architecture and delay.

The RTT in the system for a cloud gaming application is the time *T* from a user interaction to its effect being shown on the screen. That means first the signal is sent to the GS, where it is processed and the screen video updated, encoded and sent back to the UE. The time for transmission in the network is denoted by T_{RTT}. Denoting the processing time for the game by T_p and the total video coding by T_v, the total gaming loop delay is given by

$$T = T_{RTT} + T_p + T_v$$

If this total delay is at maximum $T_{max} = \frac{1}{FPS}$, the update will be not be delayed more than one sample. This is also the minimum delay that can be guaranteed for a sampled system. Hence, it is reasonable to assume that the user will not be able to notice the delay, in terms of degraded QoE, if the network delay does not exceed T_{max}. Table 6, shows typical types of games with their requirements on FPS and what it implies in terms of maximum delay.

The values in column T_{max} in the table can be considered as the requirements to achieve no noticeable delay in the loop, which should be satisfactory even for skilled gamers. For average gamers the delay can probably be set slightly higher without

any considerable quality degradations, reflected in the column $2T_{max}$. Thus, the quality should still be satisfactory as long as the delay is less than two screen updates.

Table 6 Table of maximum tolerable delay for different types of games.

FPS	Type	T_{max}	$2T_{max}$
30	Simulation, building	33 ms	67 ms
40	Sport, MMORPG	25 ms	50 ms
50	TPSG, FPSG, racing	20 ms	40 ms
60	FPSG	17 ms	33 ms

A requirement of 20 ms RTT for the gaming loop means that the DC can be placed in, or at the same level as, the Main CO, where the network RTT is in the order of a few ms from the UE when considering fibre, LTE or G.fast access technologies. This time can be considered to be negligible for the total time allowed. However, by placing it in the Core CO there is a risk that a RTT equal to or larger than 5 ms will be noticeable at the end-user's side, since such a value cannot longer be considered very small in comparison to 20 ms.

3.3.3 Cloud based services delivery in the COMBO architectures

On the verge of Internet services *cloudification*, network operators are looking for new strategies to reduce both capital and operational costs. This section describes how that NFV, a novel network architecture paradigm, can help network operators achieving their objectives in terms of costs in the future Fixed Mobile Convergence (FMC) networks. NFV is based on the concept of network functions: a network function in this context is an abstract building block performing a specific task.

Examples of network functions are Firewalls, Traffic Monitors, etc. So far, network functions have been implemented using dedicated hardware, usually referred to as *middleboxes* that are able to handle very high traffic load but are expensive and inflexible. NFV allows a move towards a *softwarization* of network functions in a virtualised environment [24]. Multiple VNFs can thus be instantiated and consolidated in the same Commercial-Off-The-Shelf (COTS) hardware that can potentially be placed in any powered location of the network. NFV also eases service deployment by exploiting the concept of Service Chaining introduced below. On one side, the optimal solution in terms of costs for a network operator would be placing the VFNs in a DC to provide all Internet services from a centralised cheap location. On the other hand, this solution might not be convenient for latency-sensitive SCs and may cause performance degradation due to excessive distance of the DC that, in some cases, can be even thousands of kilometres far from the users. Hence, the only solution is to place the VNFs closer to the users at the edge of the network.

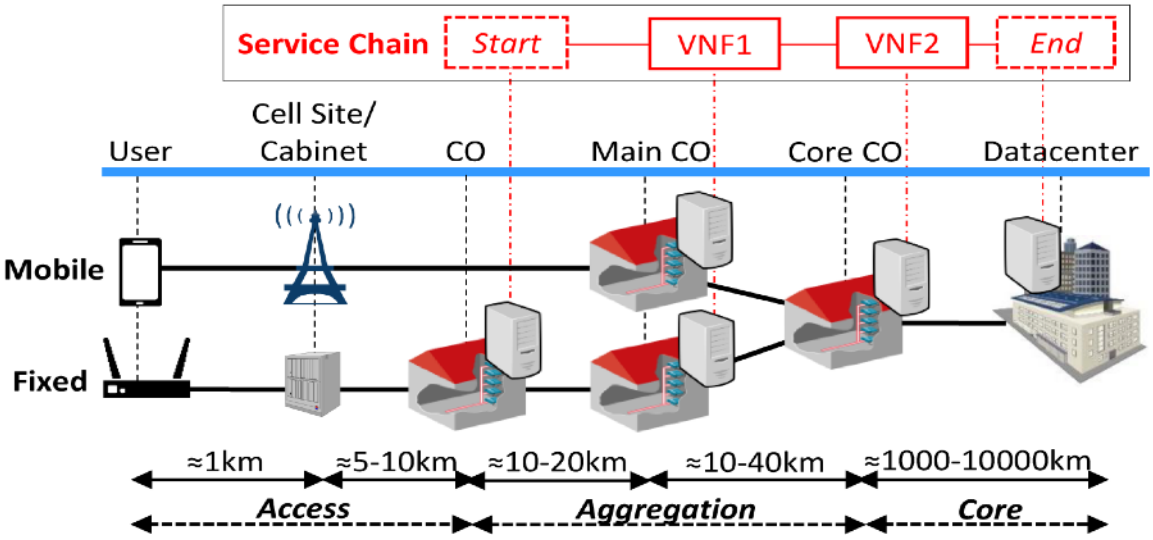


Figure 29: NFV-enabled fixed and mobile aggregation networks

The main locations to host VFN in the metro/access segment are COs at different hierarchical levels of the fixed and mobile aggregation networks (i.e., COs, Main COs and Core COs) as depicted in Figure 29. Several research efforts have targeted the definition of novel architectures for FMC networks, where fixed and mobile networks are jointly designed and optimised both from a functional (i.e., by unifying network functionalities) and structural (i.e., by sharing network infrastructures) perspective [35]. Therefore, the main objective of network operators when deploying VNFs is to find the optimal placement of VNFs that maximizes the consolidation of the VNFs. This means placing the VNFs in the minimum number of NFV nodes while meeting the end-to-end latency requirement for the service chains (SC) while satisfying processing capacity constraints for the NFV nodes. A SC is a sequential concatenation of VNFs providing a specific Internet service. Note that a VNF can be shared among multiple SCs by properly scaling up its processing requirements.

3.3.3.1 Delay requirements for Virtual Network Functions

Cloud gaming as described in the previous section is one of the most time sensitive network applications. There are others with similar behaviour, such as cloud based applications on virtual reality (VR) and augmented reality (AR), which have similar requirements [34]. In Table 7, these applications are named 5G services (5GS).

Table 7 considers a set of five different applications. It associates delay requirements for these five applications represented as chained VNFs. The SCs have been elaborated by considering similar approaches described in [82]. Web services have the loosest requirements with tolerable delays up to 500 ms. VoIP and Video conferencing (VC) systems have delay requirements of 100 ms and 80 ms, respectively, while online gaming requirements are set to 60 ms. In [49] it is seen that a noticeable effect on the game experience can be seen for fast games for a RTT in the order of 70-75 ms. Taking into account also the status the RTT, or ping time, has in the gaming community a reasonable maximum value is 60 ms. Each of the application corresponds to a specific SC. The VNFs used in the SCs in Table 7 are NAT: Network Address Translator, FW: Firewall, WOC: Wan Optimiser Controller, IDPS: intrusion Detection Prevention System, VOC: Optimisation Controller, TM:

Traffic Monitor. Each SC chains different VNFs in a sequential order, and is associated to a different end-to-end latency requirement. The VNFs are associated to a processing requirement per user, obtained by middleboxes datasheet.

Table 7 Details of the SC deployed and bandwidth and latency requirements.

Service	Chained VNFs	Latency req.
Web Service (WS)	NAT-FW-TM-WOC-IDPS	500 ms
VoIP	NAT-FW-TM-FW-NAT	100 ms
Video Conferencing (VC)	NAT-FW-TM-VOC-IDPS	80 ms
Online Gaming (CG)	NAT-FW-VOC-WOC-IDPS	60 ms
5G Service (5GS)	NAT-FW-TM-WOC-VOC	20 ms

3.3.3.2 Assessment of latency performance for cloud based services

In this subsection the case study used to assess the impact of latency on VNF distribution is briefly described. For a more detailed representation of the system model, the problem statement and the heuristic algorithm used to solve the VNFs embedding problem the reader is referred to [26]. The physical topology considered is shown in Figure 30. The coverage area of this network is in the order of the surface of a large European metropolitan city and its dimensioning is based on the urban geotype for the 2020 metro access reference network proposed in [29]. Two network architectures were compared: FMC (converged) and No FMC (non-converged).

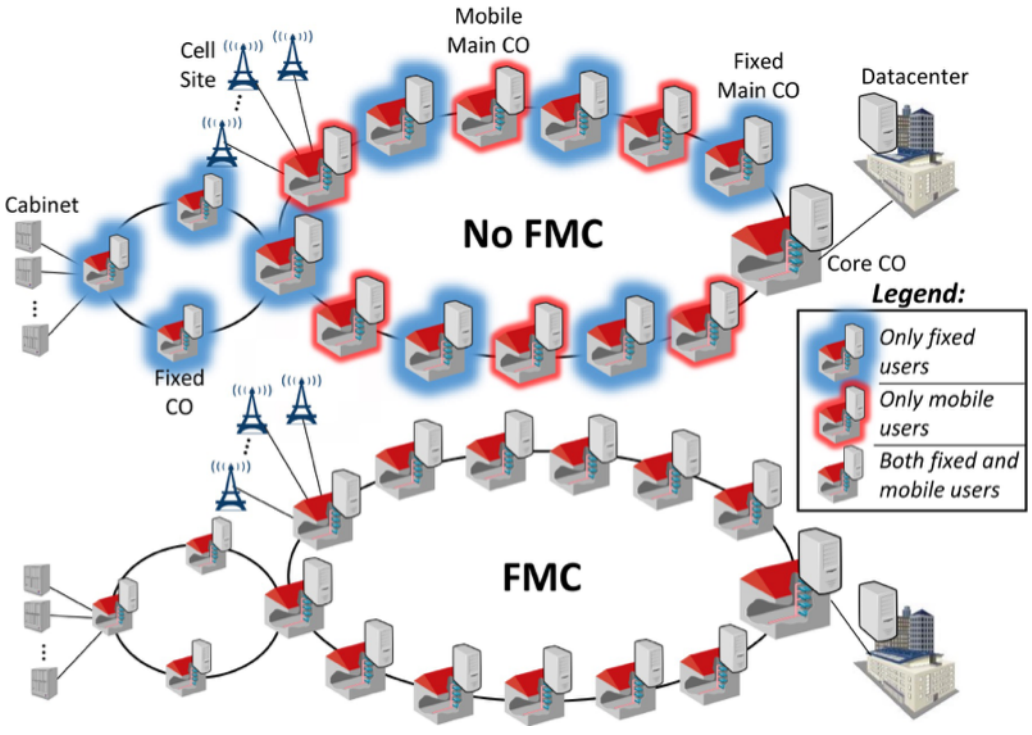


Figure 30: NFV node accessibility in FMC and No FMC architectures

In both architectures, fixed and mobile users can access VNFs placed either in the core CO or in DCs. In the non-converged architecture, fixed (respectively mobile) network users can also access VNFs that are placed into the fixed (respectively mobile) network infrastructure, i.e., in Fixed Main COs (respectively in mobile Main COs). Main COs are thus referred to as fixed (respectively mobile) NFV nodes. Note that, although in legacy LTE architectures IP flows are tunnelled up to a PGW, which is typically located in the core network, we assume here that Mobile Main COs represent the IP edge towards the mobile core network. Thus, we consider that the Mobile Main COs are provided with a PGW and a server where only mobile traffic-related network functions can be virtualized. In the converged scenario fixed and mobile users can share the network infrastructure and, thus, also the NFV nodes that are all converged Main COs.

Three different DC location configurations were taken into account: Close DC, Midrange DC and Very Far DC with latencies equal to 15, 75 and 150 ms, corresponding to a national, continental and intercontinental DC location. Moreover, five different homogeneous scenarios have been analysed in order to evaluate the impact of latency of different applications on VNF consolidation. At each iteration a single SC out of those given in Table 7 is embedded in the network. Each instance run is performed comparing three different percentage of local traffic terminating in the metro network: 0%, 50%, and 100%. The first setting (0%) represents the case where all the SCs have as destination point the DC location. In the second setting, (50%) half of the SCs have as destination the Core CO in the metro network and the remaining half terminate at the DC location. Finally, in the last setting (100%) all the SCs terminates at the Core CO (i.e., at the edge of the metro network).

Figure 31 shows the number of NFV active nodes for the various traffic configurations and network architectures discussed so far.

For the Close DC configuration, we observe that the most convenient solution in terms of VNF consolidation is to host all the VNFs in the DC for every homogeneous scenario, every architecture and every percentage of local traffic, except for the 5GS SCs for 50% and 100% of local traffic. In these cases, it is required the activation of some fixed/mobile Main COs and of the Core CO. In fact, for all the SC types but the 5GS, consolidating the VNFs in the DC, even though part or all the SCs terminate in the Core CO, is a feasible solution because the Round Trip Time to the DC (30ms) does not affect the latency requirement of the SCs. This is not true for the 5GS homogeneous scenario due to the very strict latency requirement of its SCs (20ms). In this case and in conditions of local traffic, placing all the VNFs in the DC would degrade the performance. For this reason, distributing the VNFs in the metro/access network is necessary to meet latency requirements for the SCs terminating in the Core CO. For the Midrange DC configuration, only the VNFs for the WS scenario can be all consolidated into the DC. Finally, for the Very Far DC configuration, only the WS homogeneous scenario can still be guaranteed by placing the VNFs in the DC for all the traffic conditions. For the other scenarios, the only way to meet latency requirements is to have all the VNFs placed in the metro/access network and to keep all the traffic local (100%).

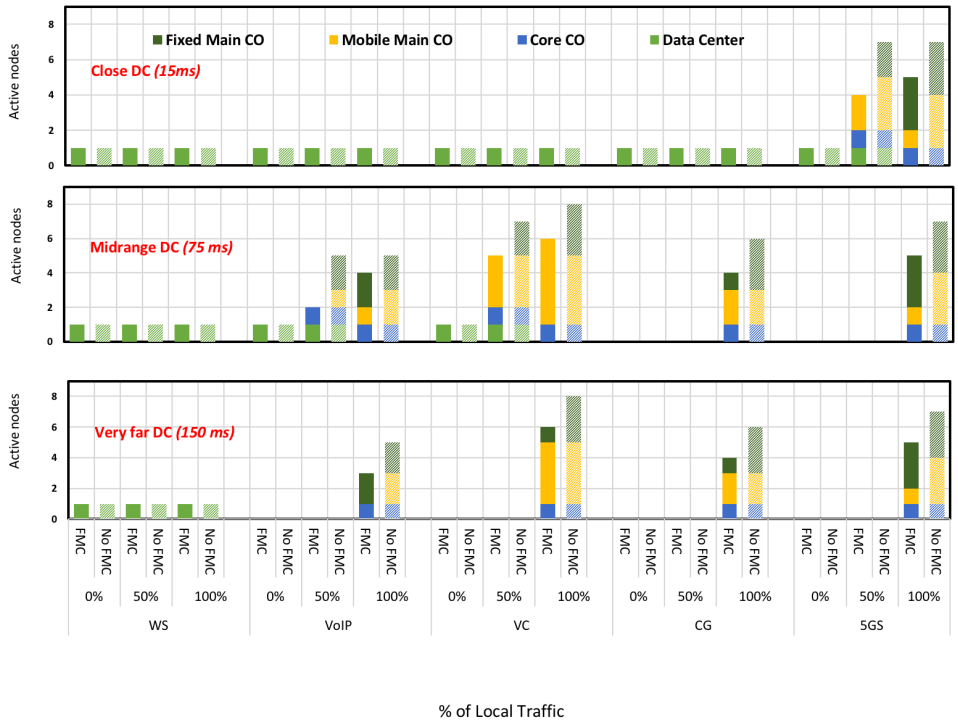


Figure 31: Numbers of active NFV nodes in converged and non-converged architectures

Moreover, the impact of latency on VNF consolidation is similar for both converged and non-converged architectures. However, when the VNFs are distributed in the metro/access network, the FMC architecture requires from 30% to 60% less NFV active nodes than the No FMC one. This means that the adoption of a FMC metro/access network can consistently improve the consolidation of VNFs. Further considerations on processing requirement description, can be found in [26].

3.3.3.3 Call Flow for VNF Placement

The aim of this subsection is to provide an overview of the call flows needed to instantiate and terminate SC instances. Every SC chains together multiple VNFs. In general, a Network Service is provided by multiple SCs. In this document, without any loss of generality, we assume that a single SC can provide a Network Service and, thus, SCs and Network Services coincide.

The starting point for the call flows defined in this document is the ETSI NFV MANO (Management and Orchestration) document [36]. We consider the centralised COMBO architecture, with split UAG and a co-located DC that accommodates either only the UAG CP (*partial model*) or both UAG CP and DP (*full model*) as in section 2.3.3).

In Figure 32 we show which are the functional entities involved in the SC instantiation/termination and show the main steps of SCs instantiation. The functional entities involved in instantiation/termination of SCs are:

- **Sender:** It is the entity requiring the instantiation/termination of the SC. ETSI NFV MANO identifies this entity with the *OSS* (Operation Support System),

which receives a service request (i.e., the deployment of a SC) between two end points.

- **Virtualised Infrastructure Manager (VIM):** The VIM is responsible for the partial control, management and allocation of virtual resources (i.e., compute, storage and network) of the Network Function Virtualisation Infrastructure (NFVI) when a VNF must be instantiated. In our vision a VIM acts inside a *NFVI-POP*, i.e., a physical location where virtual storage, compute and network resources are available (e.g., Main CO, Cabinet etc.). The overall NFVI is composed by multiple NFVI-POPs.
- **NFV Orchestrator (NFVO):** The Orchestrator coordinates NFVI resource management among different VIMs. Moreover, it is responsible for the lifecycle management of the SCs.
- **VNF Manager (VNFM):** The VNF Manager is responsible for the lifecycle management of the VNFs

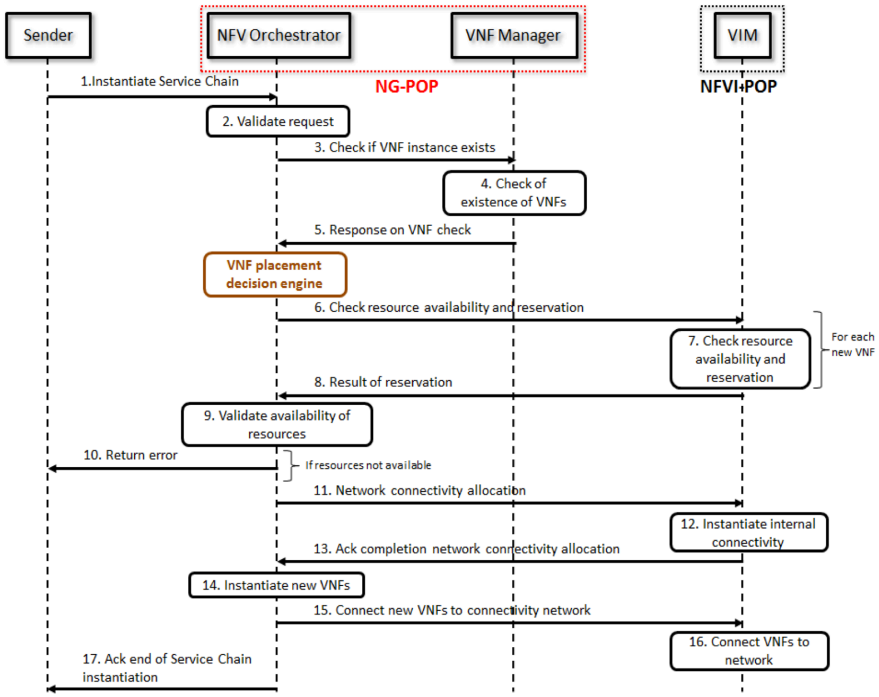


Figure 32: Service Chain instantiation operations according to the ETSTI MANO framework

In general, while multiple Virtualised Infrastructure Manager functional entities are distributed across the FMC network in the different NFVI-POPs, we assume that the NFV Orchestrator and the VNF Manager functional entities are both located in the NG-POP. According to the ETSI NFV MANO document, while instantiating a new SC, there are three possible variants:

a) None of the VNFs to be chained already exist.

b) All the VNFs to be chained already exist, because they have been previously instantiated by other SCs.

c) Part of the VNFs to be chained already exists.

Note that in the second and third cases the already existing VNFs might scale up (a single node gets more powerful to support more demand by itself) or scale out (more nodes of the same power are added to meet demand) to accommodate traffic from the new SC. However, in this document we do not focus on this aspect, and we assume that the allocated resources for existing VNFs are enough to sustain the new SC.

3.4 Supporting the IoT deployment

The basic idea of the Internet of Thing (IoT) is "the pervasive presence around us of a variety of things or objects – such as Radio-Frequency IDentification (RFID) tags, sensors, actuators, mobile phones, etc. – which, through unique addressing schemes, are able to interact with each other and cooperate with their neighbours to reach common goals" as defined in [7].

A *thing* in the Internet-of-Things can be defined as a physical or virtual entity that exists in space and time and is capable of being identified. In line with the vision of "anything connection" (ITU 2005), practically any smart object, either physical or virtual, could become a connected thing in IoT.

Authors of [13] define smart objects (or things) as entities that:

- Have a physical embodiment and a set of associated physical features (e.g., size, shape, etc.).
- Have a minimal set of communication functionalities, such as the ability to be discovered and to accept incoming messages and reply to them.
- Possess a unique identifier.
- Are associated to at least one name and one address. The name is a human-readable description of the object and can be used for reasoning purposes. The address is a machine-readable string that can be used to communicate to the object.
- Possess some basic computing capabilities. This can range from the ability to match an incoming message to a given footprint (as in passive RFIDs) to the ability of performing rather complex computations, including service discovery and network management tasks.
- May possess means to sense physical phenomena and/or to trigger actions.

Machine-to-Machine communication (M2M) is a close concept. There is no general agreement about the difference between IoT and M2M although M2M is generally considered as more focused on the communication between devices in the same system and IoT as trying to foster cooperation and interactions between disparate systems (e.g. http://www.pubnub.com/blog/iot-vs-m2m-understanding-difference/).

In this document we do not make any difference between M2M and IoT. We use the word *device* to refer to an object that has communications means.

3.4.1 Families of applications

As noted in [10] we can differentiate remote measurements and remote control though the same device sometimes manages both functions. *Remote measurements* refer to sensing physical phenomenon, storing, sending, receiving and processing of measured information. *Remote control* of devices includes access control, sending, receiving and processing of control commands.

Expected applications and the associated requirements are shown in Table 8. Possible connectivity solutions are given in the last column. Note that when 2G is mentioned, specific adaptation of LTE for M2M (LTE-M) could be also used

Applications	Number of devices	Daily traffic per devices	Critical aspect	Device Lifetime	Coverage requirement	Preferred Connectivity solution
Smart City services	++	-	Power, cost	up to 10 years	urban to underground	Several Dedicated network
Smart metering	+++	-	Power, cost	10 to 20 years	deep indoor underground	Dedicated
System monitoring	+	+	Cost, power	Several years	deep indoor	2G/4G (on main) None (on batteries)
Connected Car / Fleet mgmt.	+++	++	Cost, roaming	10 years	mobility + roaming	2G
Logistic - Asset tracking	+	-	Power, cost & coverage	month to years	mobility + roaming	RFID - none
Smart Grid	++	+	latency	10's of years	Underground	None
E-health	+	++	reliability	years		Wi-Fi / 2G / 3G /4G (LTE-M)
Security - Public safety	-	+++	Throughput - reliability	years	Urban	2G / 3G / Wi-Fi /4G (LTE-M)

Table 8: IoT connectivity requirements per application type

As presented in [14] and identified in the METIS project [12], two different Machine Type Cellular (MTC) types are considered: Critical MTC and Massive MTC.

Critical MTC types (also called uMTC for ultra-reliable) are present in Industrial applications such as smart grid, traffic safety and control and “Tactile Internet”[3]. The

[3] Tactile Internet is defined in [ITU] by having extremely low latency in combination with high availability, reliability and security. Applications range from Robotics and telepresence, Virtual and augmented reality to healthcare, road traffic, serious gaming, education and culture, smart grid.

communication network should be ultra-reliable, have very low latency (< 1ms) and offer very high availability. Reliability is defined as to the capability of guaranteeing successful message transmissions within a defined delay budget. In [9], 0.9999 is given for critical MTC. Availability is the proportion of time a device can use the service. In [9], 0.999 is considered but in 0.99999 (five 9) is mentioned in [8]. Robust transmission, Fast Channel Assignment and Multi-level diversity are required to guarantee the expected level of availability and reliability.

Massive MTC (mMTC) types are represented by a large set of applications based on sensors and actuator. The constraints are different from critical MTC: low cost, low energy (>10 years on AA battery [14]), generally no strict time constraint and small data volume (20 to 125 bytes according to TC11 [11]) but with a massive number of terminals (tens of billions of network-enabled devices, thus typically in the order of 100.000 per access point according to [12]). Protocols should thus have low overhead and be scalable.

3.4.2 Interfacing data and control planes for IoT services

3.4.2.1 Issues related to uAUT

The main security functions are: availability, authentication, confidentiality and integrity. The different types of application have not the same requirements regarding security. For many M2M applications, secure authentication is more important than keeping confidentiality. For example, it may not be confidential that a door is being unlocked, but it is crucial to accept an unlock-command from authorized senders only.

With uAUT proposed in deliverable D3.2 [3], it is possible to manage in a unified but flexible way a large variety of types of devices (unified management of information model and data model that allows the definition of multiple security policies). Specific front ends can be deployed in case specific security policies are required. For instance, if there are simple devices for which authentication and integrity are required but cyphering is not necessary; a dedicated front end with simpler procedures could be deployed. As mentioned in deliverable D3.2 [3], EAP (Extensible Authentication Protocol) is a good candidate to be the common basic protocol for all front-ends.

From an implementation point of view, the UDC database can be distributed. Hence, there is a dimensioning issue regarding front-ends and UDC to manage a huge number of connected devices but there is no specific scalability issue

One key question regarding IoT is to define how objects are identified. Object Identifiers (Object IDs) are different from addresses. They are used for uniquely identifying physical or virtual objects and not for routing. Possible object identifiers are Electronic Product Codes, UUID (Universally Unique Identifiers), MAC address, URI (Uniform Resource Identifier). It appears highly unlikely that a unique way of identifying objects will be adopted. Hence, the UDC approach with different front-ends is really well adapted.

Furthermore, COMBO proposes to make a difference between the subscriber, which is the entity or the person who pays for the access to the service, and the user, who

or which uses it. This distinction is important in the context of IoT: each user is a given device and the subscriber is the company that deploys the devices (for example, the management of a fleet of vehicles)

3.4.2.2 Issues related to uDPM

Critical MTC require a high availability (e.g. 99,999%). In most case, such a high availability can only be guaranteed if diversity is provided. This diversity can be intra-technology (e.g. each area is covered by at least two LTE base stations) or inter-technology (a device can use either Wi-Fi and LTE). In the latter case, the uDPM function is required to provide a transparent switching from one technology to another.

In a lot of cases, both critical and massive MTC are characterized by sporadic transmissions of short amounts of data. Therefore, the location of the DP of the UAG is irrelevant for IoT services. If there are gateways between the access point (LTE base station, Wi-Fi access point, etc.) and a set of devices close to each other, the traffic can be aggregated by these gateways. In that case, the access method can be connection oriented because only one connection is required for each gateway. This point is related to the addressing/identification question. M2M only requires that each device has a unique identifier (e.g. URI). The gateway can be the only one to have an IP address. Hence, from an IP point of view there is only one session for all the devices connected to the same gateway.

If devices are directly connected to the network, connectionless methods should be preferred in order to avoid a large overhead of connection signalling. It should be noted that standard LTE (i.e. up to Release 12) is not adapted because a four-phase handshake should be made on the radio interface before any transmission of data after a typical 30-second inactivity of the terminal and some connections in the network should also be set up. Hence, the radio access protocol should be modified, as is currently studied for example within 3GPP and the LORA alliance (https://www.lora-alliance.org/), but this is out of the scope of the COMBO project.

Table 9 computes, for different geotypes, the number of requests per second that IoT services generate. We consider the area of a Main CO given in section 9 of D3.4 and a number of connected devices as given in 3GPP TR 45.820 and TR 36.888. These reports only consider urban environments. We reused the same statistical sources statistical sources [73] to get an estimation of the density of households in semi-urban environment. 3GPP considers a ratio of 40 devices per household. Note that if we consider smart cities applications, a lot of devices are not related to a given household but we assume here as a first step that the ratio between the number of devices and the number of households is constant. We consider between 1 message per device every hour and 1 message every 5 minutes (see TR 36.888). We consider that each message generates only one request to the UAG.

Table 9 shows that in all cases, the number of requests is less than 7000 request/second in the worst case (sub-urban with 1 message every 5 mn), which can be easily managed. Hence, UAG with CP in the Main CO can be used from a processing load point of view.

Table 9: Estimation of the number of requests for the UAG centralised configuration for IoT

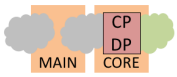
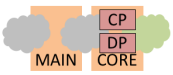
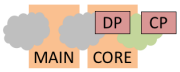
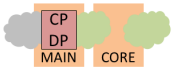
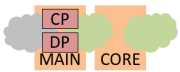
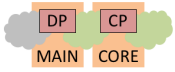
Avg. Geo-data of a typical Main CO area in Central Europe	Ultra DU	Urban	Sub-urban	Rural
Number of COs	1	2.9	5.9	10.8
Main CO area size	2 km²	15 km²	142 km²	615 km²
Household density (London for Ultra DU and DU, Wokingham for SubUrban, source [73])	4275	1517	330	30
Number of connected devices per household	40	40	40	40
Number of connected devices per CO	342 000	910 200	1 874 400	738 000
Number of connections per device per hour low-high (from TR 36.888)	1 - 12	1 - 12	1 - 12	1 - 12
Number of request per second to UAG	95-1140	253-3 034	521-6 248	205-2 460

3.5 Assessing the impact of UAG implementation option on service support

In this section, we briefly recap the qualitative analysis made previously regarding how each implementation option of the UAG on the two COMBO architecture impacts the service delivery. We focus on content delivery, real time communication, cloud based services and IoT support.

Table 10 provides a qualitative assessment of UAG architectural implementations regarding service delivery. For each service type the UAG architecture degree of fitness is stated for the CP (Command) and DP (Execution).

Table 10: Qualitative comparison of UAG implementations regarding service delivery

Centralised COMBO architecture			Distributed COMBO architecture		
Standalone UAG at Core CO (0) 	Split UAG with co-located DP and CP at Core CO (0) 	Split UAG with nonadjacent DP and CP and DP at Core CO 	Standalone UAG at Main CO (0) 	Split UAG with co-located DP and CP at Main CO (0) 	Split UAG with nonadjacent DP and CP and DP at Main CO 
Command of content distribution (3)					
+	+	+/-	+	+	++
Execution of content distribution (3)					
+	+	+	++	++	++
Command of real time communication services (6)					
+/-	+/-	-	+	+	+
Execution of real time communication services (7)					
+	+	+	++	++	++
Command of cloud-based services (1)					
+	+	+	+/-	+/-	+
Execution of cloud-based services (2)					
-	-	-	+/-	+/-	++
Command of IoT based services (4)					
-	-	-	++	++	-
Execution of IoT based services (5)					
++	++	++	++	++	++

++ = a very good fit for the architecture
+ = a good fit
+/- = some positive and negative aspects
- = a bad fit
-- = a very bad fit
IRR: irrelevant to this particular implementation

Notes:

(0) Implementing DP and CP functions in a single equipment versus co-locating them in the same location only differs in terms of scalability performance. Co-location is made possible thanks to SDN, and allows to manage the scalability of DP and CP independently from one another

(1): For cloud based services, both in terms of uAUT and uDPM, the CP functions can handle larger areas than the Main CO covers. See section 3.3.1.

(2): Especially for low latency services the NFV servers should not be positioned beyond the Main CO, where there is a risk of degradation of the QoS, due to network latency. See section 3.3.2.

(3): Deploying CP functions of content delivery services in Core CO allows to implement collaborative caching between different caches located in Main COs. DP functions deployed in Main CO will improve the reliability, scalability and performance of content delivery services. See section 3.1.4.

(4) Existing technologies are connection-oriented. The UAG CP should be able to manage the amount of signalling generated by massive MTCs. Thus, having the CP in the Main CO is recommended. See Section 3.4.2.2.

(5) The total amount of traffic for IoT is limited. All implementations of the DP can be used. See Section 3.4.2.2.

(6) uAUT and uDPM functions can be centralised, however, decentralised architectures in Main COs can provide better performance metrics for real time communications services and can handle more easily complex network scenarios. See section 3.2.1

(7) Real time communication services can be implemented following a centralised or decentralised architecture, however distributed architectures are more suitable to reduce the service latency and to enable complex network scenarios. See section 3.2.1.2

We can derive some global conclusions from the above qualitative analysis:

- The optimal location of the CP depends on the type of service to consider. Actually, when the network control has to interact with storage control (content distribution or cloud-based services), it seems more appropriate to locate the CP rather high in the network architecture, i.e. at Core COs. Such a location is acceptable also for real time communication services except for those requiring decisions that are complex or should be taken almost in real-time. The distribution of the CP functions between Main CO and Core CO could thus be an option: those corresponding to global control (e.g. the interaction with content distribution services) would be located in the Core CO, while those corresponding to fine grained control (e.g. access control) would be located in the Main COs.
- IoT support brings stringent requirements in terms of control traffic to be supported by the CP, which leads to recommend the distributed architecture, with the CP located at Main CO.
- The DP of the UAG should be located in the Main COs for all services; actually for IoT, there is no constraint related to the volume of information, but as the CP should be at the Main CO for IoT, it makes sense to co-locate the DP in the same location.

4 Network Sharing within a COMBO framework

Network sharing defined in the FMC context is all about sharing the physical network infrastructure required for supporting both fixed and mobile communications. Due to the legacy separation of fixed and mobile network infrastructures, network sharing for FMC can be a complicated and challenging task. Indeed, competing network operators need to relax their competitive concerns to adopt cooperative strategies. Network sharing presents many potential advantages such as cost-savings, fair and dynamic adjustment of resource sharing between operators and services, new use cases fostering cost sharing of resources depending on different conditions (e.g., pure resource usage or quality issues). Once the fixed and mobile network sharing is implemented in FMC, it will address two important challenges for network operators:

- Evolution in terms of network ownership.
- Operator relationships in a multi-operator FMC infrastructure attaining cost-savings from both perspectives CAPEX and more importantly long-term OPEX.

The main potential benefits of network sharing are

1. reduced OPEX through consolidation of existing networks,
2. reduced CAPEX for deploying new networks and technologies as a higher utilisation of shared resources based on sharing the rollout gains can potentially be reached in a converged architecture,
3. energy saving by using the same resources for diverse types of access networks,
4. increased competitiveness and improved time to market for the deployment and operation of new networks and technologies,
5. increased customer satisfaction since more services may be offered and be made available to more customers across a larger geography than may otherwise have been possible in a non-converged infrastructure.

The present chapter first lists various types of network sharing, which differ in terms of which network segment is shared. The next section describes how network virtualisation techniques such as NFV can be used to logically split a given architecture between “slices” that can be operated independently from one another. A “network slice” as defined by 3GPP is a collection of logical network functions that supports the communication service requirements of particular use case(s). Network resources are thus logically partitioned into “slices” that can be isolated from one another.

Then two network sharing cases are addressed:

- The first one is related to infrastructure network sharing between different operators, some of them being virtual operators (service providers that operate a service on an infrastructure operated by a different entity). Sharing (at least logically) such an infrastructure allows roaming (subscribers from one network operator supported by another operator) and offloading (network operators

that collaborate in order to optimise how the resources from different access networks are utilised).

- In the second case, a single Fixed Mobile Convergence Operator (FMCO) owns and operates a network while client network or service operators are allowed by contractual agreements (SLAs) to share it.

4.1 Different modes of network sharing

There are multiple reasons for network sharing, but reducing roll-out and operating costs is the main one. Furthermore, network sharing often enables quicker network extension in a restricted investment context. Network sharing has also other beneficial effects for the long term, e.g. with a positive impact on the global energy consumption and the carbon footprint.

Network sharing is often implemented in legacy networks, especially for Mobile Access Network sharing, Wi-Fi Access Network sharing, Fixed Access Network sharing and Backhaul Network sharing. In all these cases, it is assumed a Virtual Network Operator (VNO) operates at least a core network, and relies on SLAs with one or several other network operators to extend its coverage. In some cases, the VNO is purely virtual and totally relies on other network operators to provide the necessary resources to fulfil the services provided to their subscribers.

4.1.1 Mobile Access Sharing

There is no single solution applicable to all operators in every situation, but various mobile access sharing options. These options could involve and combine different dimensions:

- Degree of the sharing: sites, passive infrastructure, active infrastructure, full sharing
- Reach of the sharing: rural only, rural and selected urban, country wide
- Type of shared technology: 2G, 3G, LTE
- Number of sharing parties: 2, 3, 4, etc.

The choice for Mobile Access sharing depends on the competitive context and on the networks already in place.

For example, Site Sharing is a case of Passive Sharing where only the physical sites are shared. Each operator keeps its' own antennae, base station, energy/air conditioning and backhaul infrastructure. Passive Infrastructure sharing is a more advanced Passive Sharing scenario where tower, feeders, antennas, energy and air conditioning are also shared between operators. Each operator however maintains its own base station and its own backhaul network.

Active Sharing is a more advanced form of sharing between two operators, with for each site a single base station-managing subscribers for both operators. Active Sharing can typically be implemented in two main modes:

- Multi-Operator Radio Access Network (MORAN) where RAN sharing is implemented with separate management of the radio frequencies. Each Base

station transmits multiple frequencies with dedicated Public Land Mobile Networks (PLMNs).
- Multi-Operator Core Network (MOCN) also includes spectrum sharing.

National Roaming is a roaming agreement signed between operators. A single operator builds the site and routes traffic from other operators to their respective core networks.

4.1.2 Wi-Fi Access Sharing

There is also no single solution applicable in every situation. The document presents two examples of Wi-Fi Access sharing, both examples are active sharing.

A fixed network operator provides Wi-Fi connectivity on the Residential Gateway. The RGW propagates at least two SSIDs: a private SSID that provides connectivity in the customer LAN and a public SSID that provides a community Wi-Fi service. The community service can be directly operated by the Fixed Network Operator or can be operated by a third party Wi-Fi operator. The community Wi-Fi traffic is tunnelled to a Wi-Fi Gateway in the core network which provides subscriber management.

The second example deals with public hotspots such as hotspots available in the airports or hotels. The local Wi-Fi provider propagates a single public SSID. The customer connects to the corresponding portal and then chooses its community Wi-Fi service provider. Traffic is redirected from the Wi-Fi gateway of the local operator towards the Wi-Fi gateway of the customer's Wi-Fi service provider. This network sharing solution can be seen as roaming.

4.1.3 Fixed Access Sharing

The rollout of a new fixed access network for Very High Broadband represents a huge investment for an operator, especially for deploying fibre in the access network.

The co-investment for the passive infrastructure is an emerging solution, where the ducts are shared between the customer site and a mutualisation point (shared cabinet or shared optical central office) in the access network. Each operator has its own access fibre up to the customers, deploys its own ANs in the mutualisation point and uses its own backhaul network. This kind of Fixed Network sharing is known as a passive sharing.

Another scenario is an active Fixed Network sharing, where an Operator 1 deploys a full access network (ducts, fibres, splitters, ANs, energy, and physical site) in a particular area. A second operator deploys a full access network in areas not covered by the Operator 1. Both operators sign reciprocal "bitstream access" in order to provide an access service with a broader coverage.

4.1.4 Backhaul Sharing

All Access Network Sharing scenarios can be combined with Backhaul Network Sharing, which is a natural consequence of both Mobile Access and Fixed Access sharing.

Backhaul Sharing provides an extension of the access network from the ANs (Base Station, OLT/DSLAM…) up to the core network (radio controller, mobile gateway, fixed gateway…). Backhaul network can be shared either end-to end or partially, depending mainly on the transport network availability.

Backhaul sharing relies on active techniques in order to be able to monitor the links and to deliver the Key Performance Indicators (KPIs) defined in the Service Level Agreement between operators. For example, Backhaul Sharing can be implemented with a full IP packet-based network or managed fibres (Point to point Optical Ethernet, GPON, WDM for BBU backhauling).

4.1.5 SLA-based network sharing

In some cases, a VNO could be fully virtual, without any physical network to operate, and would in that case completely delegate network control to another network operator. The VNO is in this case more a service provider than a network operator. This case corresponds to network sharing managed through a business interface only and is addressed here.

In an FMC network, the infrastructure merges the fixed and mobile network access and aggregation network units. In a multi-operator environment virtual operators agree with the FMC operator to use the underlying infrastructure. Commonly, virtual network operators (VNOs) can be fixed and mobile converged or mobile only.

In the following scenarios, all virtual operators are assumed to be “fully virtual” having no network infrastructure. They do not operate an IP network and have to fully delegate networking operation to other operators that manage their customers’ traffic on their behalf.

OTT companies and content providers have to make agreements with either VNOs or directly with FMC operators when requesting dedicated bandwidth resources for their applications. Figure 33 represents the infrastructure network sharing scenarios.

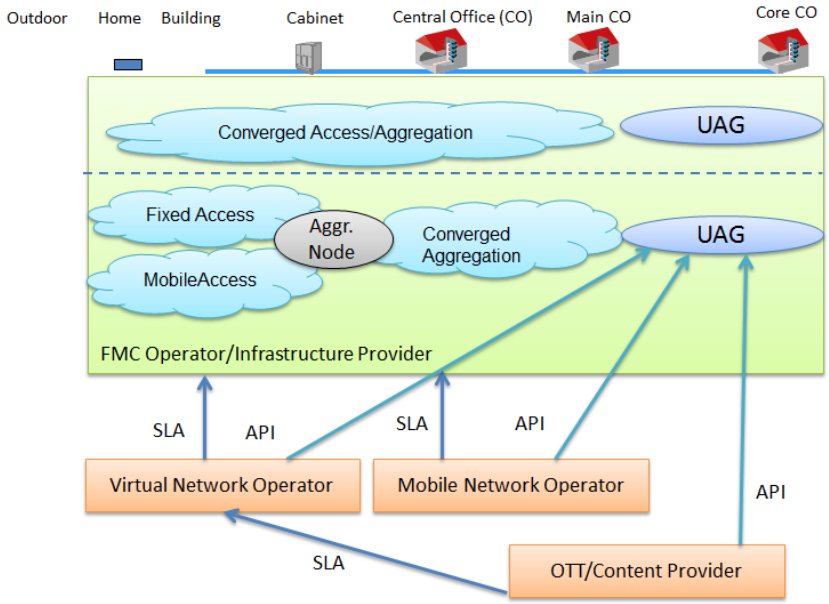


Figure 33: Actors of Infrastructure Network Sharing

A converged fixed and mobile access network is operated by a FMCO, which also is Infrastructure Provider for VNOs. The aggregation node can be deployed at different network levels, for example, at cabinet, CO, or Main CO levels. Sharing specifications are defined and declared through the Service Level Agreement (SLA).

For the operators sharing the physical network, an API has to be defined to enable the network control. VNOs can thus manage their SLAs with the FMCO as desired. Based on the selected business model, OTT/Content Providers could have access to the network management services as well, or they could rent network resources by making an SLA with the VNO.

Accordingly, in Figure 34 the converged access and aggregation network is operated by a single FMCO, while the VNOs can monitor their own virtual networks using monitoring information provided by the FMCO. The UAG is operated by the FMCO and may provide information either directly to the VNO through the API or indirectly through the management interface.

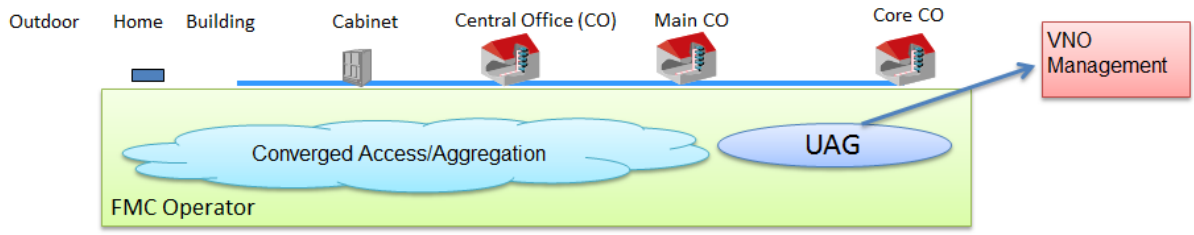


Figure 34: Network sharing with virtual operators

Over the Top (OTT) service providers and Content Providers take place in infrastructure sharing scenarios as shown in Figure 35. Here the FMC network is shared with others VNOs and OTT service providers. OTT sharing agreements include network resource reservation, QoS policies, bandwidth on demand requests and forwarding services to OTT premises.

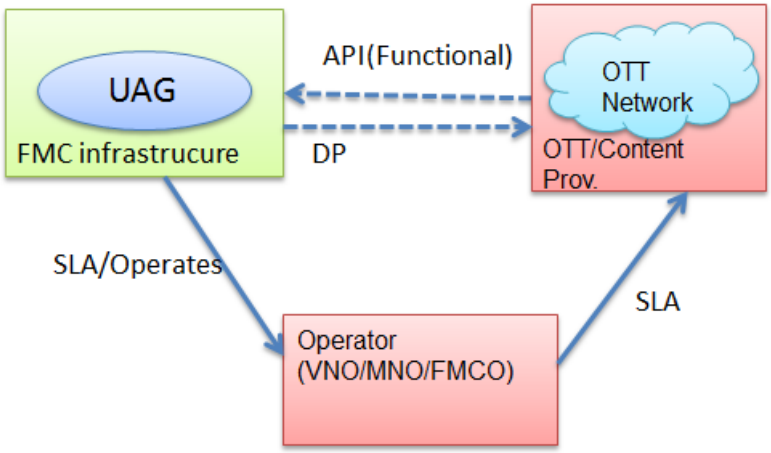


Figure 35: An OTT operator can lease converged network resources

4.2 Network Sharing relying on the network slicing approach

The concept of network slicing adopted here is that an operator (referred as “Physical Infrastructure Provider”) is able to partition its whole infrastructure to compose multiple and isolated “slices”. The infrastructure is composed of both network and cloud (IT) resources. Each slice is dedicated to a given set of applications / users / services / etc. With this approach, Virtual Network Operators can create their infrastructure (e.g. backhaul) on top of a slice built over the physical infrastructure.

4.2.1 Network Slicing Approach

Network Slicing goes beyond the general and classical notion of “multiservice” network since the targeted implementation considers the use of a dedicated CP entity, which differs among the different co-existing slices. To this end, the networking concept of virtualisation of infrastructure, SDN and NFV are adopted. In other words, the key idea is that all resources (networking and IT) are partitioned and abstracted enabling to compose and tailor an independent infrastructure dealing with the requested application / user / service demands or requirements. For the sake of completeness use cases for network slicing are virtual mobile network operators, OTT providers, IoT /smart grid applications, etc.

The use of network slicing within COMBO vision mainly pivots on how such a concept is applied considering the diverse exposed flavours for the UAG deployment. To this end, different criteria can be taken into account at the time of deciding how network slicing would be impacted by splitting UAG’s CP and DP, to locate the UAG at either Main CO or Core CO, or having a single CP entity with spread DP instances of the UAG entity. For the sake of simplification, we are considering the UAG’s DP resources are composed of both networking (i.e., forwarding, transmission and switching) as well as IT (computing, processing, storage), which may be co-located with the UAG.

Bearing the above in mind, deploying a single CP entity (e.g., SDN orchestrator) allows to better have a unified view of the resources for creating network slices on top of them. In other words, a UAG centralised CP may operate at finer granularity that fosters the partitioning and composing of network slices. However, scalability issues may appear depending on the network size. Furthermore, when the UAG’s DP (e.g., switching nodes and mini-clouds) are available at different locations of the network, this facilitates dealing with specific requirement of the network slices. For instance, if a network slice is requested and will be used to carry services with stringent latency demands, the DP resources can be allocated on the locations, which satisfy such requirements. Finally, the placement of the UAG’s DP and CP (i.e., Main or Core CO) will also noticeably impact on the network slice performance. For instance, if the UAG is placed at the core network segment (national-wide infrastructure with hundreds of nodes and large DCs), the amount of resources to be partitioned is also considerable. Consequently, this may derive on the resource granularity information to be handled by the CP instance.

4.2.2 Network Slicing in the context of the centralised network scenario

We assume in this section that a number of MNOs owning their RANs are connected to a common physical aggregation network infrastructure. Such physical aggregation

network is partitioned to compose individual vMNO backhaul tenants on top of it, adjusting dynamically their respective backhaul and EPC requirements to the actual traffic demands.

The MNO's EPC functions are as well virtualised in to the cloud (DC) connected to the aggregation network as shown in Figure 36.

In this example we have considered an approach aligned with the centralised COMBO architecture where the UAG is located in the Core CO and a centralised DC reachable via the UAG hosts VNFs such as the vEPC. For the sake of clarification, VNFs instantiated in the DC are functions and capabilities within the context of the UAG. In this particular case, the vEPC functions such as the MME and mobile gateways (i.e., SGW and PGW) are virtualised in a remote DC but still belong to the UAG entity. In this case all the functionalities of the UAG are not necessary enclosed in single box.

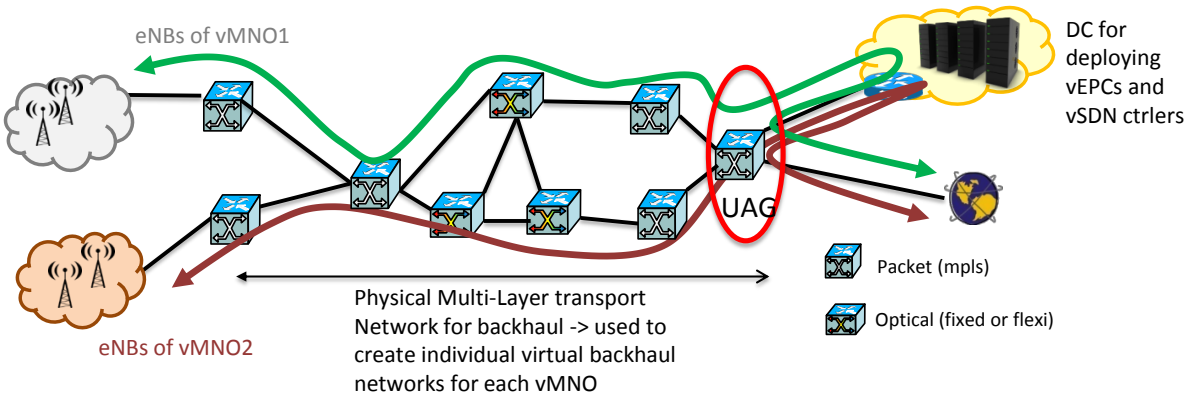


Figure 36: Deployment of vMNO backhaul for two different MNOs

4.2.3 Deployment of SDN-controlled vMNO over a physical multi-layer aggregation network

In the example below we consider a multi-layer aggregation network that combines both packet and optical switching technologies. The DC domain is located within the core network, and as commented above, is reachable via a packet network where the UAG element is located.

A MNO creating or increasing its backhaul capacity is built upon the aggregation network as interconnected virtual packet domain (see Figure 37). For the FMC purposes, observe that similarly a (virtual) network operator owning the access part (e.g., PON ONUs, DSLAMs) could request network resources over the common aggregation physical infrastructure towards (v)OLTs or vBNGs being deployed in the UAG's DC.

The MNO SDN controller's vision is an abstraction of a set of connected packet domains (via an optical connection) providing the connectivity between the RAN and vEPC at the DC. A virtual packet domain represents each abstracted packet domain whose interfaces are mapped to the physical incoming/outgoing links of a packet flow. In the example, for MNO1 the virtual packet node of the domain linked to the RAN is formed by ingress A and egress C of the corresponding physical packet network.

The network topology and packet resource status is kept in the topology database of a dedicated SDN controller per MNO backhaul tenant. This is used to dynamically set up packet MPLS tunnels for backhauling upcoming mobile LTE signalling and data bearers (i.e., S1-MME and S1-U) between the RAN and the vEPC. The vSDN controller for the vMNO backhaul is provided as a VNF in the core DC [28]. Last but not least, the connectivity within the DC is virtualised connecting the core packet domain (containing the UAG) and the deployed cloud VNFs.

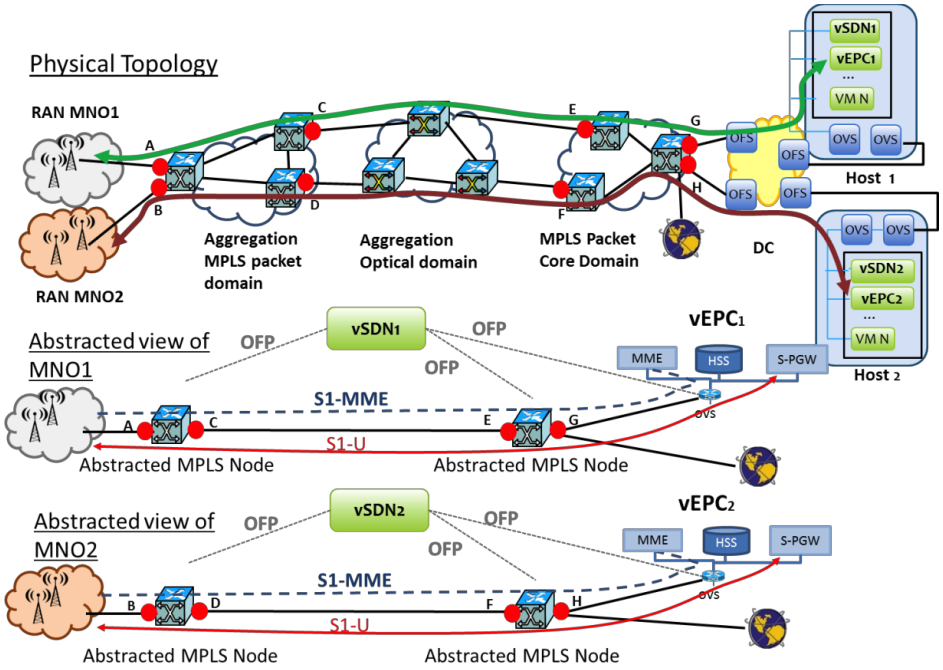


Figure 37: Physical multi-layer aggregation network connecting RANs and DCs and abstracted view of the backhaul network per MNO

4.2.4 SDN/NFV orchestration of vMNO backhaul

The SDN/NFV orchestrator architecture used to deploy vMNO backhaul (dynamically and automatically) is depicted in Figure 38. The NFV orchestrator deploys the VNFs on top of a common cloud and network platform (NFV Infrastructure, NFVI). Again, it is assumed that a single Physical Infrastructure Provider owns the NFVI. Such NFVI, defined by ETSI NFV ISG, is formed by the physical aggregation network, their heterogeneous SDN controllers per domain, and the virtual (compute) resources available in the DC.

Whenever a new vSDN controller and vEPC ar deployed, VNF managers are created to handle the VNF's lifecycle.

The Multi-Domain SDN Orchestrator (MSO) is a unified transport network operating system handling the composition of end-to-end provisioning services across multiple domains of the aggregation network at an abstract level. Another relevant element is the Multi-domain Network Hypervisor (MNH) owned by the Physical Infrastructure Provider. MNH partitions and aggregates the physical resources (i.e., nodes, links and optical spectrum, etc.) in each domain into virtual resources and interconnects them to compose vMNO backhaul tenants. Additionally, MNH is responsible for the

abstracted packet backhaul network configured by a vSDN controller. The MNH creates, modifies and deletes vMNO backhaul in response to MNO requirements.

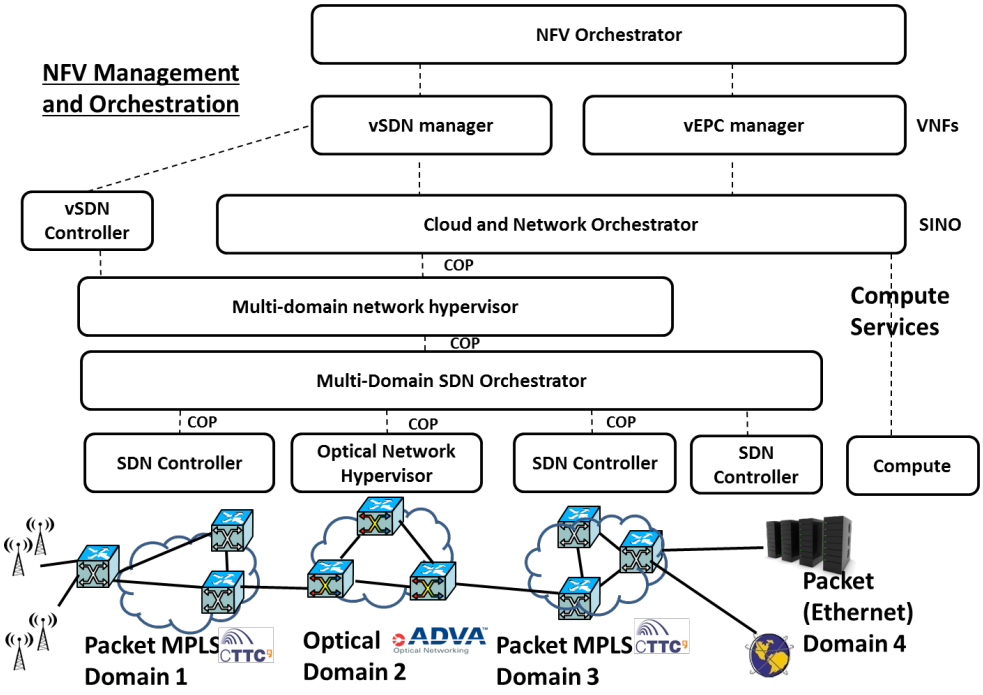


Figure 38: SDN/NFV orchestration architecture providing vMNO backhauls

The Cloud and Network Orchestrator handles the coordination and management of cloud resources (VMs) and network resources in the aggregation network infrastructure. Hence, it provides a common ecosystem for a cloud and network operating system towards deploying the vMNO backhaul and vEPC function. Finally, the NFV orchestrator manages the physical and IT resources where different slices of them can be created to support a number of applications and services. In this particular work, MNOs request dynamically the creation and/or enhancement of vMNO backhaul infrastructures, which leads also to instantiate VNFs such as vEPC and vSDN.

4.2.5 Workflow for creating the vMNO backhaul

Figure 39 shows the workflow between the involved functional blocks of the SDN/NFV orchestrator to manage the creation of an SDN-controlled vMNO backhaul and the corresponding vEPC. Step 1 allows the NFV orchestrator to request the provisioning of the vSDN controller (for the virtual backhaul) and the vEPC. This is handled by the corresponding VNF managers sending requests to the Compute controller of VMs with the respective implementation of the VNFs (vSDN and vEPC). The response determines the IP / MAC addressing of each involved element (i.e., vSDN and vEPC including MME, SGW/PGW, etc.). Next, in step 2, the creation of the vMNO backhaul is conducted. This process entails building the virtual backhaul and allowing the connectivity of the created vSDN controller to configure such an infrastructure. To do that, the MNH receives the request and computes the domain sequence within the aggregation network to connect at the packet level the MNO RAN and the vEPC. This requires that at first the traversed packet domains be

interconnected via an optical connection that is triggered by the MSO. When the optical connection is set up at the packet level all the domains are interconnected. For those packet domains the MSO subsequently requests the packet flow provisioning specifying ingress/egress links of those domains to derive the abstracted (virtual) packet node forming the targeted virtual backhaul. It is worth mentioning that this process is performed twice to support bidirectional packet communication within the backhaul. Finally, a L2 flow in the DC infrastructure (e.g., Ethernet) is created to connect the virtual (MPLS) node with the vEPC. Once the virtual backhaul connectivity is ready, this is notified to the NFV orchestrator, and at that time, the vSDN has a view of the virtual packet backhaul used to transport LTE bearers between the RAN and the vEPC.

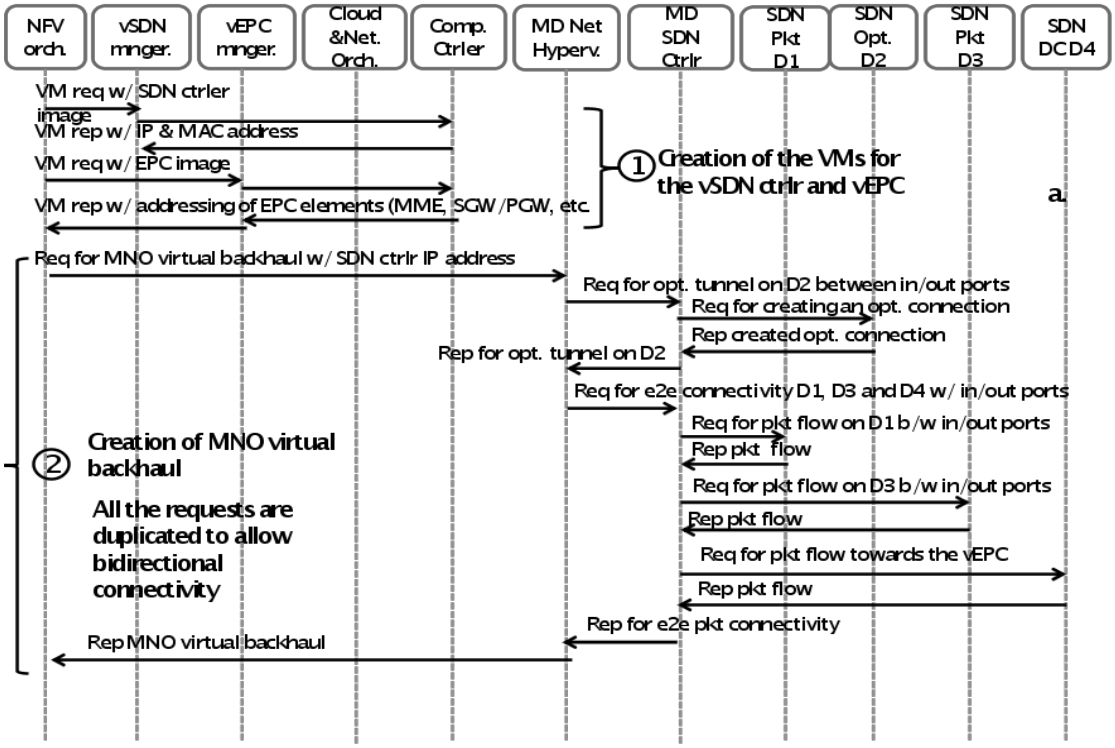


Figure 39: Workflow for provisioning vMNO backhaul network and VNFs

4.3 Roaming and offloading implementation

In FMC networks, a possible network-sharing scenario consists in sharing the physical network infrastructures required for supporting both fixed and mobile communications, leveraging roaming and offloading agreements. In such scenarios, competing network operators might adopt cooperative strategies to achieve a number of advantages brought by network sharing such as: i) OPEX and CAPEX savings, ii) energy savings iii) dynamic adjustment of resources shared between operators and services, iv) fair sharing of the capacity. These benefits will also improve the customer satisfaction due to higher capacity availability at lower costs.

In the remainder of the section we consider *i)* the optimisation of a multi-operator interface selection that minimises the total energy requirements of the MNOs while maintaining QoS and *ii)* the application of game theory for a fair offloading implementation in a multi-operator scenario.

Before we dig into the specific proposed solutions, we remark that, to effectively perform roaming and offloading, operators might leverage different kind of handover operations, namely: i) horizontal handover, ii) vertical handover, and iii) multi-operator handover. Horizontal handover occurs when a user changes its connection point within the same technology, e.g., a user previously connected to a LTE eNB connects to a different LTE eNB. Vertical handover happens when a user roams from one technology to another, e.g., a user connected to a LTE eNB connects to a Wi-Fi Access Point (AP). Multi-operator handover is performed when a user is offloaded from the network of one operator to the network of another operator. Also combinations of these three types of handover may arise. For example, a Vertical handover can occur between the networks of two different operators, e.g., a user connected to a Wi-Fi AP of operator A connects to a LTE eNB of operator B.

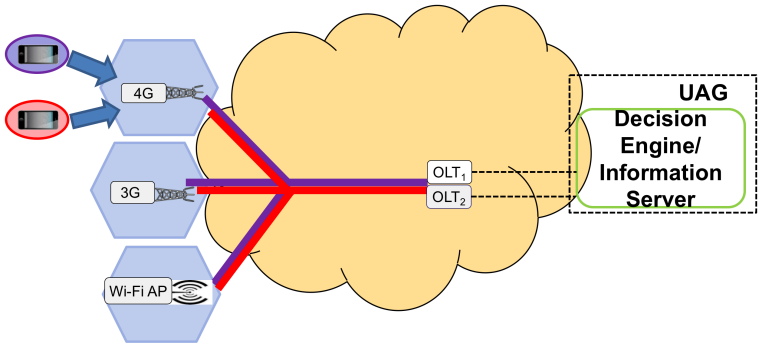


Figure 40: Multi-operator/ multi-technologies network sharing scenario

In order to effectively perform multi-operator handover (i.e., network sharing), a decision engine must take care of how the capacity of the networks of different operators can be shared. Figure 40 shows a scenario where two operators share the FMC infrastructure in the metro/access segment and different technologies (i.e., 4G antennas, and Wi-Fi APs) are backhauled over this infrastructure. The decision engine can be located in different metro nodes (e.g, Main CO or Core CO, please see Table 12 for a qualitative comparison between various implementation cases). The two operators have to send the information needed to decide if the collaboration would occur or not. The information that has to be sent to the decision engine depends on the collaboration strategy used. In the example of Figure 40, when the collaboration occurs the users of one operator are served by the bandwidth of the second operator. We show only two users: one (in blue) customer of operator 1 and one (in red) customer of operator 2. In all the approaches described in the reminder of this subsection, we assume that roaming and offloading decisions are taken in a decision engine as described above.

4.3.1 Multi-Operator Access Optimisation for Energy Efficiency, QoS and Resilience

The aforementioned three dimensions of access selection and change, namely the horizontal, the vertical and the multi-operator, all help MNOs to achieve any possible trade-off in energy efficiency, quality and availability.

The common feature of all the three dimensions is that their usage increases the redundancy of the infrastructures, and therefore, if there is an overload or a failure anywhere, there are three dimensions available for offload or protection.

On the other hand, if there is no failure, and the amount of the traffic is low, e.g., in early dawn, energy can be saved. The simplest way is to switch off parts of the network that are not used, and to switch these parts on when they are needed. This is referred to as selective switch-off. Using optimisation, energy saving can be further improved. When there are few users per cell, cells are underutilised; however, they all use energy for their operation. To switch them off, the users are redirected to other neighbouring cells. That way the load of certain cells is increased, however, there will be cells with no traffic at all, and those can be switched off. This concentration of users to certain cells is referred to as consolidation. There are constraints for consolidation. If the user is too distant from a cell it cannot connect to that cell. Or in the second dimension, if a cell is using different network technology the user cannot connect. Or in the third dimension, when a MNO has no roaming contract with another MNO covering the same area, the user cannot move to it.

On the one hand we intend to enhance the quality and the availability, while on the other hand we aim to decrease power requirement of the network. Indeed, there is a trade-off between these two objectives. It is because a better QoS and a higher availability require more parallel resources, and therefore a higher energy requirement. Conversely, when equipment are switched off or put to sleep, or in stand-by, energy usage will drop, but QoS and availability may potentially deteriorate.

Our optimisation framework, developed within the COMBO project, minimises the total energy requirements of the MNOs, of the networking technologies, and geographically, while maintaining both the QoS and the availability at a required level for all the users all the time in a fair way.

In this section we focus onto the third dimension, however we assume using the other two dimensions as well.

In [74] and [75] we propose four different methods for assigning users (User Equipment: UE) to various accesses. These methods are of different complexity, the most complex one having the most features. We have also shown [74] the impact of the topology of the fronthaul and of the backhaul access parts of the networks. By proper design of the fronthaul/backhaul topology significant QoS and availability improvement can be achieved. The topology requirement is, that all UE must see at least two such access interfaces that do not belong to the same Shared Risk Group (SRG), i.e., physically are disjoint to such extent that any single failure cannot affect both.

In [76] we focused onto the multi-operator (multi-MNO) case, where we assume that the whole area where the users move is covered by two, three or four MNOs. At the moment national roaming (moving from home MNO to competitor MNO) is not supported in majority of countries, or if supported, then with strict rules and limitations only. However, if the user goes abroad, to another country, it will be able to use any of the networks the home MNO has roaming contract with. Google has solved this national roaming issue by setting up a MVNO over two MNOs. Tax authority of Hungary has solved the same issue by using foreign MNOs SIM cards to enable

national roaming [80]. In any case, throughout our optimisations we assume that national roaming is enabled.

We also proposed two methods for handling multi-operator interface selection and change. First, when the user has higher cost when using a foreign network instead its home network. Second, when the user can use both his home and a foreign network under same conditions, however, the foreign network has limited the ratio of resources that can be used by users that are not his own users. These methods can be combined as well.

In all cases, when it was allowed for users to use not their own home networks only, but other networks as well, the energy requirement dropped and the per user throughput increased. The availability was in all cases the worse, when no network sharing between MNOs was allowed at all. However, allowing increasingly more sharing did lead to availability improvement, since intensive sharing allows serving all the users by lower number of elements switched-on, that, on the other hand, leads to availability deterioration.

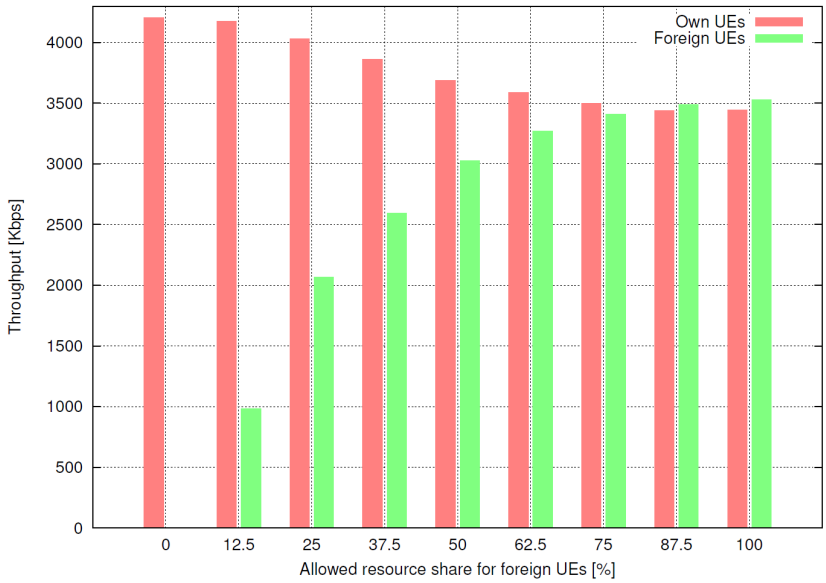


Figure 41 Throughput versus allowed resource share ratio for two MNOs in case of failure.

As an example we show in Figure 41 how the throughput of an MNO drops as the throughput of a foreign MNO grows as the ratio of resource sharing grows in case of failure. It is important to notice that the own traffic negligible dropped, while the foreign traffic was almost fully carried. Higher network sharing ratios allowed higher total throughput.

The ideas and methods described above are reported in [78]. The demonstration of the same ideas was carried out within WP6 and is described in Section 2.5.1 (Network Controlled Offload and Smooth Handover) of [6], while the results of the tests are reported in Section 4.4 (uDPM Functionality) of [6].

4.3.2 Multi-operator Game Theoretic Model for Effective Traffic Offloading

4.3.2.1 Introduction to the offloading problem: "Coopetition"

The concept of creating value through cooperation in a competing field is called "Coopetition" [29]. We propose and evaluate a multi-operator network-sharing approach based on non-cooperative game theory (GT) that fosters coopetition among competing network operators. The proposed GT-based coopetition scheme allows operators to increase the traffic served during peak-hours. At the same time, our approach avoids that one operator fills all the unused bandwidth of the second operator causing a consistent traffic loss, which can lower the QoS experienced by end-users of the second operator, and potentially cause revenue losses. GT is a suitable tool to solve the problem of sharing the network capacity among different network operators since, in this manner, operators can decide whether collaborate or not according to traffic conditions (i.e., dynamically) and not following predefined agreements.

To decide how much bandwidth is it advantageous to share with the competitors in to maintain a fair bandwidth utilisation for its own users, we use our proposed game-theoretic model for network sharing. We refer to the operator that needs to offload a part of its traffic as the *Offloader* while the operator that should serve the exceeding amount of traffic is called the *Receiver*. The Receiver can decide whether accept or not this amount of traffic using the game theoretic approach proposed in our work. If the Receiver decides to not accept, then the traffic of the Offloader cannot be served and therefore is lost. Collaboration between the two operators occurs only when one operator needs to serve more traffic with respect to the capacity of its own network and the second operators has some available capacity, and vice versa. Instead, when both operators have to serve more traffic than their capacity or when both operators are in a light loaded traffic condition, then no collaboration occurs.

4.3.2.2 Game-theoretical model for traffic offloading

Game theory is the study of strategic decision-making. It is the study of mathematical models of conflict and cooperation between intelligent rational decision-makers. The games studied in game theory are well-defined mathematical objects. To be fully defined, a game must specify: i) the players of the game, ii) the information and actions available to each player at each decision point, and iii) the payoffs for each outcome. These elements are used together with a solution concept to deduce a set of equilibrium strategies for each player, i.e., Nash Equilibrium. When the equilibrium strategies are employed no player can profit by unilaterally deviating from their strategy. These equilibrium strategies determine an equilibrium to the game, i.e., a stable state in which one outcome or a set of outcomes occur with known probability. The normal form game, also called strategic form, is usually represented by a matrix that shows the players, the strategies, and the payoffs.

A game can be cooperative or non-cooperative and can also be zero-sum or non-zero-sum (please refer to Annex 2 for more details). The game that we model in this work is a *no- cooperative, and non-zero-sum game*. In the proposed game each player chooses its strategy independently to improve its own performance (i.e., utility) or reducing its losses (i.e., costs). The game is modelled as a coopetition strategy in

which the operators cooperate for a common benefit (serving a higher amount of users minimizing the wasted bandwidth) while at the same time they compete (each operator has as a priority to satisfy its own customers). The players of this game are the network operators. The only information shared between players is the amount of traffic exceeding the capacity of the network of an operator, i.e., the amount of bandwidth that an operator wants to offload. In this game each player has two strategies that depend on the role of player: the Offloader can decide whether offload or not, while the Receiver can decide to accept or not to accept the traffic of the Offloader, and over which technology to accept the offloaded traffic. This game is played every time an operator has some traffic exceeding the capacity of its network. Therefore, the role of the player can change according to the load condition of each network. The solution of the proposed game is represented by the Nash Equilibrium (see Annex 2). In [79] we report the game theoretic model for coopetition between two operators in a multi-technologies scenario. The proposed game is a tool that can be used in order to choose when and where (i.e., over which network types, Wi-Fi or LTE) to offload exceeding traffic. The choice of the strategy of each operator depends on the values ("utilities") computed for each player and each strategy in the table.

Table 11: Proposed game theoretic strategy for multi-operator/multi-technologies network sharing

		Operator 1	
		Offload	Not Offload
Operator 2	Not Accept	$((Ex_1 - \overline{Ex_{wi,2}}) + (Ex_1 - \overline{Ex_{lte,2}}),$ $(L_{wi,1} + L_{lte,1}) - (C_{wi,1} + C_{lte,1}))$	$((Ex_1 - \overline{Ex_{wi,2}}) + (Ex_1 - \overline{Ex_{lte,2}}),$ $(C_{wi,1} + C_{lte,1}) - (L_{wi,1} + L_{lte,1}) + 1)$
	Accept on LTE	$(\overline{Ex_{lte,2}} - Ex_1 + 1, (L_{wi,1} + L_{lte,1}) -$ $(C_{wi,1} + C_{lte,1}))$	$(\overline{Ex_{lte,2}} - Ex_1 + 1, (C_{wi,1} + C_{lte,1}) -$ $(L_{wi,1} + L_{lte,1}) + 1)$
	Accept on Wi-Fi	$(\frac{\overline{Ex_{wi,2}} - Ex_1 + 1}{\overline{Ex_{lte,2}}}, (C_{wi,1} + C_{lte,1})$ $- (L_{wi,1} + L_{lte,1}))$	$(\frac{\overline{Ex_{wi,2}} - Ex_1 + 1}{\overline{Ex_{lte,2}}}, (C_{wi,1} + C_{lte,1})$ $- (L_{wi,1} + L_{lte,1}) + 1)$

Let us discuss the variables used to compute the utility of the strategies: Ex_n is the exceeding traffic of player *n* (i.e., the amount of traffic that cannot be served by player n); $\overline{Ex_{wi,m}}$ and $\overline{Ex_{lte,m}}$ are the maximum amount of exceeding traffic of player *n* that can be served by player m respectively over Wi-Fi network and over LTE network. $L_{wi,m}$ and $L_{lte,m}$ correspond to the load (amount of traffic) of Wi-Fi and LTE networks, respectively. $C_{wi,m}$ and $C_{lte,m}$ are the capacity of the Wi-Fi and LTE networks. According to the instantaneous values of such variables each player will choose a strategy. Particularly, in Table 11, player 1 (i.e., Operator 1) will choose to

offload when he has some exceeding traffic (the value of Ex_n is greater than zero). Operator 2, will not accept to serve the traffic of Operator 1 if any capacity is available on both Wi-Fi and LTE networks. If there is enough capacity on both networks, Operator 2 will choose to serve the traffic of the competitor over its LTE network, while it will serve this traffic over the Wi-Fi network only when there is not enough capacity over the LTE network. The game described in Table 11 prioritizes the utilisation of the LTE network when serving the traffic of the competitor, however, the utilities of the game can be computed in a similar manner in order to prioritize the utilisation of the Wi-Fi network, based on operators' needs.

In [79] it is shown that the GT collaboration model is able to improve the amount of traffic served by the operators with respect to the case where operators do not collaborate. The GT collaboration also has the lowest amount of traffic lost due to the collaboration (i.e., traffic of an operator which cannot be served because the capacity of the operator is filled by the traffic of other operators) with respect to all the other collaboration methods evaluated in [79]. At the same time, with GT collaboration the amount of traffic lost by the operator serving the excess traffic of another operator is only slightly higher than the case where no collaboration occurs. This means that the highest portion of the traffic lost by the receiver is lost due to the condition of the offered traffic, and not to the collaboration. The results clearly show that it is better for operators to collaborate when they have two opposite traffic profiles.

More details on the proposed methodology and a larger set of numerical results can be found in the Annex.

4.3.3 Call Flow for Multi-Operator Network Sharing

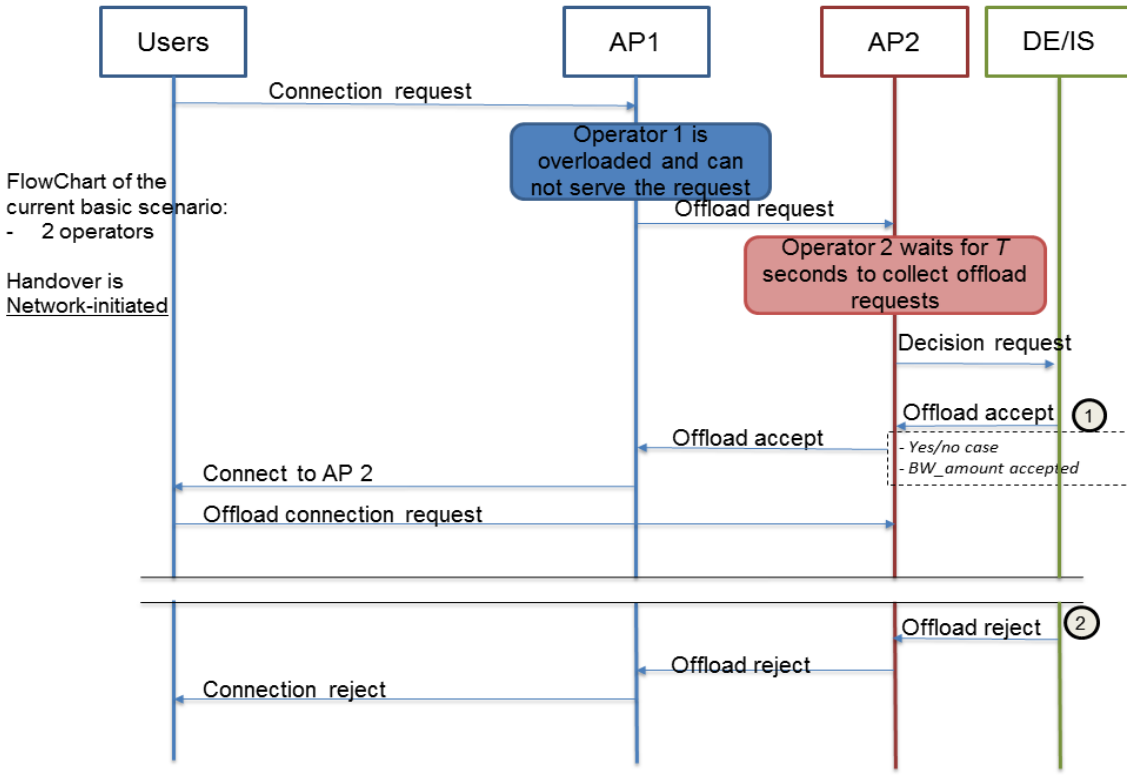


Figure 42: Call Flow for the Multi-Operator Network Sharing

In Figure 42, we present the call flow used in the Multi-Operator Network Sharing scenario to enable the communication and collaboration among different operators. The network elements involved are the Users, the Access Points of the operators (AP1 and AP2), and the DE (DE/IS) in which the decision about the collaboration are taken. We assume here and in Figure 42 that there is a single DE, whereas each operator is likely to operate its own DE. However, from a decision point of view, it is equivalent to have a single DE and to have two DEs that collaborate by sharing information, in order to implement the resource sharing policy.

The APs of the operator could be co-located. Also, the APs can be both Wi-Fi APs or LTE eNB. In particular, the decision procedure is triggered by Users' requests that arrive to one operator: when the network of an operator is overloaded, that operator starts a communication with the competitor that operates in the same area to ask some capacity to serve the exceeding traffic. This *Offload request* indicates the amount of traffic that needs to be offloaded. The operator that receives the *Offload request* forwards a *Decision request* to the decision engine including both the amount of traffic that has to be offloaded and the amount of capacity available on its network. In the decision engine the game is played and the decision for the 2 operators is taken. If the decision is to accept to serve the exceeding traffic (Case 1) then the Offload procedure is triggered (*Offload accept*, *Offload connection request*). While if the decision is not accepted the exceeding traffic (Case 2), a *Connection reject* message is sent to the end-user that will not be served in that moment. In this case the connection is lost. This high level view of the call flow needed to be used in order to enable Multi-Operator Network Sharing can for example be mapped into the Media Independent Handover (MIH) protocol. Figure 43 shows how the call flow presented in Figure 42 can be mapped over the MIH protocol: the actors DE, AP1, AP2 from Figure 42 are respectively mapped into the Information server, 3G-PoS and WiMAX-PoS in Figure 43.

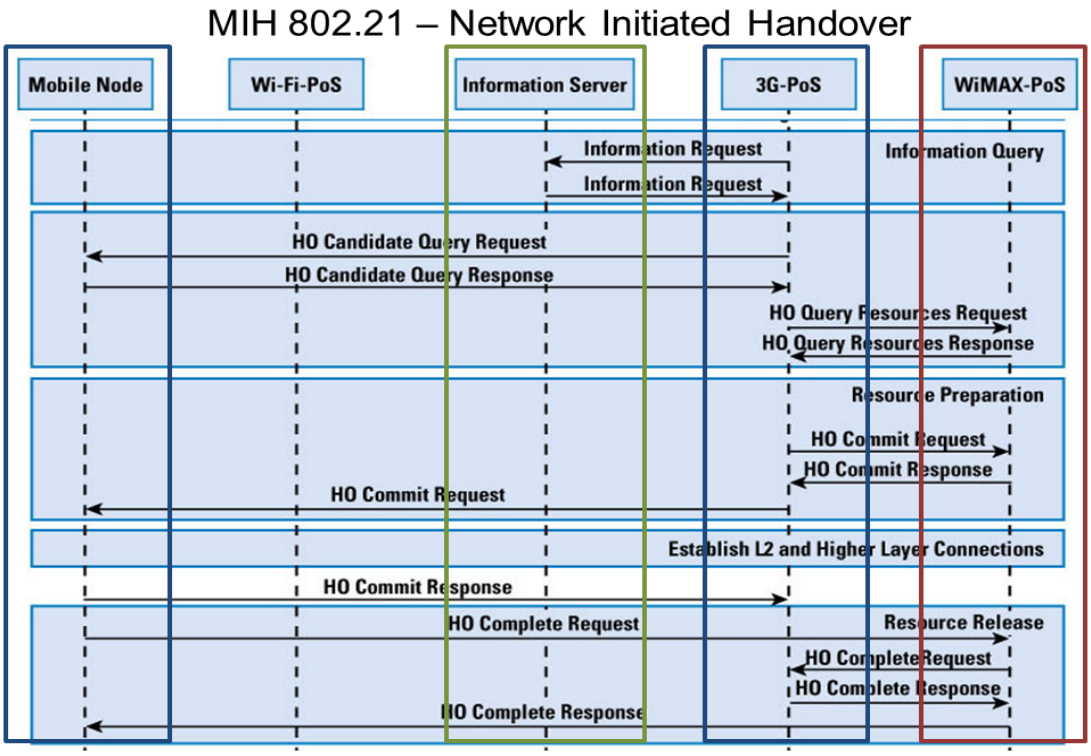


Figure 43: Media Independent Handover protocol for Multi-Operator Network Sharing

4.4 Operating SLA based network sharing

In the following scenarios, a single FMCO owns and operates a network while client network or service operators are allowed by contractual agreements (SLAs) to share it. The FMCO operates both mobile and fixed access networks, additionally with a converged aggregation network; it also relies on UAGs to control fixed and mobile networks and to ensure that resource sharing is performed according to the contractual agreements with the client network or service operators. The client network or service providers do not operate any network resources of their own and fully rely on the SLAs negotiated with the FMCO to provide the network resources necessary for fulfilling the services provided to their subscribers. This network sharing mode has been introduced in section 4.1.5.

The UAG is configured with the characteristics of the SLA such as number of subscribers, total contracted bandwidth and QoS levels. These configurations could be static or dynamic. In case of static configuration, resources are pre-reserved leading to quick reaction but possibly to low resource utilisation. In case of dynamic configuration, resources are not pre-reserved which may lead to slower reaction to requests but higher resource utilisation.

A monitoring function controlled by the FMCO records available resources per operator and per service class; this information is provided to the DE within the UAG that shall check monitored resources against the requirements agreed upon in SLAs.

In the following, several typical examples of how such a SLA-controlled network sharing scenario is operated, are described.

4.4.1 UE admission control in a SLA-based network sharing scenario

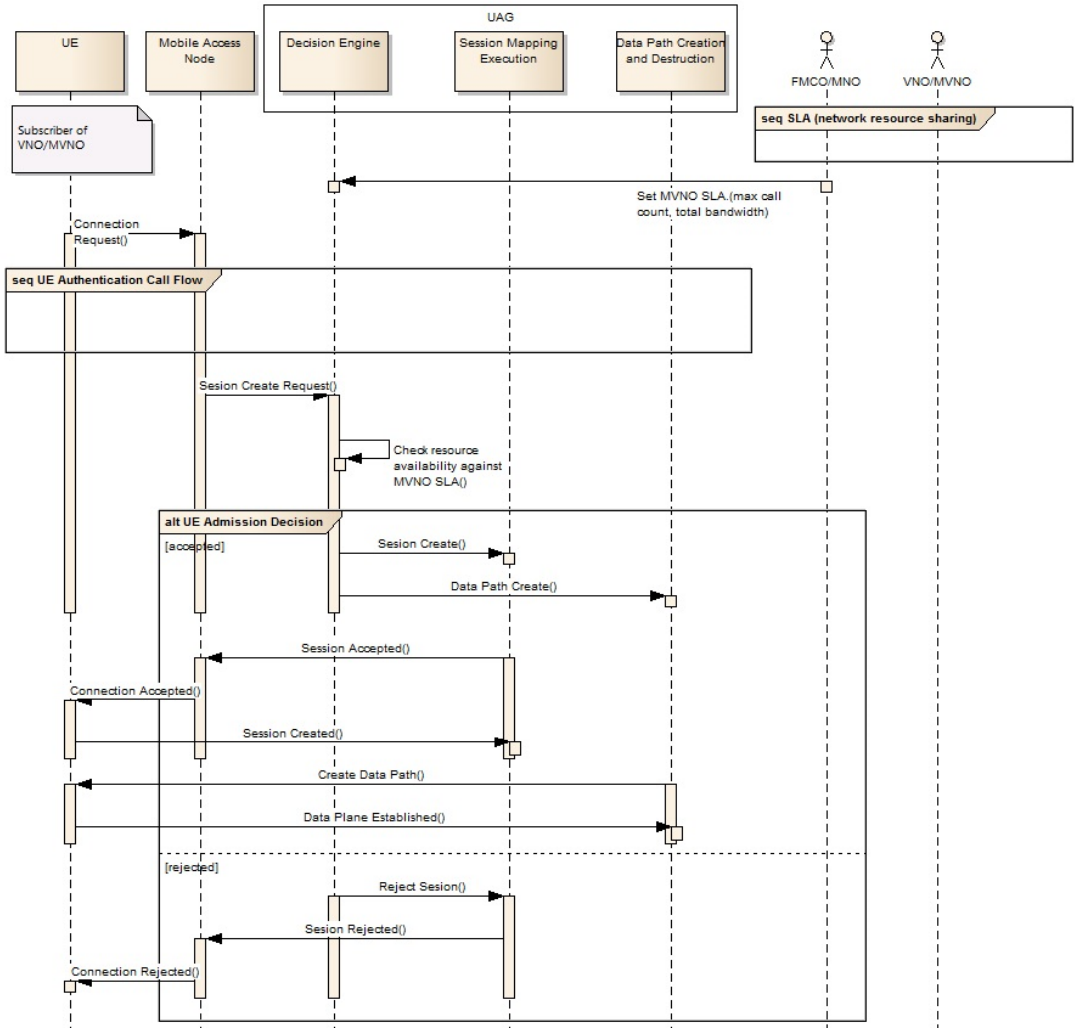


Figure 44: SLA based UE admission control flow

When the subscriber of a client operator attempts to initiate a connection to the FMCO's network, its UE tries to create a session and the DE has to accept or reject this call attempt, on the basis of the negotiated SLA. The flowchart corresponding to this acceptance control is shown in Figure 44.

4.4.2 Offloading control in a SLA-based network sharing scenario

As specified by the SLA, the client operator may have requested the FMCO to rent him both fixed and mobile access resources, and additionally to perform offload from the mobile access network to the fixed access network whenever possible.

The flowchart corresponding to offload control is shown in Figure 45. The client network's subscriber is already connected to mobile access network and its UE enters an area (e.g. home, hotel or a cafe) where access to the fixed network is possible. A monitoring process detects that offload is possible, and informs the DE. The DE has to accept or refuse the offload, based on the negotiated SLA, the global state of network access links and the current usage of resources by the client operator. If offload is authorised, uDPM functions are activated to realise it.

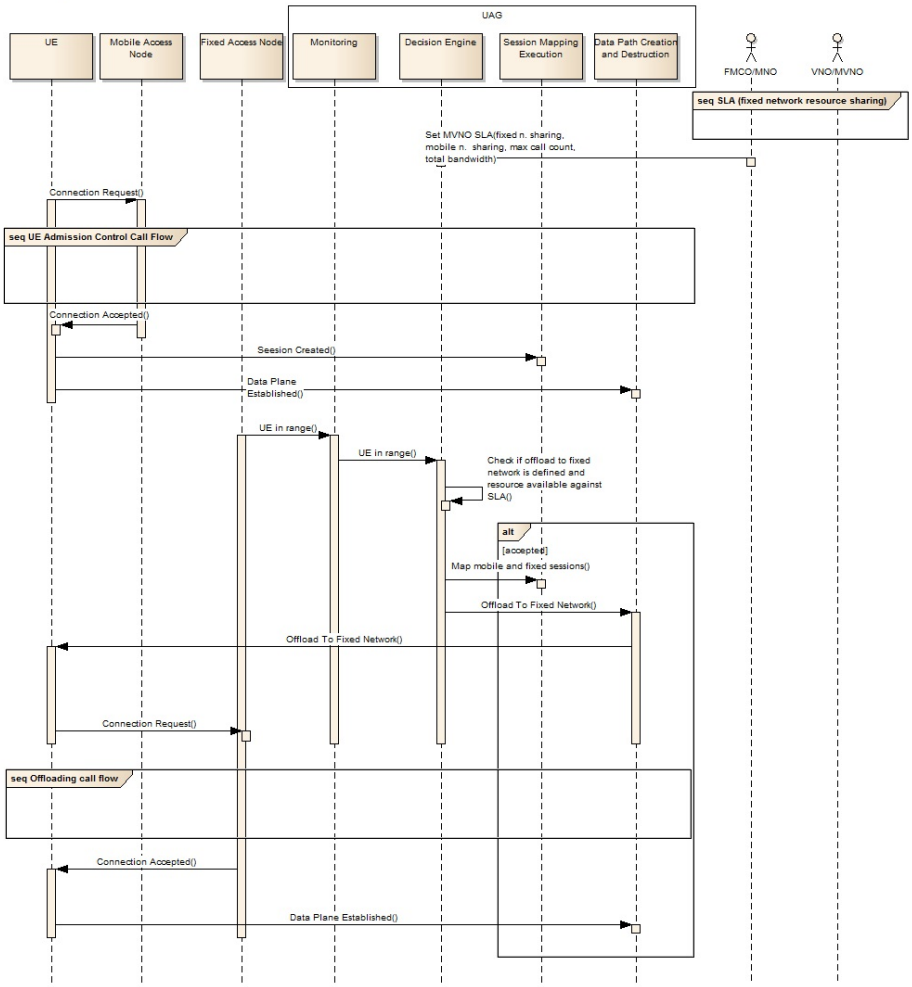


Figure 45: SLA based offloading decision flow

4.4.3 Multi-path request handling in a SLA-based network sharing scenario

As specified by the SLA, the client operator may have requested the FMCO to rent him both fixed and mobile access resources, and additionally to perform multi-path data forwarding whenever possible.

The flowchart corresponding to offload control is shown in Figure 46. The client network's subscriber is already connected to mobile access network and its UE enters an area (e.g. home, hotel or a cafe) where access to the fixed network is possible; the UE then attempts to activate another interface to take advantage of the fixed access network. A monitoring process detects that multi-path forwarding is possible, and informs the DE. The DE has to accept or refuse the request, based on the negotiated SLA, the global state of network access links and the current usage of resources by the client operator. If multi-path forwarding is authorised, uDPM functions are activated to realise it.

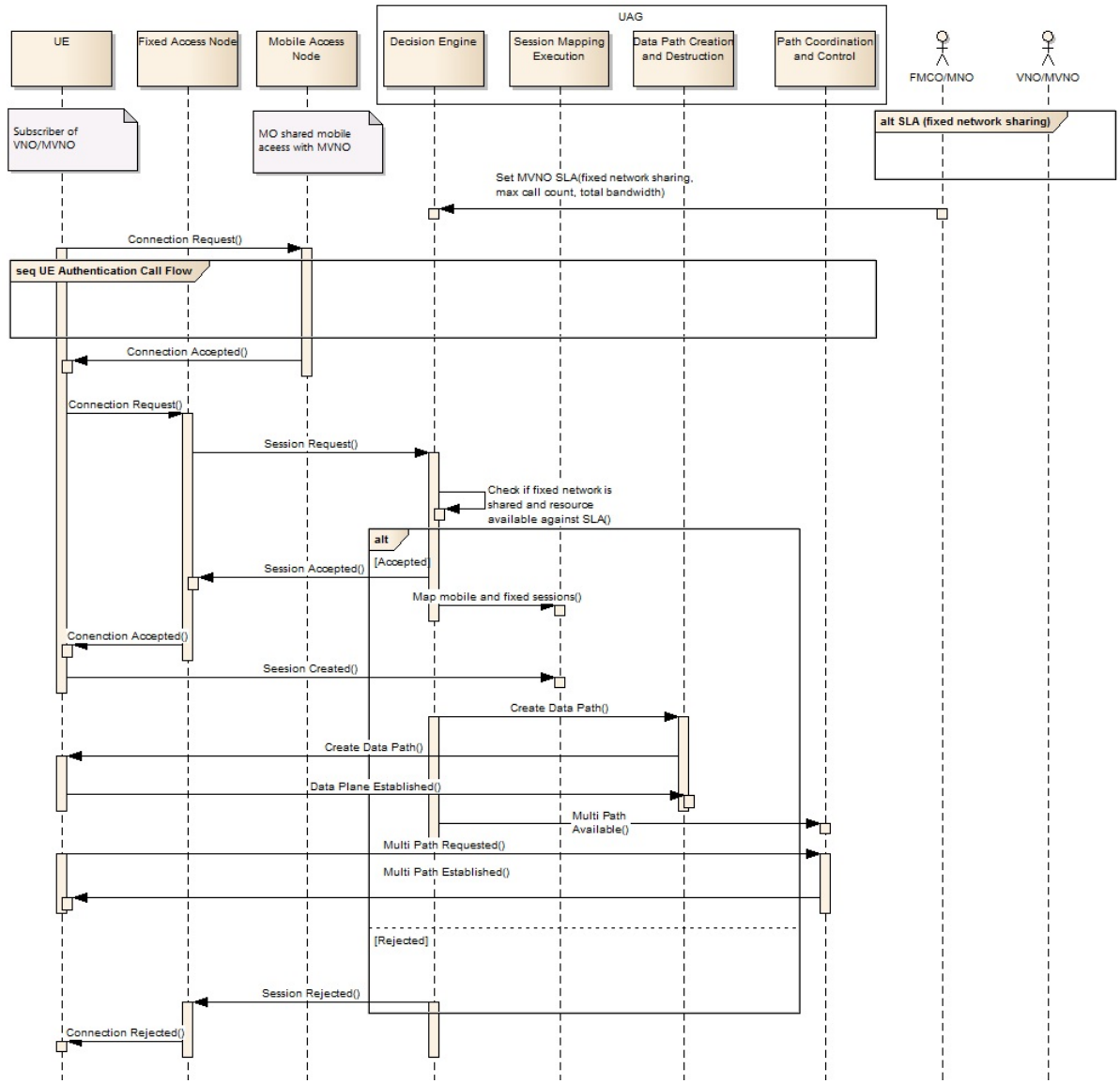


Figure 46: SLA based multi-path data send decision flow

4.4.4 Data Path Creation for an OTT service in a SLA-based network sharing scenario

In the present case, it is assumed that the FMCO has a contractual agreement with an OTT service provider.

In the scenario shown in Figure 47, a client network's subscriber is already connected to the FMCO network. An OTT application running on the UE is started and in order to connect to OTT Application Server it requests Data Path Creation from the UAG. The DE within the UAG receives the request and has to check whether the request is compatible with the SLAs contracted with the OTT service provider.

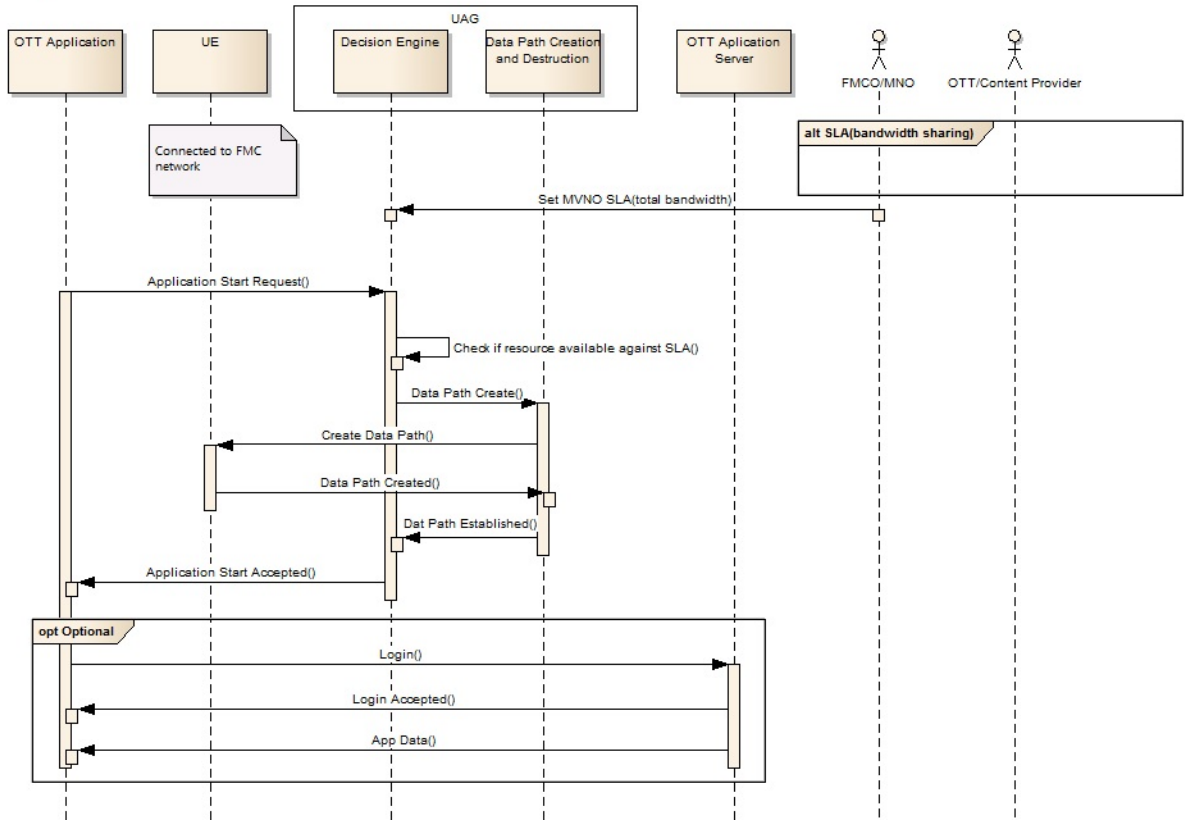
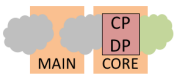
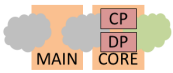
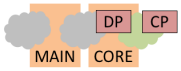
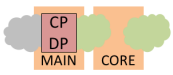
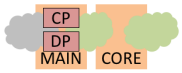
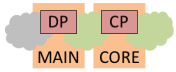


Figure 47: SLA based data path creation for OTT application flow

4.5 Qualitative assessment of implementation options on network sharing issues

Table 12 provides a qualitative assessment of the impact of each UAG implementation option on network sharing issues. Six implementation options are considered for assessing network slicing, roaming/offloading and support of agreements between OTTs and network operators. Each option of the UAG implementation is graded with a degree of fitness relative to both CP (Command) and DP (Execution).

Table 12: Qualitative comparison of UAG implementations regarding network sharing

Centralised COMBO architecture			Distributed COMBO architecture		
Standalone UAG at Core CO (0) 	Split UAG with co-located DP and CP at Core CO (0) 	Split UAG with nonadjacent DP and CP and DP at Core CO 	Standalone UAG at Main CO (0) 	Split UAG with co-located DP and CP at Main CO (0) 	Split UAG with nonadjacent DP and CP and DP at Main CO 
Control of network slicing					
+ (2)	- (3)	+ (4)	++ (2)	- (3)	++ (4)
Execution of network slicing					
+/- (2)-	+ (3)	++ (4)	+/- (2)	+ (3)	++ (4)
Control of roaming agreements (5)					
++	++	++	+	+	+
Control of roaming agreements (5)					
+	+	+	++	+	++
Control of agreements in a SLA-based network sharing scenario (1)					
IRR	IRR	IRR	IRR	IRR	IRR
Execution of agreements in a SLA-based network sharing scenario (1)					
IRR	IRR	IRR	IRR	IRR	IRR

++ = a very good fit for the architecture
+ = a good fit
+/- = some positive and negative aspects
- = a bad fit
-- = a very bad fit
IRR: irrelevant to this particular implementation

Notes:

(0) Implementing DP and CP functions in a single equipment versus co-locating them in the same location only differs in terms of scalability performance. Co-location is made possible thanks to SDN, and allows to manage the scalability of DP and CP independently from one another

(1) Scenario is composed of configuring UAG according to operator agreements which is independent of UAG implementation

(2) In the standalone UAG located at either the core or the Main CO, the adoption of network slicing is the simplest in terms of control complexity (both networking and IT resource are enclosed in single physical box with fine granularity and detailed visibility), but may not be the most appealing at the

time of creating a number of network slices on top of it (all resources are confined into a single UAG box which constrains the number of network slices to be created over the whole infrastructure). See section 4.2.1.

(3) The split UAG with co-located DP and CP presents important benefits with respect to the execution of network slicing (i.e., partitioning of resources being spread in different equipment) but severe difficulties from a control perspective. Locating resources at different locations allows dealing with different slices' service requirements (e.g., delay-sensitive, high availability, etc.). Furthermore, this distributed DP feature provides flexibility at the time of extending the resources (e.g., increasing the IT resources). On the other hand, the co-located CP instances with the DP in the UAG may complicate to keep a single and unified view of all the resources. Every CP must abstract the underlying DP resources which should be passed and collected by another higher hierarchical entity (e.g., SDN orchestrator) responsible to compute and compose the DP resources for each network slice demand. Of course, this hierarchical model despite this may scale with the size of the network, the finer granularity to handle the DP resources is lost. See section 4.2.1.

(4) The split UAG with CP remote from DP may result in the best scenario from both the DP resource allocation and control perspective of network slicing. Distributed DPs allows creating and composing network slice infrastructures with differentiated service requirements. Centralising the CP in a single element allows having a complete view of all resources, leading to handle finer resource granularity for partitioning and composing network slices. Despite these advantages, locating the CP in the core segment may present some scalability issues depending on the network size (e.g., nation-wide). See section 4.2.1.

(5) To effectively manage offload and roaming in a multi-operator scenario, we need one (or more) decision engine that takes care of that either at the Main CO or Core CO. In principle, both options (main or Core CO) are fine (+) with a slight preference for Core CO (++) as the decision engine in the Core CO would have a joint vision/control over a larger number of cells, providing chance for a more optimised control of offloading, as long as the granularity of the control does not have to be too fine. This motivates the preference for the case with UAG in Core CO. Note that from a latency point of view, the reasoning is the opposite (decision engine in the Main CO would act more rapidly), but latency difference between Core CO and Main CO is considered irrelevant. As for different flavours of CP and DP location, the reasoning is as follows: CP, for the motivation discussed above, is better if placed farther away while DP, from an execution point of view, is better closer. Finally separating CP and DP or not (1st vs 2nd column) is irrelevant.

We can derive some global conclusions from the above qualitative analysis:

- The optimal location of the CP appears to be the Core CO. In case of offload/roaming, this should provide a more global optimum than the one that would be obtained if the optimum were obtained on a smaller zone. In case of slicing, it is roughly the same argument as a global control on distributed resources is more efficient than a distributed control that would mandate a hierarchical control.
- On the other hand, there is some advantage to locate the DP of the UAG should in the Main COs for latency issues in case of roaming/offloading and for a better granularity of resource allocation in case of slicing.

5 Assessment of candidate architectures for Use Cases

This chapter performs a qualitative analysis of the functional convergence architectures regarding how they can affect the FMC use cases defined in WP2. Only the use cases affected by the functional architecture are included, namely UC1, UC2, UC4, UC6 and UC8. This chapter starts from the application of the uAUT and uDPM to the UAG functional blocks (see chapter 5 of [3]) and takes into consideration the implementation options of the UAG and the possible locations of the NG-POP. Use cases requirements specified in WP2 have been also considered to identify the implementation that could be more beneficial.

5.1 UC1: unified FMC for mobile devices

UC1 allows mobile devices to use Wi-Fi access in combination with mobile access in an FMC network with an advanced cooperation between those access technologies. This use case is supported by the core functionality of the UAG: the uAUT allows the uDPM to associate flows from Wi-Fi and mobile interfaces to a single user, whereas the uDPM performs the path diversity management according to the user profile and network status. The UE also collaborates with the uDPM to perform the local monitoring and the interface selection and utilisation. Figure 48 depicts the main UAG deployment alternatives considered of the CP and DP in the environment of UC1.

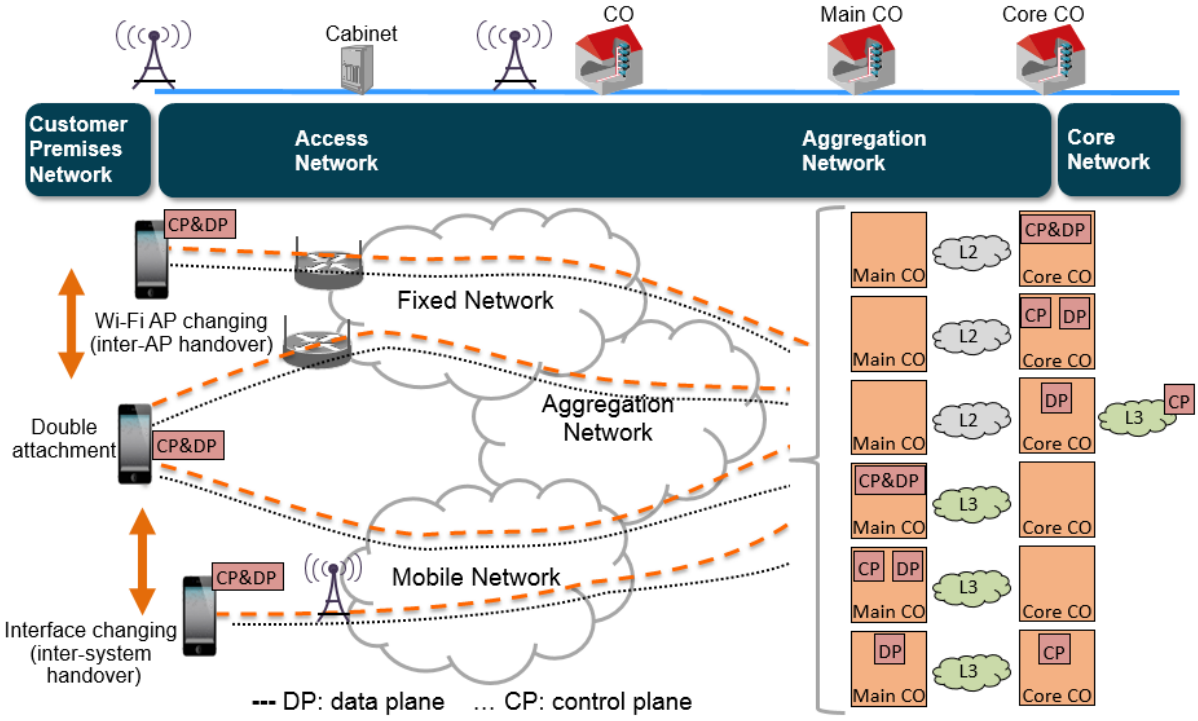


Figure 48: Implementation options for UC1

From the point of view of uAUT, a centralised approach could be preferred, as a distributed architecture is not needed for authentication nor traffic mapping procedures; however, a decentralised approach will have a similar performance with the AAA agent in the Main CO and the AAA services higher in the network.

Considering uDPM, the centralisation of the decision engine may make difficult a fast mapping of the network resources in a dynamic way and would increase control traffic in the transport network. Real time enforcement of the network management rules may require the distribution of this CP function. Path coordination and control may be compatible with both centralised and distributed approaches. Therefore, an implementation with the CP in the Main CO, or in the Core CO is recommendable for UC1, mainly depending on the granularity and dynamicity of the network resources assigned and the level of scalability desired by the network operator.

Implementing the UAG DP in the Main CO is also preferable as it allows to aggregate traffic from the available data paths on longer distances.

5.2 UC2: converged content caching for unified service delivery

UC2 provides a converged content caching solution shown in Figure 49 aiming at achieving a unified service delivery in a FMC framework. In our solution described more fully in section 3.1, we introduce a controlled content caching system including two components Cache Node (CN) and Cache Controller (CC). CN, where the storage and caching functions are enabled in customer premises network (HGW) and aggregate network (NG-POP), performs caching and prefetching functionality controlled by CC. By collaborative caching implemented in access network and aggregation network, the content can be intelligently duplicated closer to the mobile users, and efficiently delivered from the fixed network or from the mobile network. Through communicating with DE in uDPM that is aware of actual traffic load and network status, to achieve an optimal traffic offloading and content caching/prefetching decision. This use case has been reported in section 4.5 of [6].

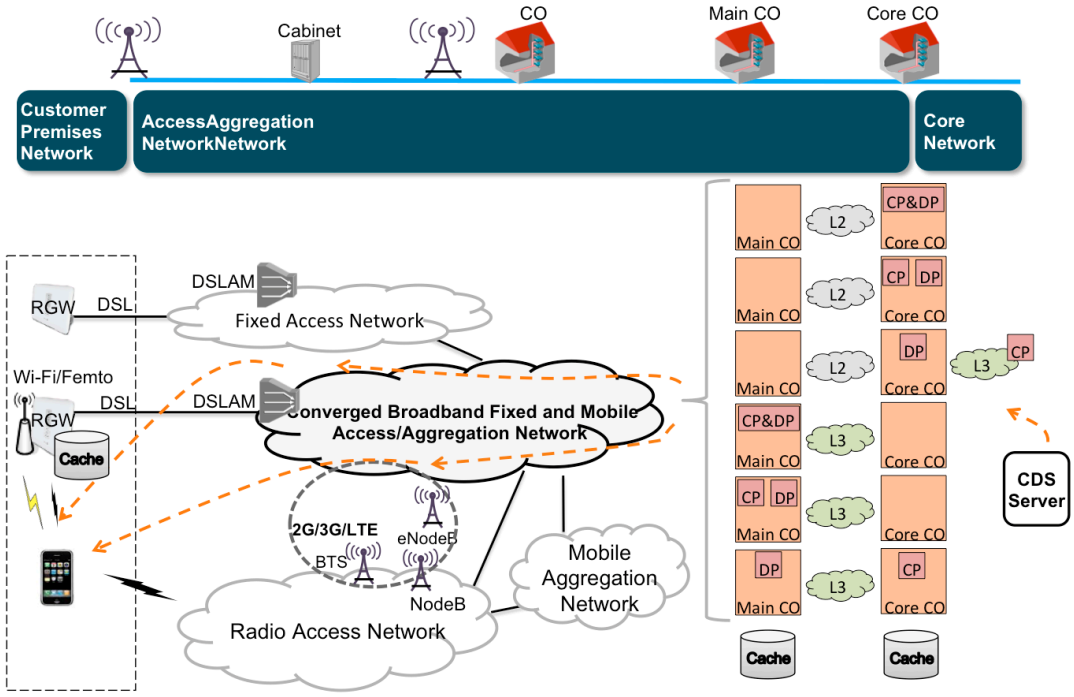


Figure 49: Implementation options for UC2

A distributed COMBO content delivery solution can improve the reliability, scalability and performance of the content distribution service (latency and throughput). Hence DP located in Main CO would be a preferred solution. However, the caching effectiveness usually gets lower if the network caches are used in the points where the user population is lower. It means if CP is deployed in Main CO, the caching benefit may not be fully achieved. The recommended solution is that CP is located in Core CO and DP is deployed in Main CO with a distributed way, and then content servers located in different parts of network collaborate among each other and these servers are managed by CC in Core CO. In this way, we can both achieve a reliable, scalable and good performance caching solution and improve the cache benefit by enabling collaboration between distributed caches.

5.3 UC4: universal access bundling for residential gateway

UC4 provides integrated functionalities into a converged network in order to provide the user with optimum bandwidth resources dynamically assigned via available fixed, mobile, and wireless technologies.

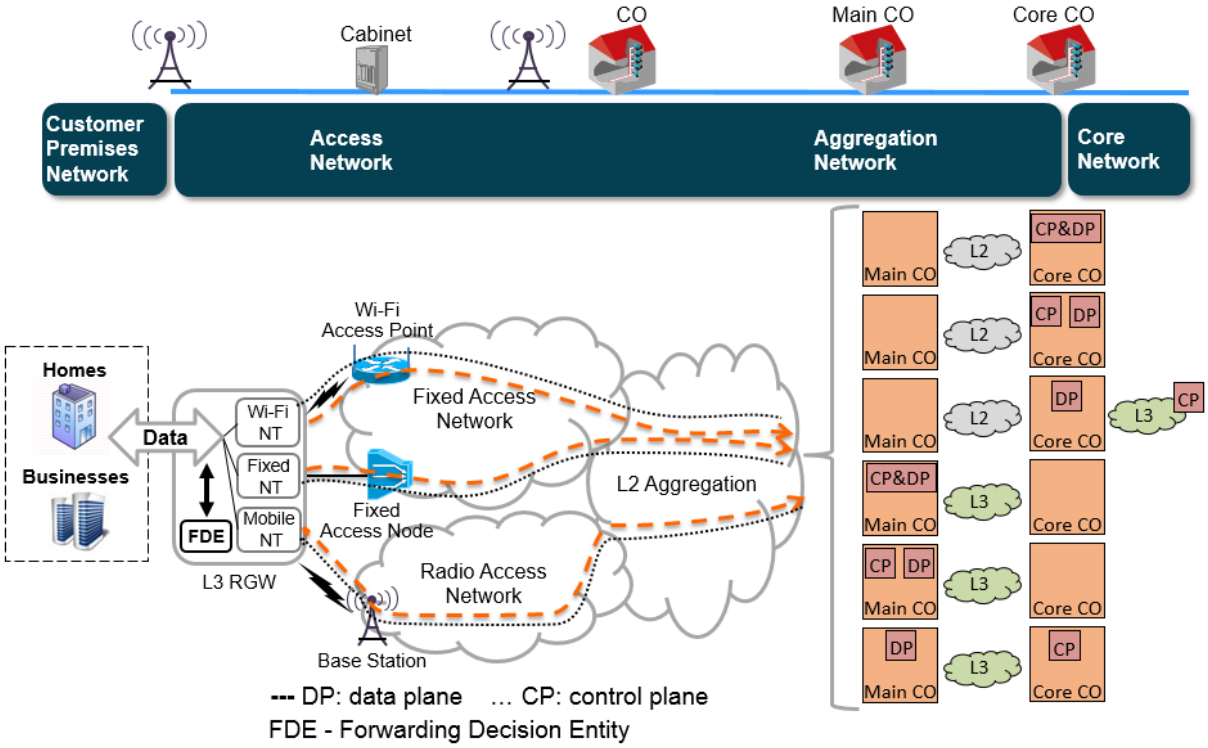


Figure 50: Implementation options for UC4

This use case allows all devices for either homes or businesses behind the RGW to use the combined available resources (Wi-Fi, mobile and fixed access) based on an advanced cooperation between the access technologies.

This use case is supported by the core functionality of the UAG: the uAUT allows to associate flows from fixed, Wi-Fi and mobile interfaces to a user behind a RGW, whereas the uDPM performs the path diversity management according to the user's profile and network status. The RGW collaborates with network elements to perform

the local monitoring and the interface selection and utilisation. Figure 50 depicts the main UAG deployment alternatives in the environment of UC4.

In general, we assume that the uDPM scheduling algorithms need to be configured for downstream and upstream traffic.

We now analyse qualitatively the impact of the UAG implementations on the UC4 for both dynamic control of the HGW and the execution of real time scheduling of data through the multiple paths.

As pointed out in section 1.2, both uDPM and uAUT functions are distributed between the UE (or the RGW in the context of UC4) and network elements including the UAG. The exact mapping of functions depends on what is required in terms of control. In case of simple decision scenarios (e.g. session based, or based on with fixed rules) the major part of the decisions could be taken in a centralised CP located in the Main CO. If more complex decision scenarios were envisaged, this would require implementing part of the function in the RGW and in Main COs. In particular, the RGW has state information about the multiple access paths (e.g. Wi-Fi link state, signal power, etc.) performance parameters.

The decision regarding the possible implementation options should consider the following aspects in addition:

- Data plane

A distributed solution would reduce the DP traffic (beyond the Main CO) in general since all the traffic has to be transported via the DP of the UAG.

The latency difference between both COMBO architectures (centralised versus distributed) is in the order of a few ms, which is not significant for the latency requirements identified previously. Indeed, a large part of the offset is due to the propagation delay, which is 1 ms for a 200 km distance. A more precise computation has been made based on a delay analysis for the German network (fibre propagation and switch processing delay considered), providing a difference in latency smaller than 5 ms (see section 3.2.2 of [4]). Even for dynamical control of tasks, this difference may not be relevant as long as the CP doesn't need to react below this value; this could be the case for some industrial application services which require switching traffic very quickly from one interface to the other.

- Control plane

When one tries to evaluate the centralised versus the distributed approach there is a large number of possible implementation options combined with the various implementation strategies of actual players.

Within this context, scalability issues might become of interest. For simple decision scenarios taking into account only a few parameters as described in sections 2.1.6.2 and 4.3.1, locating the CP of the UAG in the Core CO is feasible. Moreover, if DCs are located in Core COs or higher in the network, it makes sense to deploy the UAG CP in the Core COs, in order to favour positive scaling effects in relation to compute and storage resource. Lastly, a CP located in the Core CO can get an accurate picture of the state of core network segments in terms or traffic load, QoS parameters and others.

For more complex or highly dynamic decision scenarios, decisions should be made almost in real time, based on a large number of parameters such as e.g. the load situation of the access links and the signal power level of the LTE and Wi-Fi links. In these cases, locating the UAG CP in the Main CO could be more appropriate. The monitoring process controlled from the Main CO would also be simpler.

To summarize traffic management is less complex on shorter distance.

Lastly, locating CP and DP functions at Main COs implies that the effective number of users impacted in case of failures are limited versus a Core CO location.

Considering the influence on implementation options regarding the classical HGW versus NERG virtual HGW approach (introduced in Section 2.1.6.1) there should not be any additional issue as discussed before.

At present the influence of for example a provider agnostic implementation can't be reviewed but would have of course some influence on the decision to be made. Other business model related influencing factors (e.g. wholesale interfaces) might have an impact on selecting the optimal solution. Wholesale customers are influenced by the latter of investment barriers which means more distributed interfaces need higher own investments.

5.4 UC6: convergence of fixed, mobile and Wi-Fi gateway functionalities

UC6 aims at the integration of fixed, mobile and Wi-Fi functionalities in the same network entity. The main motivation of this UC is to realise a more efficient operation of transport/control functionalities and optimise costs by reducing the number of network elements. This UC was the origin of the development of the UAG functional entity detailed in chapters 1 and 2. The uDPM functionality inside the UAG is of special interest of this UC, as it contains the key functionality for the transport and control of all access networks, whereas uAUT is an optional requirement for UC6 targets.

Figure 51 represents the main UAG deployment alternatives considered of the CP and DP in the environment of UC6. This picture is similar to Figure 48, however, users and access networks can continue working independently of a UAG approach, and further changes are not needed. That is because UC6 target is to transform the core network, being transparent with the deployed access networks and user services.

The reduction of network elements can be achieved if the CP and DP of the UAG are integrated in the same equipment, or are co-located (compared to the situation without the UAG). Additional reduction can be achieved if the CP and DP can be separated, so the CP and DP can be dimensioned independently from one another.

Distributing the UAG DP in the Main COs can be beneficial to achieve a lower latency, reduce traffic in the transport network, enable a faster offload of mobile traffic, improve scalability and facilitate the integration with locally distributed services. If a fine-grained control of the data paths is required, the UAG CP should also be located in the Main COs.

On the other hand, locating the UAG DP in the Core COs can reduce costs as the number of complex equipment is limited. This is possible if the previous requirements on latency, traffic and scalability are relaxed. It is possible to distribute the DP in the Main COs with remote CP in the Core COs as long as the decisions to be taken (e.g. by the uDPM DE) are not too complex.

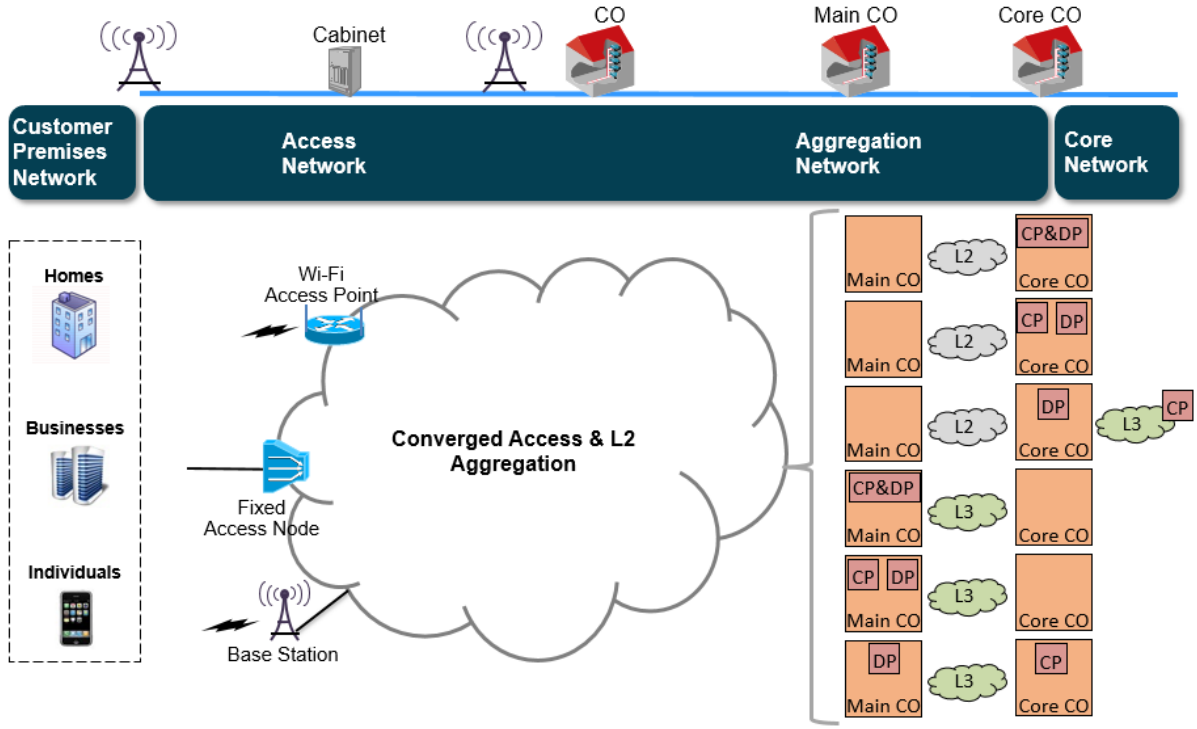


Figure 51: Implementation options for UC6

The preferable solution thus depends on the objectives of the network operator. However, in all cases, a split UAG allows to independently scale DP and CP functions, which is desirable.

5.5 UC8: network sharing

This UC aims to reduce deployment and operational costs and support more flexible business models by utilising existing infrastructure for both fixed and mobile communications as much as possible. It has both technical and business aspects for the network operator. The technical aspects necessitate a flexible integration of DP packet processing and of the CP functions of multiple operators into a converged access and aggregation framework. The business part implies that competing operators have to cooperate via a wholesale network company, which shares the network resources and operates the UAG. UC8 with the main UAG deployment alternatives is depicted in Figure 52.

From the wholesale operator point of view, a centralised solution requires less POPs, reducing connection costs and promising an easier management of the POPs. Although a distributed solution requires more POPs and thus potentially increases CAPEX, it also brings more flexibility to the control of network resources.

In an FMC network, where the main network functions converge in the UAG, the deployment implementation type and location of the UAG may facilitate the integration with the business model of the multi-operator environment. Operators, whatever their type, may have different sharing requirements. In the business models where regional operators exist and local management of resources are required, distributed UAG alternatives are recommended. On the contrary, if operators work on a nation-wide, UAG solutions located at Core CO are preferable. Finally, the UAG location has a low impact in case of business level virtual operator (see Section 4.4) because VNOs do not operate directly the shared network resources and the interaction with the wholesale network owner is made through interconnection points and signed SLAs.

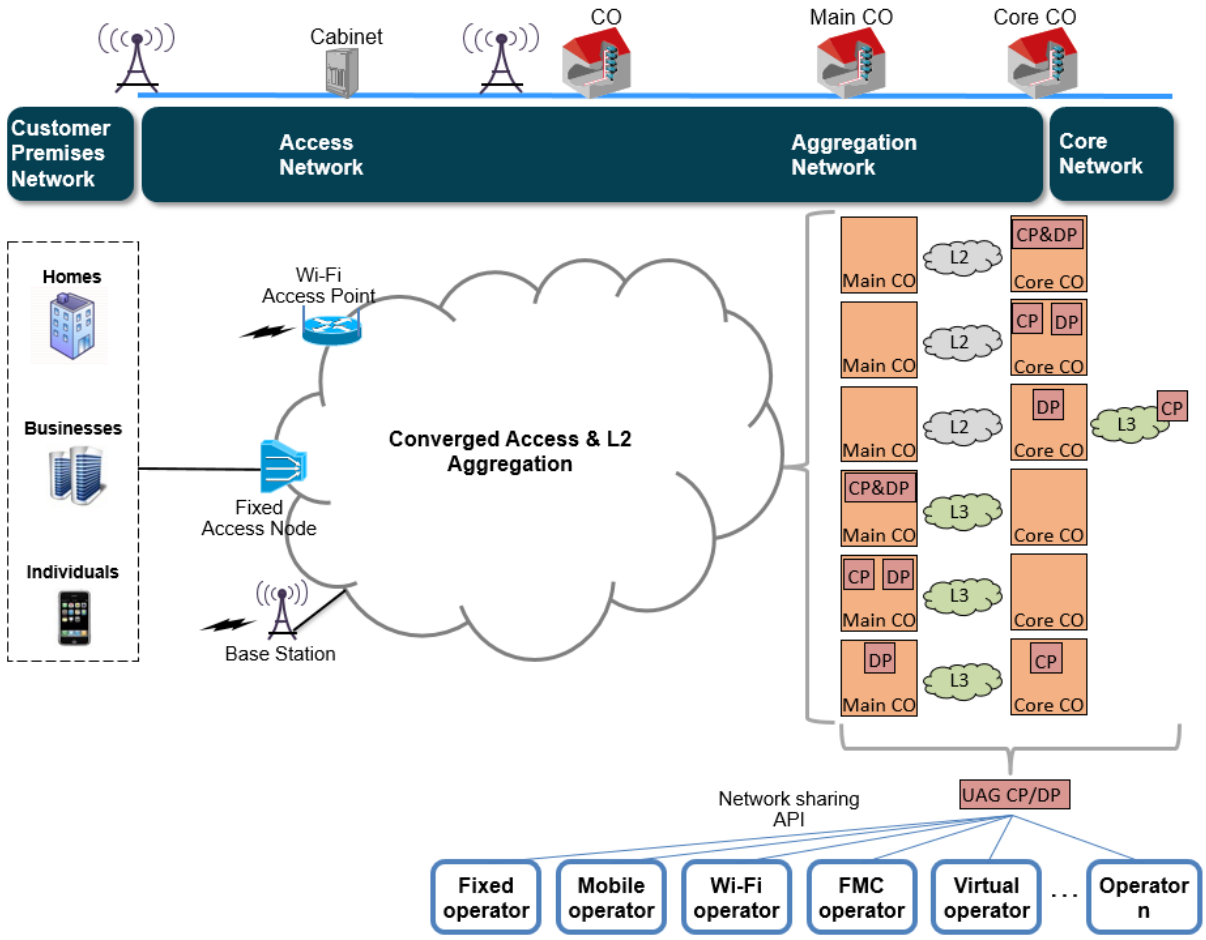
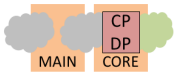
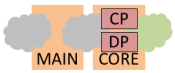
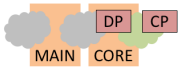
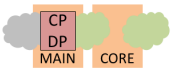
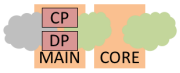
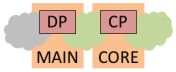


Figure 52: Implementation options for UC8

5.6 Qualitative assessment of implementation options on Use Cases

Table 13 includes the summary of the qualitative assessment of the UC. It shows the different implementation options of the UAG and the main deployment locations. For each UC, it is assessed CP and DP (execution) aspects.

Table 13: Qualitative comparison of UAG implementations regarding WP2 use cases

Centralised COMBO architecture			Distributed COMBO architecture		
Standalone UAG at Core CO (0) 	Split UAG with co-located DP and CP at Core CO (0) 	Split UAG with nonadjacent DP and CP and DP at Core CO 	Standalone UAG at Main CO (0) 	Split UAG with co-located DP and CP at Main CO (0) 	Split UAG with nonadjacent DP and CP and DP at Main CO 
Control aspects of UC1: unified FMC for mobile devices (1)					
+	+	-	++	++	+
Execution aspects of UC1: unified FMC for mobile devices (1)					
+/-	+/-	+/-	++	++	++
Control aspects of UC2: converged content caching for unified service delivery (2)					
+	+	+	+/-	+/-	++
Execution aspects of UC2: converged content caching for unified service delivery (2)					
+	+	+	++	++	++
Control aspects of UC4: universal access bundling for residential gateway (3)					
+	+	+/-	+	+	++
Execution aspects of UC4: universal access bundling for residential gateway (3)					
+	+	+	++	++	++
Control aspects of UC6: Convergence of fixed, mobile and Wi-Fi gateway functionalities (4)					
+	+	+/-	+	+	++
Execution aspects of UC6: Convergence of fixed, mobile and Wi-Fi gateway functionalities (4)					
+	+	+	++	++	++
Control aspects of UC8: network sharing (5)					
++	++	+	+/-	+/-	++
Execution aspects of UC8: network sharing (5)					
++	++	++	+/-	+/-	+

++ = a very good fit for the architecture
+ = a good fit
+/- = some positive and negative aspects
- = a bad fit
-- = a very bad fit
IRR: irrelevant to this particular implementation

Notes:

(0) Implementing DP and CP functions in a single equipment versus co-locating them in the same location only differs in terms of scalability performance. Co-location is made possible thanks to SDN, and allows to manage the scalability of DP and CP independently from one another

(1) uAUT can follow a centralised or distributed architecture with a good performance; however, uDPM may need a distributed approach to perform versatile and fast decision engine with real time network management rules. The CP will also benefit to be located close to the customer to enhance latency and scalability. See section 5.1.

(2) A centralised CP will be able to increase the caching effectiveness because it can serve a larger number of users compared to a decentralised CP. Additionally, a distributed DP enables a lower latency and a higher scalability. See section 5.2.

(3) Distributed CP and DP is preferable to enable dynamic rules in the decision engine and to reduce CP traffic. If simple control is implemented, locating the CP in the Core COs is more appropriate. See section 5.3.

(4) A distributed approach is more recommendable if very dynamic and fine-grained decisions are required, as it can deal with higher requirements regarding latency, scalability, traffic reduction, and integration with local services. On the other hand, if simpler rules with a coarser granularity have to be enforced, centralising the CP at Core CO may limit OPEX by simplifying network control implementation. See section 5.4.

(5) An approach with the UAG centralised may be more efficient in terms of cost and resources. If a more distributed DP is needed, an approach with the DP in the Main CO and the CP in the Core CO can be selected as an intermediate good solution. See section 5.5.

We can derive some global conclusions from the above qualitative analysis:

- In general, it is more appropriate to distribute the DP of the UAG in the Main COs, except for the Network Sharing UC for which deploying the DP of the UAG in the Core COs would be more cost effective. In all the other use cases, distributing the UAG CP in Main COs allow to take advantage of path diversity while efficiently aggregating traffic between Main and Core COs.
- Unless a very fine grained control is required, locating the UAG CP in the Core COs is acceptable, and desirable is some use cases in order to efficiently interact with DC management of content distribution.
- Splitting the UAG is beneficial in all cases, as it allows to independently scale DP and CP equipment.

6 Conclusion

This deliverable develops and analyses two major COMBO architectures, based on NG-POP, for implementing the universal subscriber and user AUThentication (uAUT) and the universal Data Path Management (uDPM), the key functional blocks for functional integration of fixed, Wi-Fi and mobile networks.

One COMBO architecture, called Distributed NG-POP, relies on a large number of NG-POP locations at the current Main COs, on which the common IP edges and other functional blocks would be distributed, benefiting also from optical ANs consolidation. This leads to an extension of the IP backbone towards the access network. The other COMBO architecture, called Centralised NG-POP, includes a smaller number of NG-POP locations, typically at the sites of core COs, which are the edges between current fixed aggregation network and core network.

The implementation of uAUT and uDPM in these two NG-POP architectures is formalized through the definition of a functional entity called Universal Access Gateway (UAG), the DP of which is located at NG-POPs and includes the common subscriber IP edge for all network types. Specifying the UAG CP and DP, their respective locations and how they can be deployed and implemented in the two (Distributed and Centralised) NG-POP architectures is a main achievement of this deliverable. This analysis is enriched by a study of network sharing capabilities of UAG-based networks and a qualitative assessment of distributed and centralised NG-POP architectures for implementing network services and for realising FMC use cases defined in WP2.

Final recommendations for fixed mobile network integration in 5G context will be drawn in deliverable D3.6, taking into account the main outcomes of COMBO architecture comparisons of this deliverable D3.5 and 5G transport considerations of D3.4, together with the main experimental achievements of D6.3.

The key outcomes of the developments and analyses performed in this deliverable D3.5 are summarized in the following.

FMC and the UAG concept bring new business opportunities

Functional convergence can bring benefits in many situations, either for a single integrated operator, in the framework of relationships between network operators, or lastly in facilitating the distribution of OTT services to the customers of network operators.

Integrated operators run multiple networks with different access types, enabling the highest degree of convergence, therefore with the highest benefits such as a lower latency, less physical and logical interfaces and less network equipment. Having a common IP edge (UAG) for all individual traffic flows of fixed, mobile and Wi-Fi users of a given area, allows co-locating, with the UAG, other entities such as content distribution servers. This makes the UAG the natural gateway for all services distributed over the IP layer by integrated operators.

Converged network enables (virtual) network operators ((V)NO), either mobile or fixed, to cooperate in order to extend their range. Operators enable their subscribers to access to broadband services from multiple access technology options, as allowed

by the uDPM functionality of UAGs provided by an integrated operator or infrastructure provider. A unified authentication framework, such as uAUT, embedding natively both user and subscriber authentication in an FMC environment, and making the link between users and subscribers, also opens new opportunities for collaborations between network operators and OTTs.

The UAG is a key functional entity for uAUT and uDPM

The UAG definition separates the UAG DP from the UAG CP to allow various benefits including scalability, implementation and deployment flexibility.

The UAG DP is the subscriber IP edge for all network types (fixed, Wi-Fi, mobile) and is thus playing the role of S/P-GW for 4G and of BNG for fixed networks. In addition to legacy user traffic processing functions and DP level monitoring functions, UAG DP includes the "Session Mapping Execution" functional block of uDPM, which realises the mapping and distribution of traffic between the multiple data paths of a given user session.

The UAG CP includes functions that are essential for implementing user access and session control in FMC context, namely through uAUT and uDPM. Specifically, UAG CP includes:

- a uAUT agent (proxy or client) facing the uAUT server, as a front-end of the AAA services in order to allow authenticating to multiple access networks on a single logical network;
- Network parts of uDPM, which consists in the network part of uDPM Decision Engine, network part of uDPM Data Path Creation/Destruction, Path Coordination and Control, and direct control of UE. In addition to the advanced FMC features of uDPM, these functions perform in a unified way the current control functions included in the current BNG and EPC gateways, as well as the mobility control functions included in MME.

Also, in addition to access and session control, the UAG CP incorporates service control, such as resources and policy control and charging control. This aims at controlling the DP equipment by taking into account the operator policies, the subscriber profile, the access network information, and the requested service.

The UAG can be implemented in a flexible way within FMC architecture

Two implementation models can be considered for the UAG: the standalone model where no interface is specified between CP and DP and the split model, which relies on an explicit interface defined between CP and DP. The split model will be largely enabled by SDN techniques. NFV is also relevant, particularly for the UAG CP entity, since it enables hosting related functions on commodity servers.

The standalone model can be considered for an incremental implementation of the UAG, where the current BBF and 3GPP functional entities (BNG, S/P-GW, MME, etc.) are merged into a single node, so that the UAG can be regarded as a kind of structurally converged subscriber IP edge.

The split model allows a disruptive implementation of the UAG and gives the opportunity to merge the fixed and mobile subscriber IP edge functions into common generic functions, either in DP or in CP, possibly also with common interfaces and

protocols. The UAG can be then be considered as a true functionally converged subscriber IP edge entity.

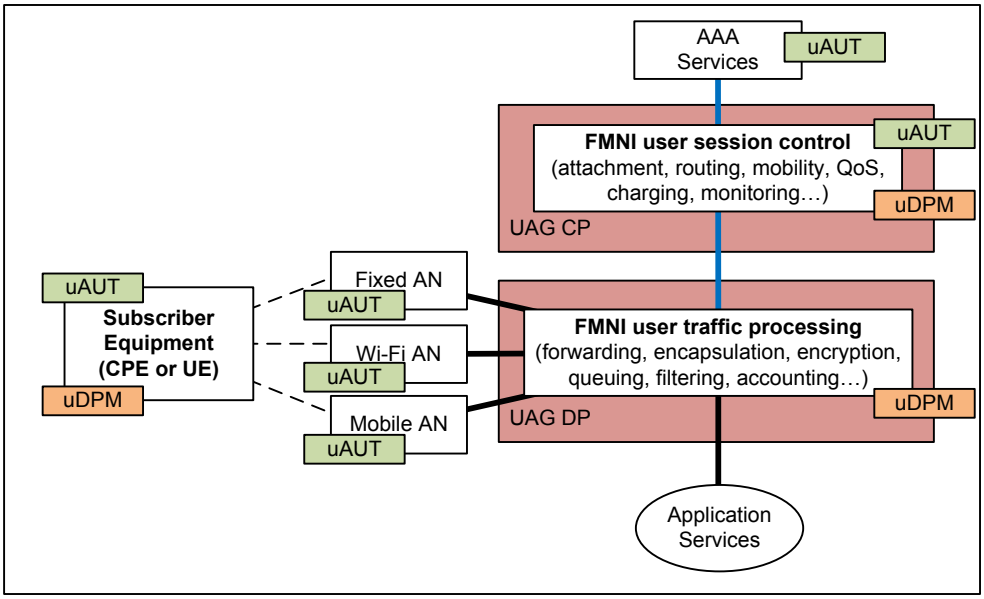


Figure 53: Disruptive implementation of the UAG as a functionally converged subscriber IP edge

In the Distributed NG-POP architecture, the UAG DP is located at Main COs, which allows low DP latencies but requires distributed mobility anchoring and the extension of the IP network down to the Main COs.

In the Centralised NG-POP architecture, the UAG DP is located at Core COs, which does not require extension of the IP network, although leading to higher DP latencies than in Distributed NG-POP architecture.

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, or located remotely to allow more centralisation of control functions. The UAG CP can even be located in very few locations inside the core network, similarly to EPC entities of current mobile networks.

The distributed NG-POP architecture with remote UAG control plane fosters advanced FMC features together with key legacy functionalities

Table 14 summarizes the overall comparative assessment of Distributed and Centralised NG-POP architectures on overall criteria including key data plane aspects. The Distributed NG-POP architecture has the best features on most of the overall assessment criteria, although the Centralised NG-POP architecture performs better in terms of network migration, cost, energy and deployment efficiency.

Table 14: Comparing the respective efficiency of Distributed and Centralised NG-POP architectures in terms of function realisation

Comparison criteria	**Centralised NG-POP UAG DP @ core CO**	**Distributed NG-POP UAG DP @ main CO**
Scalability	Good	Best
Reliability	Good	Best
Latency for network services and user applications	Low	Lowest
CAPEX, OPEX, energy and deployment efficiency	Best	Good
Network migration, changes in IP network, routing management	Moderate impact	High impact
Traffic reduction in aggregation and core networks	Good	Best
Integration with local network services	Good	Best

In addition to this overall assessment of COMBO architectures, specific network functionalities, especially control plane aspects, have to be considered. For example, uAUT implementation is not influenced by the chosen architecture, as most of the authentication procedure relies on a centralised system, and the UAG presents only a client or proxy to the centralised server. The Distributed NG-POP architecture with co-located CP and DP at main COs is the most relevant option when considering uDPM implementation, but having the UAG CP distributed at main COs would generate a lot of signalling messages related to mobility management. Table 15 summarizes the specific assessment of COMBO architectures on key control plane aspects.

Table 15: Qualitative assessment of COMBO architectures in terms of control plane implementation efficiency

Assessment on key control plane aspects	**Centralised NG-POP UAG DP @ core CO**		**Distributed NG-POP UAG DP @ main CO**
	UAG CP in core network	**UAG CP @ core CO**	**UAG CP @ main CO**
Implementation of uAUT	No impact	No impact	No impact
Implementation of uDPM CP	Bad	Good	Best
Mobility management	Bad	Best	Bad

For all these reasons, the Distributed NG-POP architecture with remote UAG CP at core COs appears as the best option allowing advanced FMC features together with key legacy functionalities.

The distributed NG-POP architecture allows an efficient support of legacy and advanced services

In COMBO architectures, the IP edges for all types of access networks are merged and located in the UAG DP: this facilitates the delivery of services which may be requested by the user on any available access network.

For example, a Cache Node can be co-located with the UAG DP, allowing improved content delivery and advanced caching and prefetching functionalities thanks to interactions between Cache Controllers and the uDPM Decision Engine. Although the operating costs could be increased in the distributed COMBO architecture compared to the centralised solution, having the UAG DP at Main COs together with Cache Nodes, while having remote UAG CP at Core COs, will significantly reduce latency and improve performance of content delivery.

Considering real-time communication services, the difference in latencies between the Distributed and Centralised NG-POP architectures is not sufficient to make a distributed architecture absolutely necessary, and the Centralised architecture is also compatible with new approaches such as Mobile Edge Computing. Nevertheless, the distributed architecture is better for handling advanced uDPM features such as vertical handover between access technologies for real-time communication services.

Some emerging cloud services, such as cloud gaming or augmented reality, require ultra-low latency in the data path, while others have more traditional service requirements. However, the CP function implementation does, in most cases, not have these delay requirements and can thus be implemented further from the user. This makes the split UAG implementation with CP at Core COs advantageous for cloud-based services, with a preference for the Distributed NG-POP architectures when considering strict DP latency requirements of 20 ms or below (as for cloud-gaming and virtual / augmented reality applications).

When considering Internet of Things, the total amount of traffic is limited and all implementations of the UAG DP can be used. Nevertheless, the UAG CP should be able to manage the amount of signalling generated by massive MTCs, which fosters Distributed NG-POP architecture with DP and CP co-located in the Main COs.

The distributed NG-POP architecture with remote UAG control plane is also promising for the network slicing approach

The split UAG with remote CP should be the best scenario from both the DP resource allocation and control perspective of network slicing. Specifically, implementing the UAG DP at Main COs allows creating and composing network slice infrastructures with differentiated service requirements. Having a remote CP at Core COs allows having a centralised and complete view of all resources leading to handle finer resource granularity for partitioning and composing network slices. Nevertheless, the functional architecture applied to each network slice itself can of course be adapted to the application targeted by the slice, e.g. an IoT slice will benefit from having UAG DP and CP functionalities co-located in the Main COs.

Both Distributed and Centralised NG-POP architectures enable FMC use cases

Table 16 summarises a qualitative analysis of COMBO architectures for realising FMC use cases. It shows that Centralised NG-POP is a good trade-off, especially when the UAG CP is co-located with the UAG DP at Core COs. This centralised NG-POP architecture already represents a significant improvement compared to the current architectures in terms of gateway distribution, i.e. scalability, reliability and latency, as the core COs are closer to the subscriber than the EPC gateways of current mobile networks. Nevertheless, the Distributed NG-POP architecture with remote CP, i.e. with the UAG DP at main COs and the UAG CP at core COs, appears even more appealing for most of the FMC use cases.

Table 16: Qualitative assessment of COMBO architectures in terms UC realisation

Use case	**Centralised NG-POP UAG DP @ core CO**		**Distributed NG-POP UAG DP @ main CO**	
	UAG CP in core network	**UAG CP @ core CO**	**UAG CP @ core CO**	**UAG CP @ main CO**
UC1: unified FMC access for mobile devices	Bad	Good	Very good	Best
UC2: converged content caching for unified service delivery	Good	Good	Best	Good
UC4: universal access bundling for residential gateway	Bad	Good	Best	Very Good
UC6: Convergence of fixed, mobile and Wi-Fi gateway functionalities	Bad	Good	Best	Very Good
UC8: network sharing	Very good	Good	Very good	Good

Furthermore, splitting the UAG CP functions between main COs and core COs, i.e. having some control functions in main COs and some other control functions in core COs can bring additional benefits to the Distributed NG-POP architecture with a split UAG model. This tuning of the location of UAG CP functions will be enabled by SDN and NFV techniques and allows optimisation of the network for a large variety of use cases and services.

An alternative, also relying on SDN and NFV, can be to specify different network slices for which the respective locations of CP functions differ.

These aspects will be further discussed in D3.6.

References

[1] COMBO deliverable D3.1: “Analysis of key functions, equipment and infrastructures of FMC networks”, V2.0, June 2014.

[2] COMBO deliverable D2.1: “Framework reference for fixed and mobile convergence”, V2.0, June 2014.

[3] COMBO deliverable D3.2: “Analysis of horizontal targets for functional convergence”, V2.0, September 2015.

[4] COMBO deliverable D3.3: “Analysis of transport network architectures for structural convergence”, V2.0, September 2015.

[5] COMBO Deliverable D6.2, “Detailed description of COMBO demonstration”, August 2015.

[6] COMBO Deliverable D6.3, “Report describing results of operator testing, capturing lessons learned and recommendations”, July 2016.

[7] Atzori Luigi, Iera Antonio, and Giacomo Morabito. "The internet of things: A survey." Computer networks 54, no. 15 (2010): 2787-2805.

[8] Osman Yilmaz, “5G Radio Access for Ultra-Reliable and Low-Latency Communications” Ericsson Research Blog, http://www.ericsson.com/research-blog/5g/5g-radio-access-for-ultra-reliable-and-low-latency-communications/, visited on 2015 July 1st.

[9] Mobile Experts, “Defining 5G: Setting Targets based on the Business Case”, Mobile Experts LLC, September 2014, http://mobile-experts.net/Documents/Tocs/ExpInsight_5G_Sept_2014_TOC.pdf

[10] Latvakoski, J.; Alaya, M.B.; Ganem, H.; Jubeh, B.; Iivari, A.; Leguay, J.; Bosch, J.M.; Granqvist, N. Towards Horizontal Architecture for Autonomic M2M Service Networks. *Future Internet* **2014**, *6*, 261-301.

[11] METIS D1.5, Updated scenarios, requirements and KPIs for 5G mobile and wireless system with recommendations for future investigations, Apr 2015.

[12] METIS D6.6 Final report on the METIS 5G system concept and technology roadmap, Apr 2015,

[13] Miorandi, D.; Sicari, S.; de Pellegrini, F.; Chlamtac, I. Internet of things: Vision, applications and research challenges. Ad Hoc Networks, vol 10, 2012, pp 1497-1516.s

[14] Afif Osseiram, “The 5G Wireless System, An introduction”, presentation at the 3rnd Internation Workshop on Next Generation Green Wireless Networks, 1-3 October 2014, Rennes,

[15] “SDN architecture”, ONF TR-502, June 2014.

[16] “OpenFlow Switch Specification ”, Version 1.5.0, ONF-TS-020

[17] JP. Vasseur (ed.) and JL. Le Roux (Ed.), “Path Computation Element (PCE) Communication Protocol (PCEP)”, IETF RFC 5440, March 2009.

[18] R. Enns (ed.), M. Bjorklund (Ed.), J. Schoenwaelder Ed.) and A. Bierman (ed.), "Network Configuration Protocol (NETCONF)", IETF RFC 6241, June 2011.

[19] B. Pfaff and B. Davie (Ed.), "The Open VSwitch Database Management Protocol", IETF RFC 7047, December 2013.

[20] "Network Function Virtualisation (NFV); Architectural Framework", ETSI GS NFV 002, 2013.NGMN. 5G whitepaper. NGMN Alliance, 17-February-2015. https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

[21] NGMN. Guidelines for LTE Backhaul Traffic Estimation, v0.4.2 July 2011

[22] "Architecture enhancements for non-3GPP accesses," Release 10, Technical specification TS 23.402, 2012.

[23] X. Lagrange, "Very Tight Coupling between LTE and Wi-Fi for Advanced Offloading Procedures". WCNC 2014, IEEE Wireless Communications and Networking Conference, 06-09 avril 2014, Istanbul, Turkey, 2014, pp. 82-86

[24] Y. Khadraoui, X. Lagrange, A. Gravey, "Very Tight Coupling Between LTE and Wi-Fi: From Theory To Practice". WD 2016, IFIP Wireless Days, 23-25 march 2016, Toulouse, France, 2016

[25] Y. Khadraoui, X. Lagrange, A. Gravey, "Very Tight Coupling between LTE and Wi-Fi: a Practical Analysis". CoRes 2016 : Rencontres Francophones sur la Conception de Protocoles, l'Évaluation de Performance et l'Expérimentation des Réseaux de Communication, 23-24 may 2016, Bayonne, France, 2016, pp. 1-4.

[26] T. Taleb, et. al., "EASE: EPC as a service to ease mobile core network deployment over cloud", IEEE Network, March-April 2015.

[27] A. Tzanakaki, et. al., "Virtualisaiton of heterogenous wireless-optical network and IT infrastructures in support of cloud and mobile cloud services", IEEE Communications Magazine, August 2013.

[28] R. Muñoz, et. al., "Integrated SDN/NFV management and orchestration architecture for dynamic deployment of virtual SDN control instances for virtual tenant networks", IEEE7OSA Journal of Optical Communications and Networks (JOCN), November 2015.

[29] D. R. Gnyawali, and B.-J. Park, "Co-opetition between giants: Collaboration with competitors for technological innovation," Research Policy, vol. 40, Issue 5, pp. 650663, June 2011.

[30] R. Mijumbi, et al., Network Function Virtualisation: State-of-the-Art and Research Challenges, in Communications Surveys & Tutorials, IEEE, vol.18, no.1, pp.236-262, First Quarter 2016

[31] S. Gosselin, et al., Fixed and mobile convergence: Needs and solutions, in European Wireless 2014; 20th European Wireless Conference, May 2014

[32] Marco Savi, Ali Hmaity, Giacomo Verticale, Stefan Höst, Massimo Tornatore: To distribute or not to distribute? Impact of latency on Virtual Network Function

distribution at the edge of the FMC network, 18th International Conference on Transparent Optical Networks (ICTON), July 2016 (Invited paper)

[33] A. Fischer, et al., Virtual Network Embedding: A Survey, in Communications Surveys & Tutorials, IEEE, vol.15, no.4, pp.1888-1906, Fourth Quarter 2013

[34] M. Savi, M. Tornatore, G. Verticale, Impact of Processing-Resource Sharing on the Placement of Virtual Network Functions, submitted to IEEE/ACM Transactions on Networking, 2016

[35] FP7 COMBO Project, Analysis of transport network architectures for structural convergence, Deliverable D3.3, Jun. 2015

[36] ETSI, Mobile-Edge Computing, Introductory Technical White Paper, Sept. 2014

[37] http://www.cloudping.info

[38] ETSI GS NFV-MAN 001V 1.1.1, "Network Functions Virtualisation (NFV); Management and Orchestration", Dec. 2014

[39] S. Taylor, "The next generation of the internet," CISCO Point of view, April 2013.

[40] M. K. Weldon, Ed., The Future X Network: A Bell Labs Perspective. CRC Press, 2016.

[41] CISCO, "Cisco visual networking index: Forecast and methodology, 2014-2019," White paper, May 2015.

[42] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, ser. MCC '12. New York, NY, USA: ACM, 2012, pp. 13–16.

[43] P. Bosch, A. Duminuco, F. Pianese, and T. L. Wood, "Telco clouds and virtual telco: Consolidation, convergence, and beyond," in Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on. IEEE, 2011, pp. 982–988.

[44] N. Fernando, S. W. Loke, and W. Rahayu, "Mobile cloud computing: A survey," Future Generation Computer Systems, vol. 29, no. 1, pp. 84–106, 2013.

[45] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," Wireless communications and mobile computing, vol. 13, no. 18, pp. 1587–1611, 2013.

[46] "MEC, Mobile Edge Computing, ETSI Industry specification group," Started Dec 2014.

[47] C. Moreno, N. Tizon, and M. Preda, "Mobile cloud convergence in gaas: A business model proposition," in System Science (HICSS), 2012 45th Hawaii International Conference on. IEEE, 2012, pp. 1344–1352.

[48] Sony Playstation Now. [Online]. Available: http://us.playstation.com/playstationnow

[49] M. Claypool and K. Claypool, "Latency and player actions in online games," Communications of the ACM, vol. Vol. 49, pp. 40–45, 2006.

[50] S. Eido and A. Gravey, "How much LTE traffic can be offloaded?" in Advances in Communication Networking. Springer, 2014, pp. 48–58.

[51] S. Eido, P. Mitharwal, A. Gravey and C. Lohr. “MPTCP Solution for Seamless Local SIPTO Mobility”. HPSR 2015 : 16th International Conference on High Performance Switching and Routing , 01-04 july 2015, Budapest, Hungary, 2015

[52] 3GPP, ”Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO),” Technical report, Release 10, TR 23.829,” 2011.

[53] 3GPP, ”Local IP Access (LIPA) mobility and Selected IP Traffic Overload (SIPTO) at the local network,” Technical report, Release 12, TR 23.859,” 2013.

[54] 3GPP, ”Evolved Universal Terrestrial Radio Access (EUTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN),” Technical report, Stage 2, Release 12, TS 36.300,” 2014.

[55] 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.

[56] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, “RFC 6824, TCP extensions for multipath operation with multiple addresses,” 2013.

[57] Broadband Forum TR-069 Amendment-5 – CPE WAN Management Protocol

[58] Broadband Forum TR-181 Amendment 5 - Device Data Model for TR-069

[59] Broadband Forum TR-098 Amendment 2 - Internet Gateway Device Data Model for TR-069

[60] Home –Gateway-Initiative: HGI-RD044, HOME GATEWAY BASE REQUIREMENTS: RESIDENTIAL PROFILE, 2 May, 2016;

[61] Motivations, use cases and Models of VCPE, draft-fu-dmm-vcpe-models-01, Fu, Ed. H. Deng, China Mobile, October 19, 2015

[62] Broadband Forum WT-348, Hybrid Access Broadband Network Architecture, Revision: 13, Revision Date: February 2016

[63] Broadband Forum TR-124, Functional Requirements for Broadband Residential Gateway Devices, Issue: 1.0, Issue Date: December 2006

[64] Erik Nygren, Ramesh K. Sitaraman, and Jennifer Sun. 2010. “The Akamai network: a platform for high-performance internet applications”. SIGOPS Oper. Syst. Rev. 44, 3 (August 2010), 2-19.

[65] Balachander Krishnamurthy, Craig Wills, and Yin Zhang. 2001. “On the use and performance of content distribution networks”. In Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement (IMW '01). ACM, New York, NY, USA

[66] Claudio Imbrenda, Wuyang Li, and Luca Muscariello. “Analyzing cacheable Traffic for FTTH Users using Hadoop”. In Proceedings of the 2nd International Conference on Information-Centric Networking (ICN '15). 2015, New York, USA.

[67] Claudio Imbrenda, Luca Muscariello, and Dario Rossi. “Analyzing cacheability in the access network with HACkSAw.” In Proceedings of the 1st international conference on Information-centric networking (ICN '14). 2014, ACM, New York.

[68] B. A. Ramanan, L. M. Drabeck, M. Haner, N. Nithi, T. E. Klein and C. Sawkar, "Cacheability analysis of HTTP traffic in an operational LTE network," Wireless Telecommunications Symposium (WTS), 2013, Phoenix, AZ, 2013, pp. 1-8.

[69] Salah-Eddine Elayoubi and James Roberts. 2015. “Performance and Cost Effectiveness of Caching in Mobile Access Networks”. In Proceedings of the 2nd International Conference on Information-Centric Networking (ICN '15). ACM, New York, USA.

[70] Sipat Triukose, Zhihua Wen, and Michael Rabinovich. 2011. “Measuring a commercial content delivery network”. In Proceedings of the 20th international conference on World wide web (WWW '11). ACM, New York, NY, USA

[71] Minlan Yu, Wenjie Jiang, Haoyuan Li, and Ion Stoica. 2012. Tradeoffs in CDN designs for throughput oriented traffic. In Proceedings of the 8th international conference on Emerging networking experiments and technologies (CoNEXT '12). ACM, New York, NY, USA

[72] Giyoung Nam and KyoungSoo Park. 2014. Analyzing the effectiveness of content delivery network interconnection of 3G cellular traffic. In Proceedings of The Ninth International Conference on Future Internet Technologies (CFI '14). ACM, New York, NY, USA

[73] 2011 Census: Population and household estimates for England and Wales http://www.ons.gov.uk/ons/rel/census/2011-census/population-and-household-estimates-for-england-and-wales/rft-h01.xls

[74] Á. Ladányi, T. Cinkler, A. Mitcsenkov, „Impact of optical access topologies onto availability, power and QoS”, DRCN2014, the 10th International Conference on the Design of Reliable Communication Networks (DRCN), Gent, Belgium, April 2014

[75] Á. Ladányi, T. Cinkler, Gy. Sallai, „Tradeoffs of a converged wireless-optical access network”, Networks 2014, the 16th International Telecommunications Network Strategy and Planning Symposium, Funchal, Madeira, Portugal, September 2014

[76] Á. Ladányi, T. Cinkler, A. Mitcsenkov, „Power saving tradeoffs in a multi-operator scenario”, HPSR 2015, the 16th IEEE International Conference on High Performance Switching and Routing, Budapest, July 2015

[77] T. Cinkler, Á. Ladányi, „Resilient access via 3D Hand-Over”, RNDM 2015, the 7th International Workshop on Reliable Newtwork Design and Modelling, München, Germany, October 2015

[78] Á. Ladányi, P. Olaszi, P. Varga, T. Cinkler, Network initiated Wi-Fi - LTE Handovers with Multipath TCP, Demo, HPSR 2015, the 16th IEEE International Conference on High Performance Switching and Routing, Budapest, July 2015

[79] A. Buttaboni, F. Musumeci, M. Tornatore, A. Pattavina, “Fostering Coopetition in Fixed Mobile Converged Networks”, submitted to Transactions on Emerging Telecommunications Technologies

[80] Tamás Koi, "Kiutaznak a pénztárgépek adatai az országból" (in Hungarian), (The data of cashier machines travel out of the country). July 17, 2013, Retreived on June 23, 2016, http://www.hwsw.hu/hirek/50619/penztargep-aee-nav-mobilhalozat-roaming-adatroaming.html

[81] 3GPP, ”Evolved Universal Terrestrial Radio Access (EUTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN),” Technical report, Stage 2. TS 36.300 version 13.3.0 Release 13.

[82] ETSI GS NFV 001 V 1.1.1, “Network Functions Virtualisation (NFV); use cases ”, October 2013

7 ANNEX: Further notes on "Multi-operator Game Theoretic Model for Effective Traffic Offloading"

In this annex, we will describe with some additional details the Game Theoretic approach for Network Sharing presented in Section 4.3.2, mostly by providing numerical examples of the game-theoretical methodology and some illustrative results obtained through simulations. Before that, in the next subsection, we explain more in detail the game-theoretic principles over which this work is based.

7.1 Game theory primer

As already mentioned in Section 4.2.2, Game Theory (GT) is the study of mathematical models representing conflict and cooperation (the "games") between rational players. To be fully defined, a game must specify: i) the *players* of the game, ii) the information and *actions* available to each player at each decision point, and iii) the *payoffs* for each decision. These elements are used to deduce a set of equilibrium strategies for each player (e.g., the Nash Equilibrium explained below). The game is usually represented by a matrix that shows the players, the strategies, and the payoffs. More generally it can be represented by any function that associates a payoff for each player with every possible combination of actions.

A game can be cooperative or non-cooperative: in a cooperative game, the players are able to form binding commitments whereas in a non-cooperative games players make decisions independently. Moreover, a game can be symmetric or a non-symmetric: in a symmetric game, the payoffs for playing a particular strategy depend only on the other strategies employed, not on who is playing them. If the players can be changed without changing the payoff to the strategies, then a game is symmetric, otherwise the game is asymmetric. Also, games can be zero-sum games or non-zero-sum games. In zero-sum games the choices of the players can neither increase nor decrease the available resources, such that the total benefit to all players in the game, for every combination of strategies, always adds to zero. In non-zero-sum game outcomes might have net results different than zero.

The game that we model is a non-cooperative, asymmetric, and non-zero-sum game, where each player chooses its strategy independently to improve its own performance (i.e., utility) or reducing its losses (i.e., costs). The players of this game are the network operators. The only information shared between players is the amount of traffic exceeding the capacity of the network of an operator, i.e., the amount of bandwidth that an operator wants to offload. In this game, the role of the player can change according to the load condition of each network: the player can be either the Offloader (and decide whether to offload or not) or the Receiver (and decide to accept or not to accept the traffic of the Offloader). This game is played every time an operator has some traffic exceeding the capacity of its network. The solution of the proposed game is represented by the Nash Equilibrium. The Nash equilibrium is a solution concept of a non-cooperative game in which each player is assumed to know the equilibrium strategies of the other players, and no player has anything to gain by changing only their own strategy.

7.2 Numerical examples of the proposed GT methodology

As an example, we represent a game with two players: player 1 (the Offloader) and player (the Receiver). In Table 17, Table 18 and Table 19, we describe three possible scenarios under different traffic assumptions and we compute the related payoffs to show numerically how the game is solved (i.e., the decision about the collaboration is taken) in different situations. Payoffs values, reported in the tables, are computed according to the equations in Table 10 (Section 4.3.2).

In Table 17 we represent a situation where the network of operator 1 (the Offloader) is lightly loaded, i.e., the load L_1 is lower than capacity C_1. In this case the solution of the game is given by a Nash Equilibrium that corresponds to the strategy where player 1 does not offload (player 2 can either accept or not the traffic).

Table 17 Players' payoffs, computed with C_1 = 100, L_1 = 60, Ex_1 = -40, Ex_2 = 0

		Player 1	
		Offload	Not Offload
Player 2	Accept	(0; -40)	(0; 41)
	Not Accept	(1; -40)	(1; 41)

In Table 18 we represent a situation where the network of operator 1 (the Offloader) is overloaded, i.e., the load L_1 is higher than the network capacity C_1. In this case the Nash Equilibrium corresponds with the strategy where player 1 offloads and player 2 accepts the traffic.

Table 18 Players' payoffs, computed with C_1 = 100, L_1 = 120, Ex_1 = 20, Ex_2 = 30

		Player 1	
		Offload	Not Offload
Player 2	Accept	(30; 20)	(30; 19)
	Not Accept	(1.96; 20)	(1.96; -19)

In Table 19, we represent a situation where the network of operator 1 (the Offloader) is overloaded but the maximum amount of traffic that can be accepted by the Receiver (Ex_1) is lower than the excess traffic that the Offloader wants to offload. In this case the Nash Equilibrium corresponds with the strategy where player 1 offloads and player 2 accepts to serve the traffic only up to Ex_1.

Table 19 Players' payoffs, computed with C_1 = 100, L_1 = 140, Ex_1 = 40, Ex_2 = 30

		Player 1	
		Offload	Not Offload
Player 2	Accept	(30; 40)	(30; -39)
	Not Accept	(1.96; 40)	(1.96; -39)

7.3 Numerical Results

To validate our proposed GT approach, we compare it against two benchmarking procedures: i) "No collaboration", where collaboration never occurs; ii) "Full collaboration", where the Receiver always accepts to serve the traffic of the Offloader. We initially consider a offered traffic scenario where the two operators have opposite traffic profiles, i.e., the traffic peak hours of the operators occur in two separate moments (shown in Figure 54). Offered traffic shown on the y-axis is normalized to the traffic of a single user. Capacity of the networks of both operators is C=100.

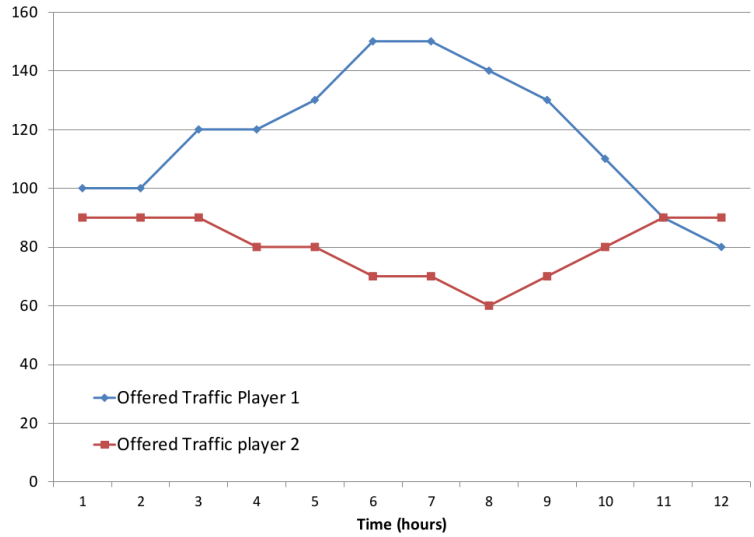


Figure 54 Daily traffic variation of the two operators

In Figure 55(a) and Figure 55 (b) we compare the average amount of traffic of one operator served by the other operator.

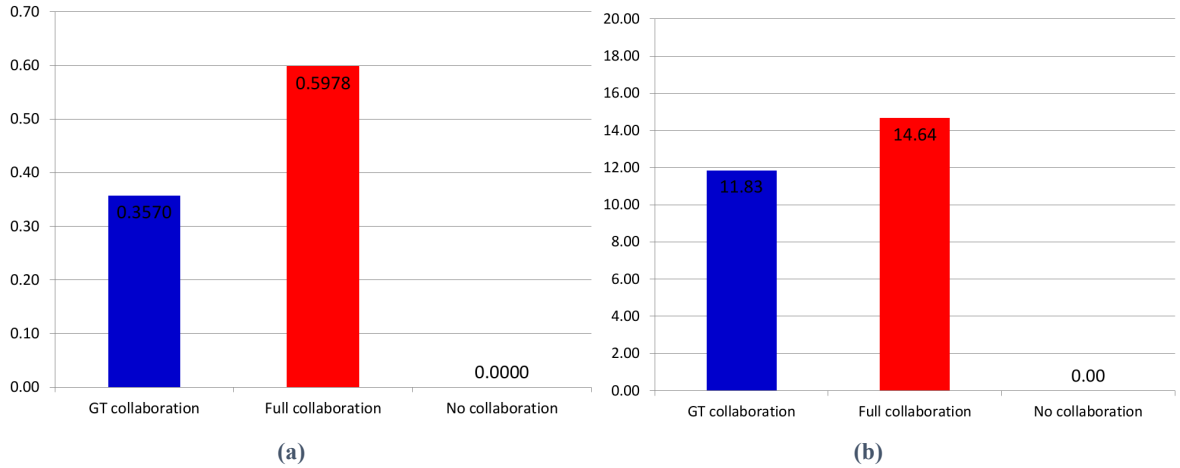


Figure 55. (a) Average traffic of Player 2 served by Player 1 and (b) Average traffic of Player 1 served by Player 2 over the day (i.e., 12 hours) in the different collaboration strategies.

In particular, Figure 55(a) shows the average amount of traffic of player 2 offloaded to the network of player 1. Since player 1 has a high offered traffic, it rarely has

available capacity that can be used to serve the traffic of player 2. In fact, the average amount of traffic of player 2 served by player 1 is always very low (less than 1), independently of the used approach. More interestingly, in Figure 55 (b) we show the average amount of traffic of player 1 served by player 2. As player 2 has much more opportunities to accommodate the traffic of operator 1, the average amount of traffic of player 1 served by player 2 is much higher than in Figure 55(a). Note that notice that offloaded for the case of "No collaboration" is always zero, as expected. Note also that the GT approach serves slightly less offloaded traffic compared to "Full Collaboration", which is exactly the objective of the GT approach, as, by serving slightly less offloaded traffic, the receiver can avoid penalizing its own traffic. This becomes clear in the next set of figures.

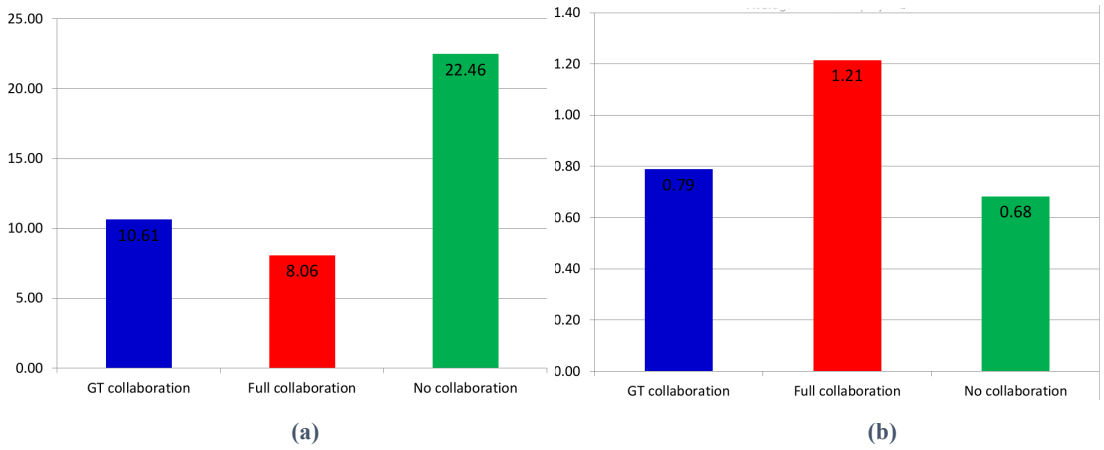


Figure 56. (a) Average traffic lost by player 1 and (b) by player 2 over the day (i.e., 12 hours) in the different collaboration strategies.

Figure 56(a) shows the average amount of traffic lost by player 1, while Figure 56(b) shows the average amount of traffic lost by player 2. In both cases the average is performed over the whole day. In Figure 56(a), the main reason for traffic loss is the high amount of offered traffic of player 1. "No collaboration" features the highest amount traffic loss, as no offload is allowed. With "Full Collaboration", the amount of lost traffic is the lowest, while "GT collaboration" only features a slightly higher loss with respect to Full Collaboration. On the other hand, in Figure 56(b), where we report the average traffic lost by player 2, loss is mostly due to an excess of collaboration. In fact, there are situations when users of player 2 cannot be served in the network of player 2 because this network is serving offloaded users of player 1. In this scenario, the GT collaboration approach is the most effective approach to protect the users of player 2, while still allowing for a very good amount of offload (comparable with the one of Full "Collaboration"). In fact, in Figure 56(b) we can see that the "GT collaboration" has the lowest amount of traffic lost. This happens because the GT collaboration approach tends to maintain a certain amount of bandwidth that can be used to serve the demands of the users that may arise in the subsequent time intervals. In conclusion, when two operators have contrasting traffic profiles, it suitable for the two operators to cooperate, and our proposed GT approach is able to maintain almost the highest possible quality of service for the users of the receiver network (almost as good as in "No collaboration") while still achieving most

of the benefits obtainable by offloading (as in the “Full Collaboration” case). These forms of network sharing are added values of functional fixed-mobile convergence. Additional results with multiple technologies are reported in our publication in [79].

- - - End of Document - - -