



Project Number:	CELTIC / CP7-011
Project Title:	<u>M</u> obile Networks <u>E</u> volution for <u>I</u> ndividual <u>C</u> ommunications Experience – MEVICO
Document Type:	D (external)

Document Identifier:	D 1.4
Document Title:	Architecture Design Release 3 Documentation
Source Activity:	WP 1
Main Editor:	Jose Costa-Requena, Jiejn HOU, Ivan Froger
Authors:	Khadija Daoud Triki, Zoltán Faigl, Michael Boc, Tero Lötjönen, Edgardo Montes de Oca, Jose Costa-Requena, Tapio Suihko, Pekka Korja, Franck Mornet, Jörgen Andersson
Status / Version:	Final
Date Last changes:	11.11.12
File Name:	MEVICO D1.3 Architecture Design

Abstract:	Based on the evolution of traffic and network usage, MEVICO proposes an enhanced mobile architecture for the LTE (Long Term Evolution) and LTE-Advanced. The architectural requirements related to different aspects (e.g. mobility, scalability, security...) are identified. These requirements allow identifying the related architecture challenges (network, applications, services...). MEVICO proposes technology solutions which address these challenges.
-----------	--

Keywords:	Mobile Broadband, traffic evolution, traffic properties, scenarios, use cases
-----------	---

Document History:	
06.11.2012	Document creation
06.21.2012	Replace Section 6 with new content
06.27.2012	Document completion and ready for review
07.20.2012	Document final version
11.11.2012	Cleanup and editing

1	Table of Contents	
1	Table of Contents	2
2	Table of figures	4
	Authors	5
3	Executive Summary	8
4	List of terms, acronyms and abbreviations	9
5	References	15
1.	Introduction	16
1.1	Business drivers for the MEVICO project	16
1.2	Overview of Evolved Packet Core (EPC) architecture.....	16
1.3	Other network elements linked to the Evolved Packet Core (EPC)	17
2.	Network Traffic and Usage Scenarios	18
2.1	Mobile traffic, service, and technology evolution 2008-2020.....	18
2.1.1	Traffic Data evolution	18
2.1.2	Services and application evolution.....	18
2.1.3	Evolution-enabling technologies.....	18
2.1.4	Aspects to be taken into account.....	19
2.1.5	Key metrics	19
2.1.6	Conclusion of the mobile data traffic evolution	20
2.2	Mobile usage scenarios	20
2.2.1	End user service scenarios	20
2.2.2	Network (operator) usage scenarios.....	20
2.2.3	Conclusions on Mobile usage scenarios.....	21
3.	Architecture Requirements	22
3.1	High-level requirements – user and operational aspects	23
3.2	Performance requirements.....	23
3.3	Network management	24
3.4	Mobility requirements	24
3.5	Scalability requirements	25
3.6	Reliability and Availability requirements	26
3.7	Security and privacy requirements	26
3.8	Charging Aspects.....	27
3.9	Energy efficiency	28
3.10	Traffic management.....	28
4.	Architecture Challenges	29
4.1	Network Topology related challenges	29
4.2	Mobility related challenges	30
4.3	Network Transport related challenges	31
4.4	Network Management related challenges	32
4.5	Traffic management related challenges	32

4.6	Network applications and services related challenges.....	34
4.6.1	M2M related challenges.....	34
4.6.2	Energy efficiency related challenges.....	34
4.6.3	Improved user experience and efficient resource usage.....	34
5.	Proposed Technology Solutions	35
5.1	Mobility.....	35
5.2	Network Transport.....	36
5.3	Traffic Management	37
5.4	Network Management	38
5.5	Network applications and services	39
5.5.1	Network functionality virtualization and realization with cloud computing.....	39
5.5.2	Network Energy Efficiency improvements by efficient capability utilization	39
5.5.3	Network efficiency improvement for Video/Multimedia Applications	40
5.5.4	Network improvement for M2M Applications.....	40
5.5.5	Application based network traffic analysis and engineering.....	40
5.6	Network Topology.....	40
6.	Architecture Approach	42
6.1	Topological models	42
6.1.1	Centralized architecture	42
6.1.2	Distributed architecture.....	43
6.1.3	Flat architecture.....	44
6.2	KPIs.....	44
6.2.1	Performance criteria	44
6.2.2	Deployment criteria.....	46
6.2.3	Validation/technology maturity criteria.....	46
6.3	Architecture options	46
6.3.1	Technologies	46
6.3.2	Architectures	53
6.3.3	Technology Coexistence analysis.....	55
6.3.4	Roaming	58
6.4	Other Architecture options	60
7	OPEX and CAPEX analysis.....	62
7.1	Model description.....	62
7.2	OPEX and CAPEX results	63
7.3	Sensitive Analysis	64
7.4	Conclusions	65
8	System validation	67
8.1	Validation Usage scenarios.....	67
8.2	Valiation results	67
8.3	Validation Conclusions.....	81
9	Conclusion.....	83

2 Table of figures

Figure 1 Problem statement of radio access and core network	16
Figure 2 Evolved Packet Core network	17
Figure 3 MEVICO possible network vision for improving video service efficiency	22
Figure 4 Network topology model.....	41
Figure 5 Centralized architecture model.....	43
Figure 6 Distributed architecture model	43
Figure 7 Flat architecture model.....	44
Figure 8 Centralized SON architecture.....	48
Figure 9 Distributed SON architecture	48
Figure 10 Hybrid SON architecture.....	49
Figure 11 Centralized architecture with selected technologies	53
Figure 12 Distributed architecture with selected technologies	54
Figure 13 Flat architecture with selected technologies.....	54
Figure 14 Roaming reference points.....	59
Figure 15. Centralized Architecture with all MEVICO technologies.	60
Figure 16. Distributed Architecture with all MEVICO technologies.....	61
Figure 17. Flat Architecture with all MEVICO technologies.....	61

Authors

Partner	Name	Phone / Fax / e-mail
AALTO	Jose Costa-Requena	Phone: +358 9 470 26257 e-mail: jose.costa@aalto.fi
France Telecom	Ivan Froger	Phone: +33 1 45 29 49 65 e-mail: ivan.froger@orange.com
France Telecom	Didier Becam	Phone: +33 2 96 05 32 02 e-mail: didier.becam@orange.com
France Telecom	Khadija Daoud Triki	Phone: +33 1 45 29 54 57 e-mail: khadija.daoud@orange.com
France Telecom	Philippe Herbelin	Phone: +33 1 45 29 45 87 e-mail: philippe.herbelin@orange.com
France Telecom	Jiejing Hou	Phone: +33 1 45 29 89 79 e-mail: jiejing.hou@orange.com
France Telecom	Benoit Lemoine	Phone: +33 2 96 05 04 99 e-mail: benoit.lemoine@orange.com
France Telecom	Pascal Pagnoux	Phone: +33 1 45 29 40 70 e-mail: pascal.pagnoux@orange.com
CEA	Michael Boc	Phone: +33 1 69 08 39 76 e-mail: Michael.Boc@cea.fr
CEA	Alexandru Petrescu	Phone: +33 1 69 08 92 23 e-mail: Alexandru.Petrescu@cea.fr

CEA	Mounir Kellil		
		Phone:	+33 1 69 08 22 74
		e-mail:	Mounir.Kellil@cea.fr

MIK	Zoltán Faigl		
		Phone:	+36 1 463 2499
		e-mail:	zfaigl@mik.bme.hu

MIK	László Bokor		
		Phone:	+36 1 463 2499
		e-mail:	bokorl@hit.bme.hu

Montimage	Edgardo Montes de Oca		
		Phone:	+33 5 380 3577
		e-mail:	edgardo.montesdeoca@montimage.com

Nokia Siemens Networks	Tero Lötjönen		
		Phone:	+358 40 5747842
		e-mail:	tero.lotjonen@nsn.com

Nokia Siemens Networks	Pekka Wainio		
		Phone:	+358 40 5811211
		e-mail:	pekka.wainio@nsn.com

Nokia Siemens Networks	Pekka Korja		
		Phone:	+358 40 7665979
		e-mail:	pekka.korja@nsn.com

Nokia Siemens Networks	Wolfgang Hahn		
		Phone:	+4989515924122
		e-mail:	wolfgang.hahn@nsn.com

Turk Telekom	Ahmet Serdar Tan		
		Phone:	+90 212 3099975
		e-mail:	ahmetserdar.tan@turktelekom.com.tr

AVEA	Engin ZEYDAN		
		Phone:	+90 216 987 6386
		e-mail:	engin.zeydan@avea.com.tr

AVEA	Çağatay EDEMEN		
		Phone:	+90 216 987 6386
		e-mail:	cagatay.edemen@avea.com.tr

Alcatel-Lucent	Jean-Luc Lafragette	Phone:	+33130772738
		e-mail:	jean-luc.lafragette@alcatel-lucent.com
Alcatel-Lucent	Erick Bizouarn	Phone:	+33130772724
		e-mail:	erick.bizouarn@alcatel-lucent.com
Ericsson AB	Rashmi Purushothama	Phone:	+46 10 715 5964
		e-mail:	rashmi.purushothama@ericsson.com
Ericsson AB	Jörgen Andersson	Phone:	+46 10 719 7013
		e-mail:	jorgen.andersson@ericsson.com
Ericsson AB	Conny Larsson	Phone:	+46 10 714 8458
		e-mail:	conny.larsson@ericsson.com
VTT	Suihko Tapio	Phone:	+358 09 451 1111
		e-mail:	Tapio.Suihko@vtt.fi
Artelys	Franck Mornet	Phone:	+33 1 44 77 89 01
		e-mail:	franck.mornet@artelys.com

3 Executive Summary

This document defines the Architecture Design Release 3 for the Celtic MEVICO project. The scope and the context of MEVICO project are summarized to recall the business drivers. An overview of Evolved Packet Core (EPC) architecture and the related network elements are described. The evolution of the network traffic and usage scenarios are the guidelines of MEVICO to generate a more efficient mobile architecture for the LTE (Long Term Evolution) and LTE-Advanced radio access systems of 3GPP.

The MEVICO architecture is based on requirements related to different aspects (usage and operational, performance, network management, mobility, scalability, reliability, availability, security, charging, energy efficiency, traffic management).

These requirements allow identifying Architecture Challenges related to the different topics (network topology, mobility, network transport and management, traffic management, applications and services).

This release of document includes technology solutions selected based on the relevant KPIs. The technologies are mapped into different architecture topologies. This release contains the coexistence analysis of the technology solutions, but not yet the further analysis which leads to the final architecture evolution recommendations.

4 List of terms, acronyms and abbreviations

Generally the 3GPP used terms are used in this document [1].

Clarification of used terms in the document

Access Point Name	In 3GPP, Access Point Name (APN) is a reference to the Gateway GPRS Support Node (GGSN) or Packet Data Network Gateway (P-GW) to be used. In addition, Access Point Name may, in the GGSN or P-GW, identify the packet data network and optionally a service to be offered [2]
Application agnostic group communications	The group communications will include a variety of multimedia application types so the solution that enables the group communications shall be application-agnostic.
Busy Hour	In a communications system, the sliding 60-minute period during which occurs the maximum total traffic load in a given 24-hour period.
Connected subscription	A subscription that has one IP address assigned to enable always-on feature.
Device	A physical entity with communications interface that requires an active subscription to networking infrastructure to establish a connection. There is an endless list of devices e.g. smart phones and other mobile phones, laptops with USB dongle or integrated wireless interfaces, vehicular network with several multimedia devices, home network with sensors, actuators, home devices such as picture frame, Video-on-Demand players, Home GWs, etc., vehicular devices such as in-car multimedia player, game console, etc., other devices associated to the user such as personal sensors, body network, etc.
Dynamic resource allocation	The network shall dynamically reconfigure providing additional bandwidth to traffic demands.
Fixed broadband data connection	Wireline connection enabling speed >1Mbps per user.
Hyperconnectivity	Use of multiple means of communication, such as email, instant messaging, telephone, face-to-face contact and Web 2.0 information services. Also a trend in computer networking in which all things that can or should communicate through the network will communicate through the network.
Offloading	The traffic offloading in this document means routing away the traffic originating from the EPS/mobile network/mobile device onto some other network such as WLAN.
macroscopic traffic management	It includes all mechanisms with the primary objective to improve efficient usage of network resources. Parameters for optimization describe traffic patterns without detailed knowledge of individual flow attributes.
microscopic traffic management	It is associated with all mechanisms with the primary objective to improve performance of individual flows based on application type, user profile and other policy related information.
Mobile broadband data connection	Wireless connection enabling speed >256kbps per user and wide user mobility. Technologies include CDMA2000 EV-DO, WCDMA/HSPA, LTE, Mobile WiMAX, and TD-SCDMA.
Mobility type support	Host mobility (a host changes its point-of-attachment), user mobility (user moves from one host to another) and session mobility (old session is restored when the user moves to a new host) shall all be supported e.g. via aggregation of mobility protocols or a single protocol.
Moving network	The network and its subsequent mobility protocol(s) must support network mobility i.e. moving networks such as bus, cars, aircraft, PAN, etc.
Multi-homed Devices	Terminals with several interfaces up that allow mobility between any IP address currently bound to the device. Multihoming is a technique that allows to be connected to several networks; it can be used to avoid the single point of failure for the network connectivity. Most of the time, the implementation is realized through use of multiple interfaces

Provider Edge	Provider Edge devices is standard layer 2 (L2) Ethernet, which is paring with the Customer Edge (CE) through a User-Network Interface (UNI)
Scalability	Scalability in a network is the ability to adapt to a change of order of magnitude of the demand. So it is its ability to increase its capacity while maintaining its features and performances. Not to make the confusion with congestion issues. A congested network might be scalable but just needing a capacity upgrade via existing equipments upgrade or new equipments integration. On the other hand, a network we can not upgrade anymore without loosing performances or revenues is not scalable.
Small cell	Small Cells are low-powered in-building or outdoor radio access nodes that operate in licensed and unlicensed spectrum that have a range from 10 meter upto few kilometers. Types of small cells include micro, pico and femto cells, distributed radio systems with remote radio heads and Wi-Fi hotspots. Small cells are used by mobile operators to extend the wireless service coverage and/or increase network capacity, both indoors and outdoors
Subscriber	A Subscriber is an entity (associated with one or more users) that is engaged in a Subscription with a service provider. The subscriber is allowed to subscribe and unsubscribe services, to register a user or a list of users authorized to enjoy these services, and also to set the limits relative to the use that associated users make of these services. [1]
Subscription	A subscription describes the commercial relationship between the subscriber and the service provider. [1]
Network topology	Network topology represents the layout of the interconnection between network elements e.g. routers, switches or other communication elements.
User	End user, an entity, not part of the (3GPP) System, which uses (3GPP) System services.[1]
User Equipment (UE)	In 3GPP System, allows a user access to network services. A User Equipment can be subdivided into a number of domains, the domains being separated by reference points. Currently the User Equipment is subdivided into the UICC (Universal Integrated Circuit Card) domain and the ME (Mobile Equipment) Domain. The ME Domain can further be subdivided into one or more Mobile Termination (MT) and Terminal Equipment (TE) components showing the connectivity between multiple functional groups [1]. In this document UE and Mobile Device are used parallel.
vertical handovers	Vertical handover is a handover between two different radio access technologies that do not share the same radio infrastructure. For example, a handover between 3G and Wi-Fi is a vertical handover, but a handover between GPRS and HSDPA is not a vertical handover, it remains a horizontal handover. Usual handover with the same radio access technology is horizontal handover.

List of abbreviations

3GPP	3rd Generation Partnership Project, based on GSM Technology
AAA	Authentication, Authorization and Accounting
AKA	Authentication and Key Agreement
ALTO	Application Layer Transport Optimization
ANDSF	Access Network Discovery and Selection Function
AP	Access Point
APN	Access Point Name
ARP	Allocation and Retention Priority
ARQ	Automatic Repeat-reQuest
BAT	Bulk Analysis Tool
BER	Bit Error Rate
BS	Base station
BTS	Base Transceiver Station
CAPEX	Capital Expenditure
CBS	Committed Burst Size
CDN	Content Delivery Network
CES	Customer Edge Switching
CET	Carrier-Ethernet Transport
CIR	Committed Information Rate
CMIP	Common Management Information Protocol
CN	Core Network
CoMP	Coordinated Multi-Point
CSCF	Call Session Control Function
DDMM	Distributed and Dynamic Mobility Management
DHCP	Dynamic Host Configuration Protocol
DHT	Distributed Hash Table
DL	Downlink
DMA	Distributed Mobility Anchoring
DNS	Domain Name Server
DPI	Deep Packet Inspection
DWDM	Dense Wavelength Division Multiplexing
E2E	End-to-end
EAP-SIM	Extensible Authentication Protocol - Subscriber Identification Module
EBS	Excess Burst Size
EIR	Excess Information Rate
eNB	Evolved Node B (eNodeB)
EPC	Evolved Packet Core
ePDG	Evolved Packet Data Gateway (ePDG)
EPS	Evolved Packet System
ETSI	European Telecommunications Standards Institute
E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
EVC	Ethernet Virtual Connection
FI	Future Internet
FTTA	Fiber To The Antenna
Gbps	Giga Bit Per Second
GBR	Guaranteed Bit Rate
GGSN	Gateway GPRS Support Node

GHz	Giga Hertz
GPRS	General Packet Radio Service
GTP	GPRS Tunnelling Protocol
GUTI	Globally Unique Temporary ID
GW	Gateway
HeNB	Home eNB
HetNet	Heterogeneous Network
HIP	Host Identity Protocol
DEX	Diet Exchange (HIP DEX AKA)
HNP	Home Network Prefix
HO	Handover
HSPA	High-Speed Packet Access
HSPA+	Evolved HSPA (3GPP release 7, including I-HSPA)
HSS	Home Subscriber Server
HTTP	Hypertext Transfer Protocol
HW	HardWare
I-CSCF	Interrogating-CSCF
ID	Identifier
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IFOM	IP Flow Mobility
I-HSPA	Internet HSPA
IKEv2	Internet Key Exchange, version 2
IM	Instant Messaging
IMS	IP Multimedia Subsystem
IMT	International Mobile Telecommunications
IMT-A	IMT Advanced
IP	Internet Protocol
IPsec	Internet Protocol Security
IS	Intermediate System
IS-IS	Intermediate System to Intermediate System
ISP	Internet Service Provider
IT	Information Technology
ITU	International Telecommunication Union
KPI	Key Performance Indicator
L2	Layer 2
L3	Layer 3
LAN	Local Area Network
LFN	Local Fixed Node
LSP	Label-Switched Path
LTE	Long Term Evolution
LTE-A	LTE Advanced
LMA	Local Mobility Anchor
LU	Location Update
M2M	Machine-to-Machine
MAC	Media Access Control, a low layer protocol
MAG	Mobile Access Gateway
MASE	Media Aware Serving Entity

MBH	Mobile BackHaul
MBR	Maximum Bit Rate
MCCS	Multi-Criteria Cell Selection
MIH	Media Independent Handover
MIMO	Multiple Input Multiple Output
MIP	Mobile IP
MLB	Mobility Load Balancing
MME	Mobility Management Entity
MN	Mobile Node
MNO	Mobile Network Operator
MTM	Microscopic Traffic Management
mP4P	Mobile P4P
MPLS	Multi-Protocol Label Switching
MPLS-TP	MPLS Transport Profile
MPTCP	Multi-Path TCP
MPTCP-Pr	MultiPath TCP - Proxy
MR	Mobile Router
MRO	Mobility Robustness Optimization
mRVS	mobile Rendezvous Server
m-SCTP	mobile-SCTP
MSO	Multimedia Streaming Optimizations
MTC	Machine-Type Communications
NB	Node B
NB-IFOM	Network-based IP Flow Mobility
NEMO	Network Mobility
NETCONF	Network Configuration Protocol
NG	Next Generation
NIMTC	Network Improvements for Machine-Type Communications
NMIP	Not Mobile IP
NW	Network
O&M	Operations & Maintenance
OC	Optical Carrier
OPEX	Operational Expenditure
OTT	Over The Top
P2P	Peer-to-Peer
P4P	Proactive Network Provider Participation for P2P
PBB-TE	Provider Backbone Bridge Traffic Engineering
PBM	Policy-based Management
PCC	Policy and Charging Control
PCRF	Policy Control and Charging rules function
PE	Provider Edge
PDN	Packet Data Network
P-GW	Packet Data Network (PDN) Gateway
PMIP	Proxy Mobile IP
PMIP-RO	Proxy Mobile IP – Route Optimisation
PoP	Point of Presence
PPP	Point-to-Point Protocol
PS	Packet Switched

QCI	QoS class identifier
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RFC	Request For Comments
RGW	Residential Gate Way
RLF	Radio Link Failure
RNC	Radio Network Controller
ROF	Radio Over Fiber
RPC	Remote Procedure Calls
SA	Security Association
SAE	System Architecture Evolution, LTE's core network architecture
SAIL	Scalable and Adaptive Internet Solutions
SCTP	Stream Control Transmission Protocol
SDH	Synchronous Digital Hierarchy
SGSN	Serving GPRS Support Node
S-GW	Serving Gateway
SIM	Subscriber Identity Module
SIMTC	System Improvements to Machine-Type Communications
SIP	Session Initiation Protocol
SNMP	Simple Network Management Protocol
SNR	Signal-to-Noise Ratio
SON	Self Organizing Network
SW	Software
TCP	Transmission Control Protocol
TDD	Time-Division Duplex
TEHO	Traffic Engineered Handovers
THP	Traffic Handling Priority
TM	Traffic Management
TRILL	Transparent Interconnection of Lots of Links
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
USB	Universal Serial Bus
VLAN	Virtual Local Area Network
VoD	Video-on-Demand
VoIP	Voice over IP
VPLS	Virtual Private LAN Service
VPN	Virtual Private Network
WDM	Wavelength-Division Multiplexing
Wi-Fi	"Wireless Fidelity" a trademark of Wi-Fi Alliance (IEEE 802.11 certified devices)
WiMAX	Worldwide Interoperability for Microwave Access (IEEE 802.16 standard)
WLAN	Wireless Local Area Network
WMN	Wireless Mesh Network
XML	Extensible Markup Language

5 References

3GPP standards

- [1] 3GPP, “Vocabulary for 3GPP Specifications”, TR 21.905, release 10.
- [2] 3GPP, “General Packet Radio Service (GPRS); Service description; TS 23.060, release 10.
- [3] 3GPP, “Network architecture”; TS 23.002, release 10.
- [4] 3GPP, “General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access”; TS 23.401, release 10.
- [5] 3GPP, “Architecture enhancements for non-3GPP accesses”; TS 23.402, release 10.

Publications / IETF standards

- [6] European Commission: A Digital Agenda for Europe (May 2010):
http://ec.europa.eu/information_society/digital-agenda/documents/digital-agenda-communication-en.pdf.
Digital Agenda web page: http://ec.europa.eu/information_society/digital-agenda/index_en.htm
- [7] Cisco Visual Networking Index: Hyperconnectivity and the Approaching Zettabyte Era (June 2010):
http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/VNI_Hyperconnectivity_WP.html
- [8] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011–2016:
http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html
- [9] EARTH (Energy Aware Radio and neTwork Technologies) Deliverables: <https://www.ict-earth.eu/publications/deliverables/deliverables.html>
- [10] GSMA Association: IR.88. LTE Roaming Guidelines. Version 7.0. 31 January 2012.
- [11] SAIL (Scalable and Adaptive Internet Solution) web page: <https://www.sail-project.eu>
- [12] Data Centers to Finland (DC2F): <http://dc2f.comnet.tkk.fi/>

1. Introduction

This document defines the Architecture Design Release 1 Documentation for the Celtic MEVICO project. This is the D1.3 Report, part of the Work Package 1. This study is conducted in 2010, 2011 and 2012. The content is contributed by the whole MEVICO project consortium. The document is constructed in a logical progression depicted below.

Section 1 introduces the scope and the context of MEVICO project. Section 2 describes the future mobile network traffic and usage scenarios that MEVICO architecture shall support. Section 3 identifies the architecture requirements based on the trends and scenarios studied before. Section 4 describes the architecture challenges. Based on these challenges, Section 5 proposes technology solutions. The choice of MEVICO architecture approaches is described in Section 6.

1.1 Business drivers for the MEVICO project

Affordable, truly accessible mobile broadband has matured with HSPA (High-Speed Packet Access), HSPA+ (3GPP release 7, including I-HSPA), and LTE (Long Term Evolution). It has blurred boundaries between mobile/fixed and voice/data for end-users, operators and application developers.

Mobile data traffic is expected to grow faster than the fixed Internet for the coming years and with the same rate as fixed Internet in the long term. Radio access and core network must be scaled to accommodate the expected traffic growth, especially if we consider limited revenue growth. It will lead to access and core networks cost pressure.

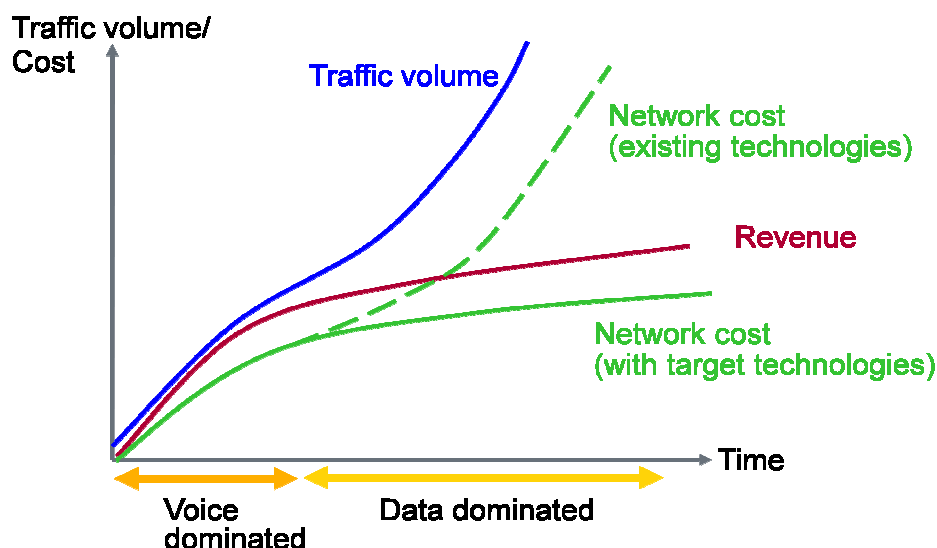


Figure 1 Problem statement of radio access and core network

The operators have to satisfy the demands of the new services and data volume growth, in order to remain competitive. New business models are required and redefining business priorities might also impact the selection of network infrastructure.

1.2 Overview of Evolved Packet Core (EPC) architecture

In 3GPP release 8, LTE and SAE (System Architecture Evolution) work resulted in the specification of the E-UTRAN (Evolved UTRAN) and in the specification of Evolved Packet Core (EPC); both components form the EPS (Evolved Packet System). LTE-EPC is the name for the long term evolution of UMTS.

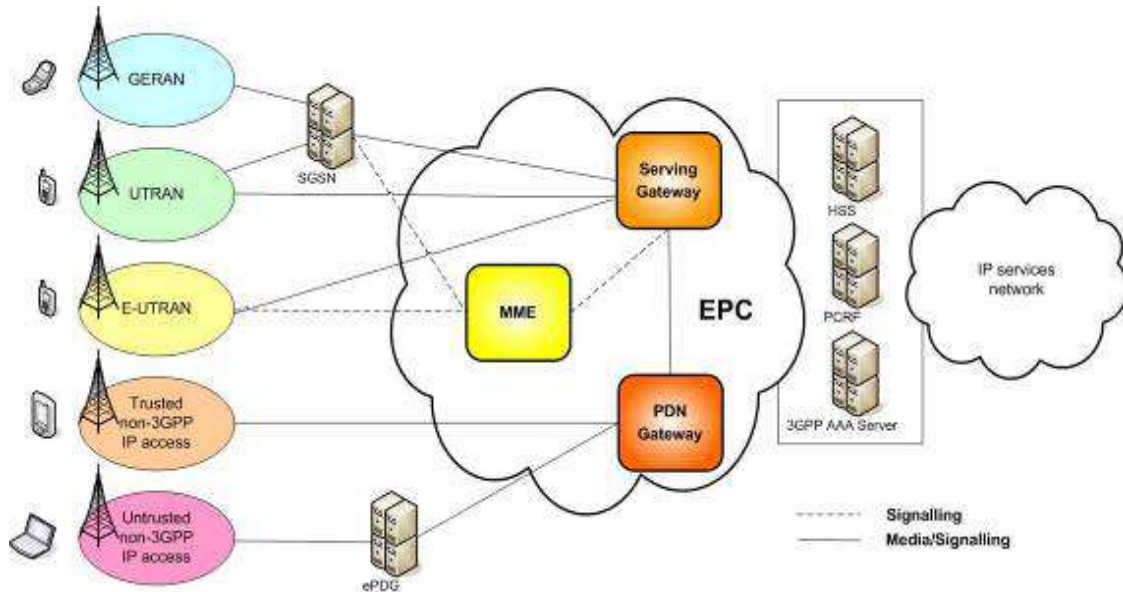


Figure 2 Evolved Packet Core network

The Evolved Packet Core is made of three main network entities, described in [3], [4], and [5]. The user plane consists of two types of nodes, the **Serving Gateway** (S-GW) and the **PDN Gateway** (P-GW). The control plane is made up of a separate **Mobility-Management Entity** (MME) :

- The **MME** manages all the signalling (control plane):
- The **S-GW** terminates the user plane interface towards E-UTRAN:
- The **P-GW** terminates the user plane interface towards one or more Packet Data Networks:

1.3 Other network elements linked to the Evolved Packet Core (EPC)

The other network elements linked to EPC are the following:

- Legacy 3GPP access: Serving GPRS Support Node (SGSN),
- Non-3GPP access: Evolved Packet Data Gateway (ePDG), 3GPP AAA server;
- Evolved UTRAN (E-UTRAN);
- Home eNodeB;
- Policy and Charging Control architecture.

2. Network Traffic and Usage Scenarios

2.1 Mobile traffic, service, and technology evolution 2008-2020

This section collects traffic analyses from the last couple of years and provides some traffic forecasts up to 2020. The evolution scenarios for the growth of traffic volumes and the number of users as well as the impacting application, service and technology evolution scenarios are covered.

2.1.1 Traffic Data evolution

The mobile broadband subscriber and traffic volume increase is inevitable and the future network architecture has to be designed to cope with it. The mobile traffic global increase is a consequence of several factors: growth of the mobile subscriptions (e.g. growth of population, improving living standards), evolution of the mobile networks/ devices and services (e.g. affordability of capable devices, enabled connection speeds, low cost flat rate data plans, easier usage, evolution of communication needs). And there is a huge increase potential of devices/subscriptions/traffic with the Machine-to-Machine (M2M) communications.

The network needs to be optimized to maximize the end-user mobile broadband experience, minimize the mobile device battery consumption and ensure efficient, congestion-free network performance. Because the available mobile network frequency bands are scarce and the utilized spectral efficiencies are tending to the theoretical limit, several other methods to cope with the increasing capacity demand need to be utilized. Even though the regulation is planning to open new Digital Dividend frequencies in the coming years, this alone will not be able to totally solve the problem.

Some regulatory or public funding drivers can have an additional impact on the operator interest to invest to expand the network capacity. There are some guidelines drawn in the European Commission Vision 2020 related to Digital Agenda work [6], for example, guidelines defining the minimum connection speed targets for broadband Internet.

2.1.2 Services and application evolution

The most remarkable mobile user application challenges in the future are expected to come from video, social networking and M2M types of services, which exponentially will increase the traffic volume.

Video: The sum of all forms of video (including Internet TV, Video on Demand, interactive video, and Peer-to-Peer (P2P) video streaming, mobile 3DTV, etc.) will account for close to 90 percent of consumer traffic (fixed and mobile) by 2012 [7]. The evolution of the Content Delivery Networking (CDN) and intelligent data caching technologies in the fixed network side might have impact on the mobile network architecture, mainly by bringing the content lower in the network and enable efficient usage of several parallel flows from different content sources.

Social networking: Consumers are more and more using a variety of services to communicate (e.g. email, instant messaging, twitter, Facebook, video, VoIP, and a host of other social networking applications) that use a mix of voice, video and messaging.

M2M: M2M communications have enormous potential (tens of billions of devices to be connected) to become the leading traffic contributor. These types of services will also generate different traffic time variations than those due to human activity (non-busy hours, strict latency requirements, initialization/synchronization after recovering from a network failure). Note that M2M devices might have longer life cycles than the ones of handsets. This could be a factor limiting new technological advances, replacing them to reclaim spectrum can be infeasible.

Mobile Gaming: As the handheld devices are equipped with better hardware, online mobile gaming traffic is expected to become a significant traffic contributor. Maintaining game stability among several mobile users necessitates the transmission of state updates between each mobile device with low latency.

To efficiently cope with the challenges related to these services, there is a need to consider the mobile network architecture optimization to allow efficient use in heterogeneous network environments and understand the impact of CDN technology evolution.

2.1.3 Evolution-enabling technologies

The main evolutions are related to:

- Bandwidth needs in radio technologies become similar to fixed network;

- Miniaturization of radio technologies facilitates the deployment of devices with only basic connectivity functionality increasing the growth of machine to machine communications; Evolution of processing and networking technologies: there is no clear indication whether in the future the applications will be used on the end device or on the network.

2.1.4 Aspects to be taken into account

The following aspects should be taken into account:

- Flat rate pricing in mobile broadband networks have stimulated many users to change their fixed broadband access to mobile.
- Users should obtain similar bandwidth capacity regardless whether the underneath technology is wireless or wire line.
- Traffic growth in mobile broadband networks is mainly due to the evolution of the mobile networks, devices and services. And to a certain extent it is due to a smooth migration of users from fixed broadband networks.
- New traffic patterns and exponential traffic increase originated from new devices that incorporate mobile broadband connectivity (e.g. sensors, home appliances).
- The evolution of User Interface and ways of interacting with the mobile devices will open up the demand for new applications which require higher bandwidth and generate more traffic.
- Traffic balance in uplink/downlink in the future. Currently mobile devices are mostly consuming content, but there are signs that real-time sharing (e.g. livecast video) can put significant traffic demand for uplink.
- The new service levels enabled by the hyperconnectivity will place huge capacity demands on the networks. The four key growing enablers of hyperconnectivity are: (a) the growing penetration of high-speed broadband, (b) the expansion of digital screen surface area and resolution, (c) the proliferation of network-enabled devices, and (d) the increases in the power and speed of computing devices.
- Context-aware mobile computing, in which applications can discover and take advantage of contextual information (such as user location, time of day, nearby people and devices, and user activity), can introduce new challenges to system infrastructure.
- All available capacity will be exploited, with affordable pricing. In mobile networks, different charging models (with respect to fixed broadband) shall/could be exploited, in order to share the limited radio access capacity, since flat rate alone should not be the most suitable model.
- Net neutrality has to be respected in the service delivery and quality.

2.1.5 Key metrics

There are general challenges in the mobile network future trends identified in the studies carried out in MEVICO project: Increase of subscriber amounts (with huge potential of M2M), high increase of the data amounts (driven by video delivery), always on applications, availability of heterogeneous network and multiple types of interfaces in User Equipments (UEs).

The results of the preliminary studies¹ done in MEVICO project show that there are several drivers that will cause network scalability and optimization challenges. The current architecture needs to be scaled according to the growth of:

- data traffic volume per user by about 3-10 times by year 2020 compared to 2010
- number of mobile broadband subscriptions and end users increase by 8-12 times by year 2020 compared to 2010, according to our internal traffic forecasts.
 - when including the M2M devices even to 50 times
- mobility rate (users changing their location during the active broadband communication) will remain around 20-25%, so most of the mobile broadband usage takes place in stationary location.
- number of network nodes, due to densification/frequency overlay/small cell needs, heterogeneous networks

¹ Analysis of the traffic evolution over the past years based on statistics and data collected from different sources. This information has been combined, together with forecasts from research studies and public reports, in order to provide the estimated traffic growth up to year 2020. The main references are Ericsson and NSN internal materials in addition to [8].

- network signalling load. Even mobile core network signalling load share of the total traffic amount is estimated to increase moderately from 2% to 3% from 2010 to 2020, but with the estimated total traffic, subscription and network node increase the signalling will increase considerably

Therefore, the future mobile architecture should deal among others with the challenges associated with the increase of traffic, mobility and signalling traffic while keeping the OPEX under competitive levels for operators.

2.1.6 Conclusion of the mobile data traffic evolution

Mobile broadband usage has taken off in the last couple of years due to several factors such as improved network capabilities, affordable data plans and the evolution of end-user devices. The main drivers for the future mobile broadband traffic growth are the increase of global subscribers number, the evolution of the user devices that enable easy usage of data hungry services and the evolution of the network functionality that enables operators to provide high speed mobile services with attractive pricing. It is foreseen that these factors, in conjunction with the evolution of the applications in new areas, will tempt more users to utilize new devices and consume more data. Most of the data hungry applications are related to entertainment content, like video streaming, social networking, mobile gaming, and thus the enthusiasm to use them depends largely on the service costs. The end result is that mobile broadband traffic volume will increase in the future and the network architecture evolution has to be optimized to cope with it.

2.2 Mobile usage scenarios

The target of this section is to identify the trends, the new technologies and drivers having an impact on mobile core network architecture. Scenarios bringing new requirements to mobile networks in terms of latency, mobility, traffic management, etc. have been identified.

2.2.1 End user service scenarios

Fixed-Mobile Convergence

The Fixed-Mobile convergence section addresses usage scenarios where there is no expected QoE difference for the end user on whether the communications are done over fixed or mobile networks. The following three use cases listed are already defined in 3GPP.

- Internet access with Parental control and personal firewall,
- Voice/Multimedia and Charging,
- Video.

Another use case is mass delivery of real-time multimedia content which has specific requirements.

M2M communication and wireless sensor network scenarios

The machine to machine scenarios include the followings:

- Remote healthcare
- Smart metering / industrial monitoring
- Mass monitoring, mass remote control, Tracking objects
- Automotive connectivity traffic scenarios
- Internet of things and future Mobile networks

2.2.2 Network (operator) usage scenarios

The Network (operator) usage scenarios include the following:

- Energy saving improvements
- Virtualization and Cloud computing
- Seamless user experience of mobile Internet over multiple data GWs and multiple interfaces
- Small cell deployment
- Secured access by design to limit unwanted traffic to mobile clients
- Mass event coverage and capacity enabling with wireless mesh transport
- Automatic and Secure Layer 2 Virtual Private Networks

2.2.3 Conclusions on Mobile usage scenarios

The traffic analysis based on aggregation alone is not enough in the future application contexts. Per user and per application analysis is needed for a better understanding and optimization of traffic. Detailed knowledge of traffic patterns, including packet size and time intervals, are needed to improve resource allocations and obtain the required end-user's QoE. In order to manage the increased traffic and new applications with new requirements, LTE-EPC technologies have adopted an all-IP architecture that integrates a more distributed management and QoS strategy. This architecture simplifies the network stack, but makes the management more complex.

3. Architecture Requirements

The architecture in MEVICO focuses on the evolution of the mobile packet core network for the LTE (Long Term Evolution) and LTE-Advanced of 3GPP. MEVICO will study and define system concepts to enhance the Evolved Packet Core (EPC) of 3GPP in the mid-term in 2011-2014 towards the requirements that are challenging the packet connectivity capabilities.

The project will focus on the network aspects to complement the research and standardization (3GPP) already ongoing for defining and standardizing a new radio system LTE-Advanced as the next step of the LTE radio technology in 2010. The project will not address the radio interface aspects, but will rather enhance the network architecture, higher bit rates and higher capacity. Nevertheless, the peculiarities and limitations of the radio portions are reflected into the core network and those impacts will be therefore addressed in the project.

As an example of requirements, we will focus on the following illustrative and challenging video services to show that the architecture covers all the spectrum of potential services, namely Internet TV, VoD, Personal Broadcasting and Interactive Video. The MEVICO network will exploit heterogeneous wireless access to deliver media content to mobile customers. MEVICO will focus on LTE Advanced and Wireless Local Area Network (WLAN) access co-operation as depicted in the figure below:

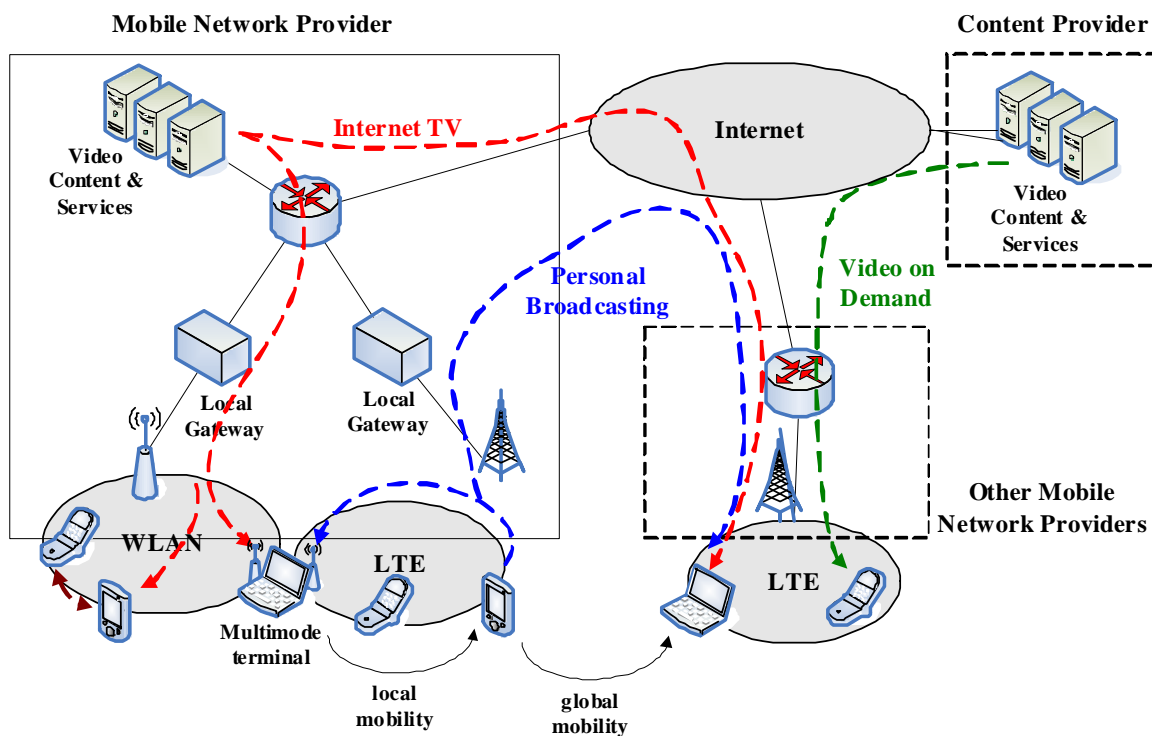


Figure 3 MEVICO possible network vision for improving video service efficiency

The main requirements for the future mobile networks are the following:

- Enabling the efficient use of the heterogeneous network capabilities, like multi-access (several simultaneous parallel paths, fixed-mobile, convergence) and multimode (several overlapping alternative Radio Access Technologies (RATs)).
 - This requires an efficient and optimized way of selecting/utilizing the multiple available paths, because, until now, the mobile device is not able to be active simultaneously on all RATs.
 - Use of multiple interfaces brings new challenges in different functions: Identity Management, security/privacy-preserving methods, charging, lawful interception, etc.
- End user Quality of Experience is the key driver for architecture evolution.
 - Architecture related capacity bottlenecks shall be avoided (i.e., the scalability has to be ensured).
 - Latency and throughput need to be kept optimal when traffic load is high.

- “Always On” applications need to be handled optimally, without causing extensive load to network signaling.
- Implications from the new, very high capacity radio access network topologies – like LTE and Wi-Fi 802.11n - shall be taken into account.
- Cost optimization needs to be addressed with care since the operator’s revenue increase will be modest due to the widespread flat rate model...
- Traffic optimizing concepts under study for the Future Internet, e.g. to access the content cached near to the user, Information Centric Networking/Content Delivery Networking concepts, shall be studied to understand their impact for optimizing the mobile network architecture.

In the subsequent sections, we will list all the requirements. The requirement numbering stems from MEVICO-internal requirement structuring. Only the requirements with very high and high priority are taken into account:

- A refers to Architectural requirement
- F refers to Functional requirement
- N refers to Non-functional requirement

3.1 High-level requirements – user and operational aspects

The requirements in this section shall secure that the network provides mobility functionality within and across the different access systems.

A 5.1.1: Heterogeneous transport technologies

The packet transport service architecture must support heterogeneous transport technologies at all levels of network hierarchy. Multiple technologies - some of those should provide long-term compatibility for M2M where UE usable life time can be significantly longer than mobile phone or smart phone.

A 5.1.2: Multiple operational domains

i.e. segment, technology and operator domains.

A 5.1.4: Topology diversities for network architectures

Network architecture must allow EPC functions distribution to be close to the devices from the access point and the possibilities of different topology architectures must be from fully distributed (close to the end user) to centralized.

A 5.3.1: QoS guarantee

In the network it must be possible to guarantee QoS to support e.g. Voice/Multimedia and Video according to service level agreement policies.

A 5.5.3: Support for Disparate Wireless Technologies

The networks must support roaming over heterogeneous access systems, systems built upon e.g. inter-joined cellular, WLAN, Bluetooth, and satellite networks, within one operator domain

A 5.6.1: IPv4 / IPv6 Cross-Family Communication

The network must support interoperation between IPv4 and IPv6 i.e. support cross-family communication.

3.2 Performance requirements

Network performance needs to satisfy the demands of the new services and data volume growth, if operators want to remain competitive.

A 6.1: Optimized architecture for content delivery

The mobile network architecture shall be optimized for content delivery methods.

A 6.4: Low latency

The network shall provide low latency to enable real time network functions. This is specified in 3GPP as “The maximum delay should be comparable to that for fixed broadband Internet access technologies”.

A 6.7: Synchronization

The network shall provide Clock synchronization signal transport over packet network to enable accurate synchronization of mobile Network Elements (NE).

A 6.8: Mobility type support

Host mobility (a host changes its point-of-attachment to the network), user mobility (user moves from one host to another), and session mobility (old session is restored when the user moves to a new host) shall be supported. Session mobility (old session is restored when the user moves to a new host) shall be supported e.g. via an aggregation of mobility protocols or a single protocol.

A 6.9: Device characteristics

The network shall allow taking advantage of M2M devices with pre-defined characteristics like Mobility; fixed device, devices with low or high mobility; and Traffic profile.

F 6.1: Small cell signalling optimization

It shall be possible to minimize the signalling load caused by deployment of small coordinated cells in the network architecture.

F 6.6: Multi-elements connectivity

Transport service shall support efficient low-latency partial mesh connectivity between Mobile Network elements.

N 6.1: Dynamic resource allocation

The network shall dynamically reconfigure providing additional bandwidth to large short-lived traffic demands.

3.3 Network management

Network management deals with operation, administration, maintenance, and provisioning of the network. On top of these normal tasks special attention should be paid to the large number of new device types attached to the network and energy efficiency aspects in the core network although the largest potential here is in the access networks.

F 7.7: Flexible network operation

The management solutions shall enable and support the network to adapt to the changing network usage. E.g. flexible bandwidth allocation would make it possible to adapt the network resource usage efficiently to the varying traffic load.

F 7.8: Efficient network monitoring

The monitoring solutions shall provide accurate measurements for (self-organizing) management solutions. The tradeoff between accuracy and monitoring traffic shall be maintained.

3.4 Mobility requirements

Mobile-Fixed network mobility, Multi-radio (LTE, HSPA, Wi-Fi) and Multi-layer (Macro, Micro, Pico, Femto, multifrequency) support in combination with the traffic growth adds complexity to the mobility functionality and this should be reflected in the mobility requirements.

A 8.1: Seamless Handovers

Network handovers between different accesses shall be, when required, fast enough to support the applications without change in service capability, security or QoS.

A 8.2: Optimized mobility protocols

Mobility protocols shall bring optimized routing and minimize data overhead.

A 8.3: Protocol interoperability

The mobility and multi-homing protocols shall ensure interoperability with IP based protocols.

A 8.5: Context Transfer

Networks shall provide transfer of session parameters to the new roamed network without interruption in service and re-initialization of the session parameters e.g. QoS, Security.

A 8.7: Selection of Mobility protocols

Several mobility protocols might co-exist. It shall be possible to select the mobility protocol for one traffic flow, to dynamically select and activate a mobility anchor; to dynamically configure the anchor selection criteria.

A 8.9: Small cell mobility

The network architecture shall be able to support UE mobility and service continuity between small coordinated cells (such as picocell), and between small coordinated cells and large coordinated cells.

A 8.11: Mobility between heterogeneous radio technologies

Mobility between WLAN and LTE wide area must be supported (where WLAN could be considered as uncoordinated cells).

A 8.12: Support of moving networks

The network and its subsequent mobility protocol(s) must support network mobility i.e. moving networks such as bus, cars, aircraft, etc.

3.5 Scalability requirements

It is important to have a scalable solution to be able to take care of the different traffic growth scenarios.

A 9.1: Small cell support

The network architecture must be able to support a large number of small coordinated and uncoordinated cells.

A 9.4: Signalling scalability

The signalling traffic induced by the mobility protocols shall be independent of the number of traffic flows per user.

A 9.5: Robust network

The network must be robust, optimized and designed to handle future mobile data bandwidth consumption and growth, driven by QoS-aware services.

A 9.6 Device addressing

The network must be able to uniquely identify a huge number of devices, i.e. this may require a new addressing method.

N 9.3: Scalability of management solutions

The number of nodes in the LTE-EPC is very high and the management solutions shall operate so as to enable scalable operation.

3.6 Reliability and Availability requirements

The increasing possibilities to connect people, things, etc, to distribute functions e.g. cloud computing, add more and more services that rely on the network and thus put increasing requirements on the network reliability in a broad sense. The new architecture should meet the needs of these new ways of using the network.

A 10.4: Application agnostic group communications

The group communications will include a variety of multimedia application types so the solution that enables the group communications shall be application-agnostic.

A 10.6: Interfaces availability information

There shall be a mechanism, either mobile or network originated, to find available interfaces.

A 10.9: Support for Multi-homed Devices

The mobility protocol shall address multi-mode terminals (i.e. terminals with several interfaces up) and allow mobility between any IP address currently bound to the device.

A 10.11: Routing loops avoidance

The architecture shall prevent routing loops in case single routers and terminals have multiple attachment points to the network.

F 10.2: IP flows routing

Packet content determines the associated flow and the network should enable routing based on IP flow.

3.7 Security and privacy requirements

The emergence of the new networks comprising converging technologies, different access technologies and environments mixed of computation and communication, requires new and strong security solutions (including privacy, authentication, need for encryption,...)

A 11.2: Protection against cyber-attacks

The system must provide a protection mechanism to mitigate various types of cyber-attacks: Denial-of-Service (DoS), Man-in-the-Middle (MitM), IP address spoofing, replay and redirection attacks, and identity theft of a host.

A 11.3: Strong Mutual Authentication

The communicating hosts (i.e. terminals and network nodes) shall be mutually authenticated as belonging to and allowed to join the network by a trusted third party.

A 11.5: Address Ownership

The system shall verify address ownership of each newly claimed address before using it, to prevent from possible address stealing and redirection attacks.

A 11.9: Location and Identity Privacy

Third parties must not be able to keep track of a host to know its present location and past history of activities. Nor must the identity be revealed to possible listeners of network traffic. This requirement must have exceptions as described in A 11.10

A 11.10: Lawful Interception

The system must take into account a possibility of required legal interception of traffic in the network.

A 11.15: Network isolation

Transport service shall assume that there will be complete isolation between client and transport networks + between different client networks.

A 11.17: Node identity

Identity of a device, and therefore its authentication shall be based on globally unique identifier.

A 11.18: Control plane security

The transport service shall be able to protect the integrity and confidentiality of control plane traffic.

A 11.19: Secure Zone-Based Authorization

The network shall authorize access to the users based on preset secure zone definitions and their access policy rules.

A 11.20: User profile based secured zones

The architecture shall allow definition of security zones where users are granted access based on their profile information.

F 11.1: Ensure Network neutrality

Network neutrality according to country specific legal requirements shall be ensured. The network should be flexible to fulfil specific legal requirements.

F 11.5: Device disconnection

Allow M2M disconnection of devices when tampering, theft or fraud detected.

F 11.7: Unwanted traffic avoidance

Packet network shall support security to avoid unwanted traffic e.g. spam or traffic generated from malicious nodes

F 11.8: Multipoint VPNs

The network shall support multipoint-to-multipoint (i.e. full mesh) VPNs

3.8 Charging Aspects

The network shall support various charging models including all those supported by the 3GPP system contained within TS22.115 and be able to support introduction of new charging schemes including online and offline schemes, and charging schemes for the multi-access system environment.

A 12.1: User profile extension

The user profile shall include information about charging schemas access type, technology preference and location.

F 12.1: M2M Charging

Introduce Machine Class Subscription Identifiers. M2M charging model should allow reduced overhead for small payloads. Count traffic to and from the servers at the network boundary. Allow charging for groups of devices.

N 12.1: Operator legal aspects

Preserve the ability of an operator to fulfil obligations towards regulators and government authorities to guarantee secure authentication and billing

3.9 Energy efficiency

The specification of new architecture design must take into account energy-efficiency issues. Access, core network and backend "cloud" efficiency should be considered to total energy efficiency.

F 13.1: Minimize device battery consumption

The network architecture shall minimize the mobile device battery consumption.

A 13.2 Low consumption mode

The network elements should be able to go into low energy consumption mode when possible.

3.10 Traffic management

Traffic management and engineering cover all measures to dynamically control and optimize traffic flows in a network domain or in a global view of the Internet, aiming at ensuring a maximum throughput and sufficient QoS/QoE for the users. In order to achieve this goal, traffic management includes methods and schemes for dimensioning, admission control, service and user differentiation and failure resilience as well. The specification of the MEVICO architecture should meet these needs.

F 14.1: Application-awareness

The traffic management must be able to provide means for traffic classification based on application types.

F 14.2: Support for macroscopic traffic management

E2E traffic steering, usage of proper on/off switching cost function, uplink bottleneck detection in cell breathing,

F 14.3: Support for microscopic traffic management

Support for multipath flows, QoE support in roaming case, cross-layer interference detection by traffic monitoring.

F 14.4: Improved content resource selection & caching

Peer or storage selection optimised for mobile networks, resource partitioning, unfavourable resource usage detection, P2P transit traffic reduction.

F 14.5: Support for deployment of new network resources and upgrading processes

Whenever possible, new resources should be integrated in a self-organizing and seamless way. The transmission capacities should be easily adaptable to steadily increasing traffic within a sufficiently wide scalability range.

4. Architecture Challenges

Based on the requirements (section 3), the architecture challenges, taken into account in MEVICO project, are related to the overall mobile architecture not only traffic engineering, and deals with the following aspects, introduced briefly here and discussed further below:

- Network topology (scalability challenge)
 - Target: flexible topology
 - Constraints: heterogeneous access networks, SGi interface, PCC architecture...
- Mobility
 - Target: diversity of connected devices and access networks handling...
 - Constraints: load balancing, heterogeneous network accesses, multipath routing...
- Network transport
 - Target: overcome cost crisis, new synchronization requirements support, multiple traffic flows differentiation...
 - Constraints: strict mutual timing requirements between BTS, various migration paths support with different technologies...
- Network management
 - Target: SON implementation, common management system for different RA technologies.
 - Constraints: SON potential conflict with transport network management. Co-existence of different technologies. Distribution of the architecture...
- Traffic management
 - Target: QoS differentiation; connection management over multiple flows, massive multimedia transmissions optimisation, efficient offloading techniques, switch on / off schemes of networking equipments...
- Network applications and services
 - Target: M2M related challenges (group addressing...), energy efficiency challenges, efficient resource usage...

In general there is always some trade-off between performance and functionality.

4.1 Network Topology related challenges

The most important challenges concern the scalability of the network that can be ensured among other ways by an adaptive/flexible topology, against different parameters such as traffic load, subscriber density, number of network connections and signalling transactions.

- Due to LTE radio throughput enhancements and new smart phone applications, the mobile network busy hour data traffic volume is expected to increase up to the ten fold in the next 10 years, so that a new backhaul /core network topology might be required to increase the network throughput capacity.
- In the same time, due to the increase of the mobile broadband subscribers number and due to the introduction of M2M devices with mobile network connection, the network topology will have to be adapted against the density increase of attached User Equipment (UE) and that especially as a default EPS bearer will be systematically created for each new attached LTE UE.
- As the new EPC network aims at supporting both conversational and classical data traffic, Quality of Service mechanisms will have to be able to handle this higher number of network connections per UE. This might induce a change in network topology.
- If the increased need for access network throughput and session handling capability leads to cell density increase (huge number of Femtocells for instance), this will cause potentially bottlenecks for the

communication or the signalling transactions within the centralized gateway and servers that handle mobility and service provisioning, once reached the network elements capacity upgrade limits.

The scalability issues of the mobile network will depend on the capacity evolution of the EPC nodes² compared to the network load increase for each of the above parameters. An additional challenge is to identify from CAPEX and OPEX point of view the most appropriate EPC nodes localization from a centralized to a distributed architecture, and that will enable to eventually distribute further the EPC architecture. The proper distribution of the EPC nodes will have to take into account the following parameters:

- In the case of heterogeneous access networks, the core nodes optimal positioning for LTE networks and non 3GPP networks might not be the same since the session amount, the traffic bandwidth, the handover frequency or even the service types profile might change when a UE is connected to 3GPP and non 3GPP access network.
- Content and cache servers are getting deployed at the edge of the fixed networks, so that a distribution of the mobile network could permit to merge fixed and mobile content and cache servers.
- Centralized mobile networks permit the use of customized accounting devices in order for a mobile operator to propose offers for its customers, whereas the distribution of the mobile network requires the use of less costly and by the way less accurate and more standard accounting features.
- The complexification of sGi interface/APN management/PCC architecture/company connections...

4.2 Mobility related challenges

The increase in the number of connected devices, diversity of access networks, and the resources limitations pose real challenges on how the network will handle security, users, and flows contexts. This together with the expected data traffic growth will have a serious impact on flows performances when considering the current centralized architecture approach and existing mobility management and routing procedures (e.g., bottlenecks, overloaded access networks). Specifically, the following challenges are of crucial importance:

- Anchor-based mobility management protocols (Mobile IP, Proxy Mobile IP, etc.) for non-3GPP accesses rely on centralization of all traffics towards a unique anchor wherever UEs are currently attached (potentially far from the anchor). Mobility of UEs will lead to considerable amount of traffic routed throughout the core network to the centralized anchor.
- Some UEs applications need regular access to the network, which is often referred to as the “always-on” mode even if no user data is to be transmitted. This means that some traffic still passes through the P-GW and P-GW load balancing is difficult to perform.
- Current network devices may have several interfaces able to get access to different types of network (3G, Wi-Fi, etc.). When one access network is overloaded, it might be possible to redirect traffic to other access networks³ or to perform multipath routing.

A decentralized architecture with multiple external gateways is a relevant approach to distribute network resources and to handle scalability issues. However, it is expected that the multiplication of GWs will also lead to more frequent inter-GW handovers. Therefore, mobility management solutions and security mechanisms have to be adapted to cope with this phenomenon.

- In some cases, UEs communications do not require: (1) the support of a specific (L2, L3) mobility management protocol because it is handled at another layer (e.g., application layer with SIP) or because UEs are mainly static (e.g., M2M devices, sensors, home location, etc.), (2) seamless mobility support as transport protocols are able to handle packet loss (e.g., TCP). However, at attachment all UEs are handled automatically by most current mobility management protocols, leading to wasted network resources for the above depicted.

When moving from one un-trusted access system to another (like between two different WLAN networks), a considerable delay is introduced by setting up a new security association. Furthermore, the current use of IKEv2 in EPC can lead to overlapping (encapsulated) IPsec connections. E.g., in case of initializing an IMS session through a 3GPP Wi-Fi access, an IPsec association is established both on the network level and on the SIP signalling level, resulting in overprotection and signalling overhead between the UE and ePDG. In case of trusted WLAN access IKEv2/IPsec is not used because the security control is made on layer 2 between the UE

² Note that several vendors already present high capacity figures, there is an order of magnitude of one million for simultaneous active users per MME and/or S-P Gateway.

³ Inter-RAT traffic steering/load balancing can enable this by querying the radio Call Admission Control of the nodes in the other RAT (requires that CAC is implemented and enabled in those systems) or by collecting load indicators used by the inter-RAT Load Balancing (requires that LB is implemented and enabled in those systems); the challenge is then the multi-vendor/multi technology environment. See also IFOM, PMIP, ANDSF, SON and the type of architecture.

and the non-3GPP access point. Hence for that scenario the IPsec overhead problem does not exist. It is up to future scenarios whether the handover between untrusted non-3GPP access is relevant. There are still use cases where the operator is not aware of the available Wi-Fi APs. It may be the case for enterprise networks for example.

- Paging enables to reduce energy consumption as it is not necessary for the MN to be permanently connected to the network. In EPC the paging requires that the MN has a specific allocated ID (the GUTI most of the time). In a distributed architecture, this ID might change frequently. Current paging procedures – operating when packets are coming in - would be non optimal with distributed MME or even unfeasible. Paging and Location Update (LU) procedures should take into account the upcoming multiplicity of gateways and interfaces per active UE to extend and improve the performances of idle mode management procedures.
- To improve user experiences, the EPC might propose/enforce its policy of vertical handovers towards networks with higher available resources. Some UEs are able to connect to several types of accesses or networks (LTE, Femto, Pico, WLAN, etc.) and so, the operator could have in the core network, functions to support smart vertical handover. Some of these functions already exist (e.g., 802.21, and for selection/reselection: ANDSF) but they need to be upgraded accordingly.

To leverage all those benefits, mobility management protocol should be extended to support new types of UEs (moving networks, M2M, etc.) and optimized to reduce routing path lengths. Meanwhile, new routing solutions should be overseen to better handle UEs mobility.

- Existing mobility management protocols do not all support moving networks (train, bus, aircraft, cars, boats, etc.). Those networks are interfaced by one or more mobile routers and provide connectivity to several UEs
- Anchor-based mobility management solutions suffer from triangular routing (the routing towards the anchor when two UEs are close to each other). Such sub-optimal routing has to be handled to improve network resources usage.
- Future routing solutions may require new locator namespaces and routing mechanisms. Introduction of new locator types and routing mechanisms specific to the intra-domain should be supported independently from the identifiers used in the service stratum, and without influencing inter-domain routing.

4.3 Network Transport related challenges

Due to Internet and peer to peer services, the traffic has increased heavily but the flat rate tariff prevents revenues to grow in similar pace to cover the increasing costs. New innovative Mobile Transport solutions, possibly optimized together with future mobile systems (LTE-Advanced and beyond), are needed to overcome the cost crisis from transport point of view. These changes consist of certain architecture changes as follows.

A **flat architecture** of LTE, i.e. moving radio controller functions to the BTS, affects a lot the quality and performance requirements to MBH transport. The delay sensitive loops do not necessarily exist anymore between a BTS and its controller. On the other hand new synchronized air interfaces may need very strict mutual timing requirements (microsecond level of phase/time synchronization accuracy) between base stations.

Some **transport node functionalities can be integrated to the base station** (e.g. Ethernet switching), when BTS functions like a part of normal MBH solution. Switching and some network management functions are physically inside a BTS but they are part of E2E MBH concept.

Some BTS internal interfaces can be brought out and extended by fibre where available; i.e. **BTS is split in two parts** – centralized base band processing node (BB Hotel) and distributed antenna RF heads.

The **new LTE X2 interface** between the adjacent base stations (eNBs) of LTE architecture is used for handover (HO) negotiations (control plane) and data forwarding (user plane) caused by the handover process.

The energy consumption, CAPEX and OPEX will increase with the bandwidth.

The size of addressing and routing tables will increase with the number of end points, thus the signalling overhead will also increase.

The transport network needs to support various migration paths with different technologies (e.g. Carrier Grade Ethernet, MPLS-TP, IP/MPLS, PBB-TE, etc).

Multiple topologies (i.e. centralized or distributed) should be supported by the transport network.

The security concept management should be developed in a more dynamic environment (re-negotiation of security parameters etc.)

The importance of the horizontal X2 interface might increase

- Some RAN related new features might need to increase the transported data and set strict latency requirements for the X2 interfaces between the eNBs

- The amount of X2 peers increases with the increasing amount of eNBs

The Network Transport related challenges are the following

- The transport network should allow the possibility to be shared with co-sited base station from different operators.
- New synchronization requirements have to be supported.
- In order to share the transport network between multiple operators or in order to differentiate various traffic flows both IP addresses and VLANs play an important role so the transport network should be able to handle them efficiently.
- Ethernet is widely deployed in core network and mobile backhaul so it does not only require point to point or point to multipoint but also multipoint to multipoint connections.
- The sharing of functionalities between L2 (i.e. Ethernet switch) and L3 (i.e. IP router): routing and addressing are based on IP at L3 but Ethernet is implementing part of that functionality.
- The transport network needs to provide secure communications so Ethernet needs to address security or IPSec is handling the security, thus it has some performance impact.
- The Transport network needs to differentiate multiple traffic flows with different QoS.
- The transport network needs to support the required capacity according to the expected traffic as well as maintaining acceptable delay.
- The transport network should provide plug and play functionality to allow integrating new eNB when needed due to capacity or when changing technologies in the existing eNB via SW upgrades.

4.4 Network Management related challenges

The evolution of the RAN introduces new requirements (for instance, CoMP and strict requirement on the X2 interface) and increases the complexity of network management that needs to deal with the co-existence of different technologies, e.g., RATs (HSPA, HSPA+, LTE, LTE-A, Wi-Fi). In addition, distribution of the architecture increases the number of network elements to be managed. SON (Self Organizing Network) for LTE RAN provides several features optimizing radio resource usage and automating radio network setups. SON will help simplifying transport setup for network elements (e.g. femto/micro cells). Related to SON's features, the following items bring some challenges to the mobile network architecture:

- The radio SON features are transport agnostic, thus their operation may unintentionally conflict with transport network management.
- The number of nodes in the EPS (eNBs, GWs, IMS servers, content servers, routers, switches, Femtocell GWs) is very high and the management solutions should enable scalable operations. Furthermore, the increase of M2M subscriptions and M2M generated traffic monitoring requirements might set new challenges to scalability.
- With the common EPC for multiple radio accesses, intra-3GPP inter-RAT handovers need to be improved for balancing load between RATs coexisting over the same coverage area. In the same way, hand-overs between 3GPP and non 3GPP accesses (e.g. LTE – Wi-Fi) can provide boost in performance of the network as perceived by the user and improve resource usage (see load balancing).
- Different RA technologies (HSPA, HSPA+, LTE, LTE-A) will coexist for a significant time and it would be uneconomical to build separate mobile backhaul or management system for each of them, thus these should be able to handle the traffic and management functions of the RA technologies efficiently and allow sharing of the bandwidth without unwontedly privileging one customer over another.

4.5 Traffic management related challenges

As mobile and wireless communication architectures evolve toward broadband multiplay and multimedia networks, the demands for solutions on the infrastructure increase. Legacy voice, and novel data, video and other applications are to be served on the same network, simultaneously. Advanced terminals (i.e., smart phones, tablet PCs and other mobile devices) are spreading and consuming more and more network resources by running their multimedia applications and services. Consequently, the needs for available wireless bandwidth will constantly increase and LTE/LTE-A networks will likely follow the same path as wireline networks in the past resulting in a significant expansion of CDN and P2P traffic volumes.

In such a fast development it is essential that the network must be aware of each application's traffic type and enforce traffic management and control (i.e. priority, routing, bandwidth, etc.) required for ensuring improved Quality of Experience for every user anytime and anywhere. Assuring that mobile and wireless communication systems are application-aware, operators can achieve flexible adaptation to any new application and traffic pattern as soon as they emerge in the future. Operators need to install effective management tools to control every traffic component using QoS policies, prioritised access and admission control, bandwidth allocation schemes, traffic shaping and rate control, and flow based processing. Only such an active, advanced traffic management will ensure that operators can provide cost-effective data transfer with real-time multimedia information over heterogeneous access architectures of future networking schemes.

Based on these considerations the main traffic management related challenges identified in project MEVICO are the following:

- Satisfy user experience with minimum of infrastructure resources and still be flexible to handle the possible large variation of traffic patterns over time.
- Initiate handovers of sessions and/or flows not only based on signal degradation, costs, etc. but also based on a possible threat of congestion or any other threat on the QoS-QoE conditions.
- Provide QoS differentiation based on both applications and user profiles and ensure an appropriate scheme of user and application prioritization and differentiation which is not limited to forwarding behaviour but may consider access control as well.
- Split and manage connections (e.g., TCP sessions) over multiple flows inside the network.
- Optimize P2P and massive multimedia transmissions over the network
- Solve the problems of existing combinations of link layer ARQ and TCP (unnecessary TCP retransmission causes unwanted traffic through the network and reduces application throughput and response times).
- Optimal design and efficient management of Content Delivery Networks in an operator's infrastructure (e.g., identify suitable locations for caching, select suitable locations for content, detect unfavourable resource usage, redirect requesting node to alternative resource, etc.)
- Implement efficient offloading techniques, access network/core network elements (re)selection schemes in order to effectively distribute users' data traffic through localized wireless access points (femtocells or WLAN) and to locate service gateways (breakout points) near to those access points (aiming to avoid non-optimal routing and overloading of the network elements).
- Supply switch on / off schemes of networking equipments with traffic management aware decision algorithms.
- Anticipate applying an intelligent planning process for extending the available resources (i.e., design optimal or near-optimal capacity extension procedures which are able to cope with the enormous traffic volume evolution).
- Enable fast re-active mechanisms based on detection of application and network layer events to accomplish rate adaptation for multimedia streaming application and synchronization with resource management in EPS networks.

Traffic management functions tackling the above challenges usually require access to higher layer user plane data, i.e. IP packets, TCP segments and application layer protocols. Placement of such functions at SGi interface (co-located with PDN-GW) or S5 interface (co-located with S-GW) are possible options, since GTP tunnelling is terminated at these locations. In the following, a brief analysis is done on possible impacts. Positioning of TM functions at S-GW implies that all user plane data can be managed by the considered function unless there is handover between 3GPP and non 3GPP access network. In such case user plane traffic could not be processed or handled by the same node, hosting the TM function. If this can't be avoided, possibly different instances of the TM function have to coordinate in order to ensure continuous TM operation, in case such feature is supported by the TM function in consideration. Positioning the TM function at SGi interface – e.g. co-located with PDN-GW – may cause problems if user data is transferred using different access point names (APN). This usually implies that data paths stretch along different SGi interfaces. It is common practice in currently deployed networks to allocate the same APN to a user for all OTT services. However managed operator services may use different APNs. As a consequence the same TM function may not be used for connection via different APNs in case of multiple distributed PDN-GWs. Indeed this would require changes in network elements, e.g. a master TM or communications between TM elements. This situation would increase equipment cost (CAPEX) as well as operational cost (OPEX). As a consequence, the suitable location of TM functions depends on mobility aspects (whether a TM function needs to be supported after 3GPP-non3GPP handover) or connectivity aspects (whether the same TM function shall be in usage for services using different APN).

4.6 Network applications and services related challenges

This section describes some of the challenges that new applications and services will bring to the mobile architecture based on their specific needs and requirements.

4.6.1 M2M related challenges

Machine to Machine (M2M) service evolution will set challenges for mobile network functionalities:

- Individual M2M device addressing, global addressing, group addressing, device vendor based provisioning.
- Network selection mechanisms, location tracking, steering of roaming for MTC devices.
- M2M specific properties: MTC group concept, MTC monitoring, MTC time controlled and time tolerant functions, MTC low mobility, MTC small data transmission.
- Charging mechanism specific for M2M communications.

4.6.2 Energy efficiency related challenges

Heterogeneous overlapping networks and potentially more distributed architecture might increase the total energy consumption and might be underutilized with the lower traffic times or unevenly utilized. Then the optimum and controlled resource utilization can provide some energy savings.

Network controlled reduction of energy consumption in the devices for extending the battery life might be challenging, in the heterogeneous radio network environment. Scanning of the possible radio interfaces might be able to be optimized based on the network delivered information about availability of other 3GPP or non-3GPP accesses.

4.6.3 Improved user experience and efficient resource usage

This set of challenges is associated on one hand with poor quality of experience for running multimedia (streaming) applications on mobile networks. Secondly unexpected traffic patterns, like caused by flash mobs and other events may significantly contribute to decline the amount of potentially available resources. Some of the following aspects are not restricted to streaming applications but it is assumed that this class of applications needs a special focus in the project with respect to traffic management.

- How to achieve acceptable QoE for OTT (Over The Top) content (located in external CDN / network)?
 - Some content may not be cached within the domain of the MNO, but inserted from a 3rd party content / CDN provider.
- Detection and localization of high traffic load within local domain or external network:
 - timely detection of problems and proper reaction mechanisms.
- Increasing amount of traffic in upstream direction:
 - usually there is less capacity on the path in upstream direction – some user hosted content might be shifted into the network,
 - some applications (like video conferencing) require QoS support in both directions of flow.
- Align resource selection principles from application with constraints from the network:
 - detection of and reaction to unfavourable selection,
 - how to manipulate resource location information (DNS, etc.) based on resource selection principles,
 - influence resource selection in external network / CDN.
- Improve QoE by caching popular content from Internet:
 - analysis of promising caching strategies especially for mobile access according to content popularity over time, of the possibility for content partitioning and other factors.
- Inefficient content delivery to mobile devices:
 - optimization for unicast and multicast streaming applications,
 - take into consideration capabilities of mobile devices, wireless access and subscriber profile.
- How to go from required QoE for the user to QoS support in the network?
- Support of QoE in roaming case:
 - transfer of content without dedicated QoS control like the one provided in the home network.

5. Proposed Technology Solutions

This section describes the MEVICO Proposed Technology Solutions to cover the Architecture Challenges identified in section 4. Each sub-section includes high level description of the related problem statement and focuses on the aspects that MEVICO project will address.

5.1 Mobility

Facing the traffic evolution trends, higher network throughput and better scalability and flexibility of the core network functions are required as was concluded in the network topology related challenges in Section 4.1. All challenges under the mobility topic described in Section 4.2 are connected to this previous goal. The main challenges are to elaborate appropriate mobility management and path selection strategies facing the foreseen trends of traffic demands and user behaviours. This topic focuses on user terminal and EPC element aspects. The proposed solutions in the focus of this project for the above mentioned challenges are the following.

Smart traffic steering

Smart traffic steering with multi-access terminals and multipath protocols will enable better load distribution considering user, network and application preferences. The functions needed for smart traffic steering are the followings.

Smart traffic steering decisions: the most important selection problems considered are access interface selection, gateway selection, source address selection during terminal attachment and session establishment. For the support of terminal and flow mobility, mobility anchor selection in a distributed mobility management scenario requires novel classification algorithms as well. Enabling technologies investigated in the project will be the IEEE 802.21 Media Independent Handover protocol which provides a framework for transverse information services, physical and link layer resource monitoring, reservation and release. The 3GPP Rel-8 Access Network Discovery and Selection Function (ANDSF) describing the access interface selection policies must be further improved to provide an optimized set of rules to the UE.

Multipath technologies: the Multipath TCP (MPTCP) can transmit one TCP flow over multiple interfaces, and can balance the load between subflows. Stream Control Transmission Protocol (SCTP) supports multistreaming, i.e. several streams related to the same application can be handled by one SCTP stream, and backup SCTP associations. Both technologies will be analyzed and further improved to enable multipath communication.

Flow mobility: The performance of 3GPP Rel-10 IP Flow Mobility (IFOM) will be evaluated. IFOM enables smart IP flow allocation.

Offloading techniques: offload the EPC and LTE through non-3GPP networks could further improve the overall network throughput and quality of service. Access offload through IEEE 802.11 managed by the 3GPP operator will be evaluated.

Distributed and dynamic mobility management

These solutions cover terminal and flow mobility, and reachability of multi-access devices on L2 and IP level. The technologies developed in this project aim to achieve the following properties: increase network throughput by the support of a dynamic activation of mobility signaling and by providing distributed, anchorless or partially anchorless solutions.

Mobility management technologies include Session Initiation Protocol for SIP-based services and SCTP for non SIP-based applications. These technologies basically can provide end-to-end, anchorless mobility, but they will be applied in a flat or distributed approach in the EPS.

Proxy Mobile IPv6 (PMIPv6) will be extended with route optimization procedure among the Mobility Anchor Gateways.

The project also covers how to adapt the Host Identity Protocol to provide distributed mobility management in EPS. HIP by default follows an end-to-end approach, hence could provide an anchorless solution.

A new Ethernet-level mobility management solution will be developed and evaluated, that could replace the GTP concept of EPC by Ethernet VLAN tunneling, hence reduce the overhead.

Evolution of the current 3GPP based model (GTP tunnels) with the dynamic and distributed mobility principles will be studied.

An anchorless mobility solution for TCP sessions called NMIP (TCP rehash) will be evaluated.

DMA (Distributed Mobility Anchoring) has been initially discussed in IETF to improve MIP/PMIP by distributing mobility anchors and use as much as possible a local, not tunnelled addresses, see also DDMM.

Here the technology intends to optimize the EPC based on the ideas of the DMA, but utilizing existing 3GPP protocols like GTP with as less as possible changes, to enabling SW upgrades to optimize the usage of existing resources. An underlying assumption is that a GW distribution brings certain benefits. A proposal is to change PGWs using intelligence in the PGW or changing SGWs for routing optimization.

The first DMA solution applies after a UE has moved into a new “gateway area”. The PGW selects IP (PDN) connections for what a new IP address and service interruption may be acceptable from application point of view and forces a reconnection that allocates a new more optimal PGW and new IP Address. This leads to more optimal routing and savings in transport networks.

A second DMA solution proposes to relocate the SGW to achieve maximal SGW-PGW collocation in a distributed architecture when UEs use different PDN connections. This saves at the end GW capacity. Different gateway locations may result from the fact that a UE may connect to local and/or central networks or Internet providers.

Access and network security

For untrusted non-3GPP access, the existing user access security procedures must be revised, and the communication protocol might be further optimized to the new distributed EPC architecture, aiming to reduce overprotection and decrease L2 and L3 re-authentication times during handover. The investigated technologies will be the Internet Key Exchange v2 protocol, HIP and HIP Diet Exchange that is a lightweight version of HIP.

Trusted WLAN solution solves already some level of overprotection. See 3GPP SaMOG (S2a mobility based on GTP & WLAN access to EPC).

Bootstrapping

Configuration of multi-access terminals might lead to conflicts in case of parameters that have wider than interface-level scope. These conflicts must be discovered and resolved. ANDSF policies will be investigated from that aspect.

5.2 Network Transport

The next billion Internet users will connect primarily through mobile networks. Therefore, mobile networks have to support constant growth of traffic and increase the throughput from 1Gbps to tens or hundreds of Gbps already in the near future. Ethernet-based technologies have several features that make them especially interesting. Therefore, Ethernet is a natural solution for replacing legacy SDH and other older technologies, and the energy consumption of L2 switching is an order of magnitude lower than IP routing.

The proposed solutions to cover the above mentioned research paradigms that will be deployed in this project are the following:

Carrier Grade Ethernet with inbuilt O&M

The objective is to provide Carrier Grade Ethernet and overcome the limitations of using Ethernet for large scale networks. In order to provide reliability and robustness required for Carrier Grade Networking an O&M mechanism is required. Therefore, the goal is to enable routed based Ethernet where the O&M functions will provide the necessary routing optimizations and bootstrapping algorithms, as well as the link break detection and route recovery mechanism.

Ethernet Mobility to the Edges based on TRILL

TRILL leverages IS-IS routing protocol to achieve Ethernet frame shortest path routing with arbitrary topologies. In this research item the goal is to utilize TRILL extended with DHT to deploy mobility in the network edges. The goal is to combine the advantages of bridging and routing and fully distributed mobility mechanism implemented in the Link layer (i.e. Ethernet). In order to increase the available throughput we consider that is necessary to move towards lower layer switching and minimize processing per packet.

Customer Edge switching

The unwanted traffic is one of the reasons for inefficient usage of resources (i.e. radio spectrum, routers, bandwidth, etc). Unwanted traffic includes port scanning, intrusion attacks, viruses, email spam, traffic to reallocated addresses, and generally traffic from sources that the user does not want communication with. The unwanted traffic has to be filtered before entering the operator network to avoid waste of transmission capacity (bandwidth, resource usage). Inbound packets should only be forwarded if the user expects them, either by having an ongoing session or by running a server (e.g. a SIP UAS) expecting traffic. In this objective we propose to deploy a Customer Edge Switching (CES) element that will enable setting up an end-to-end trust connection for traffic where the sources has been verified. The CES operates similarly to a NAT or firewall, but with added functionality for accepting inbound connections and with traffic control based on policies. The CES does not only improve security but also extends the amount of available addresses (similarly to NATs)

while enabling inbound traffic. The CES interacts with other systems such as Deep Packet Inspection (DPI) and reputation systems performing part of the filtering functions and sharing trust information.

Automatic and Secure L2 Virtual Private Networks (VPNs)

Virtual Private Networks (VPNs) are popular among network providers that wish to separate multiple LAN domains across a single network infrastructure. One VPN technique is Virtual Private LAN Service (VPLS), a layer 2 (L2) solution that connects several physically separated LAN segments in to one logical LAN segment, i.e. emulated LAN or VPN overlay. This research item is interested in investigating so called “bump-in-the-wire” customer VPLS solutions in which the VPN service is overlaid on top of a provider network combined of IPv4 and/or IPv6 hybrid segments. In particular, the research item studies how identities can be utilized to mutually authenticate the PEs as belonging to a certain overlay and facilitating the renumbering of the PE devices.

Wireless mesh networks for mobile backhaul first mile access

With the introduction of WiMAX and LTE the need for mobile backhaul transport capacity grows rapidly, to the level of Gbps. Fibre media is able to provide the high data rates but fibre is not available everywhere either for technical or commercial reasons. Additionally, more base station sites, with different cell sizes, must be provided to meet the capacity and coverage requirements. Therefore, new wireless solutions are needed for the backhaul, especially in crowded areas characterized often by lack of available frequencies. One feasible solution providing sufficient transport bandwidth and capacity is E-band (71-76 GHz, 81-86 GHz) microwave radio with Ethernet connectivity. A wireless mesh backhaul can be used for small cell, high capacity base station first mile access and for other high capacity packet connections (e.g. office and home access), traffic management.

Relaying

Relaying techniques are considered as an alternative solution to enhance capacity for the cellular networks, to extend coverage in specific locations, to increase throughput in hotspots and to overcome excessive shadowing. It gives important advantages such as ease of deployment and reduced deployment costs and decreased output power compared to deploying regular Base Station (BS). Moreover, there is no need to install a specific backhaul in the network. It is an important aspect and one of the key technologies taken into consideration during the standardization process of 4G technology LTE-Advanced. RNs are also envisioned to be transparent to UE. In other words, the UE is not aware of whether it is connected to RN or a conventional base station. This ensures backward compatibility with previous LTE releases 8/9. Therefore, gradual introduction of relays without affecting the existing structure of UE's can be ensured.

Relaying promises coverage-area extensions and high data rates for the cell edge users. This is especially useful because LTE will operate on high carrier frequencies, i.e. 2.6 Ghz which will result in ultra-dense deployment of network nodes, the transmit power is limited when transmitting broadband at the cell edge and the most of the traffic is generated indoor. It can also be used as a capacity improvement with load balancing and cooperative relaying techniques.

Current relay architecture in 3GPP LTE Release 10 assumes fixed relays. However, handover of a relay from one donor eNodeB to another donor eNodeB should also be supported in future network architectures and releases which will be a consequence of mobile relaying.

5.3 Traffic Management

Section 4.5 introduces the main traffic management related challenges which are connected to the MEVICO goals and motivations. In order to tackle these challenges, techniques operating within different traffic management building blocks must be considered.

- First, mechanisms with the primary objective is to improve performance of individual flows based on application type, user profile and other policy related information must be incorporated. Such solutions are belonging to the microscopic traffic management (1) building block.
- Second, the macroscopic traffic management (2) must also be introduced in the network with the primary objective to improve efficient usage of network resources. Parameters for optimization in this case describe traffic patterns without detailed knowledge of individual flow attributes.
- In addition to microscopic and macroscopic traffic management, a third group is improved resource selection and caching (3). The associated mechanisms address the selection of resources in distributed data management systems (P2P, CDN, caching), if necessary. This building block may rely on services of both microscopic and macroscopic traffic management. These could be in place without dependence to other traffic management building blocks. Cross-layer P2P is a novel technique where the ISPs can have control over the non-optimized

P2P traffic. Proactive Network Provider Participation for P2P (P4P) is a promising solution to non-optimized and self-organizing P2P.

- The fourth building block is called as application supported traffic management (4) which tries to optimize performance from end user perspective of certain, widespread applications (e.g., based on CDN and P2P) without getting support from network elements.
- The fifth building block is more relevant from business perspective without too many technical aspects: steering user behaviour (5) is mainly used by network operators and by possibly other stake holders as well in order to influence user behaviour by defining certain constraints for usage of networks / services and certain incentive to comply with the usage constraints.
- The last building block in this enumeration is about capacity extension in case the available network is regularly in high load conditions. It is the challenge to apply an intelligent planning process for extending the available resource (6). In addition to the building blocks there are some common functions like policy control and traffic monitoring.

Even though the above building blocks and the associated mechanisms / possible technical solutions should represent a functional decomposition of traffic management in EPS on high level it is assumed that some of the mechanisms to some extent are dependent on each other. It is one of the most important efforts of MEVICO to map the diverse mechanisms into functional components for the traffic management architecture and elaborate the dependencies between the building blocks and the containing functional components based on the various design options that have been identified.

5.4 Network Management

As indicated in section 4.4, the network management related challenges identified by MEVICO are mainly: avoiding conflicts when introducing SON features, finding solutions to scalability and heterogeneity requirements and managing intra 3GPP handovers and handovers between 3GPP-non 3GPP accesses for optimizations and load-balancing.

The common characteristic of all the alternative/complementary transport solutions is that, in each case, long-lasting connections (OC-x, LSP, EVC) are configured between the EPC nodes and that QoS schemes are applied either at the IP/MPLS or CET layer. These connections are configured either manually or via management tools that provide some level of automation when, for instance, the network is extended with new eNBs. In addition to the route/path of the connections and parameters needed for the connectivity (e.g., VLAN tags, IP addresses), other technology specific transport level parameters must be configured to define the amount of reserved/granted resources (as in the case of CET: CIR, EIR, CBS, EBS) or the level of service granted to a specific traffic class: scheduling weights, buffer allocations, etc.

To cope with the network management issues in EPC and heterogeneous networks, several topics and technologies have been identified that need to be addressed to make management more efficient.

Managing heterogeneity:

Heterogeneity involves managing, using the same management system, different co-existing network technologies; for instance, CET/DWDM, IP/Ethernet/NG SDH and also different radio technologies sharing the same transport network, such as HSPA, HSPA+, LTE, LTE-A. In order to maintain heterogeneity in the network an open standard layered network architecture for co-existing network technologies can be introduced.

Adapting EPC/Network Management to LTE-A features:

LTE-Advanced introduces several new features for enhancing the peak rates and service quality experienced by the user. Managing these features requires more strict synchronization and lower delays in the network. The scalable system bandwidth would put the requirement for more flexible resource allocation solutions and new management solutions might be required, for instance, to avoid introducing serious load on the X2 interface.

To validate that SON functionality and policies are working correctly it will be necessary to examine SON related messages from the S1-MME/S10/S11 interfaces to determine that the appropriate Network Elements are selected by the eNBs and S-GW. For validating the SON functionality and eNB algorithms and for optimizing the core element usage, normal signalling KPI's shall be used to determine if the network is equally loaded.

SON features in radio and transport Network Management

SON in LTE can efficiently improve the management and resource utilization of RAN but it is necessary to investigate the impact of radio SON features on transport network management and to find efficient global solutions.

The Mobility Robustness Optimization (MRO) SON feature aims, first, to reduce the number of handover-related radio link failures (Too Early HO (handover), Too Late HO) and, second, to automatically adjust the HO parameters to avoid incorrect HO parameter setting that can lead to inefficient use of network resources due to unnecessary or missed

handovers. For this, neighboring eNBs need to exchange certain information, e.g., through Radio Link Failure (RLF) reports.

Load balancing algorithms for EPC

The Load Balancing (MLB) SON feature aims to dynamically and automatically balance the traffic. The different hand-overs should be considered: hand-overs between eNBs, hand-overs between eNBs and HeNBs, inter-RAT hand-overs (e.g., LTE – HSPA, LTE – Wi-Fi). An eNB must know its own load and the load of its neighbouring cells. This information is exchanged through the X2 interface.

Network monitoring for EPC

To be able to correctly manage the networks using classical or SON techniques, precise information is needed at all times on the state of the network and estimations of traffic evolution for different types of traffic should be rendered possible. For this, network monitoring needs to be adapted to EPC and heterogeneous network constraints and several topics need to be addressed including : traffic analysis and capturing performance, time-stamp accuracy, protocol stack support, interface requirements and SON support as well as satisfying all the dependability requirements.

For MEVICO WP5 focused on the three main monitoring topics of interest...

- **End-to-end monitoring** to evaluate the QoS/QoE of applications and services,
- **Deep Packet Inspection (DPI)** for the identification and the classification of protocols and applications, and,
- **Monitoring of SON activities** for both testing the SON features and verifying them during operation.

Energy saving and impact on network monitoring and management

The Energy saving needs to be taken into consideration and, from a monitoring point of view, it is very important to provide network measurements to optimize the energy saving policy, and to test that the implemented energy saving policy behaves as expected.

5.5 Network applications and services

5.5.1 Network functionality virtualization and realization with cloud computing

Scalability and optimization of the mobile network architecture for high traffic demand are major challenges in the future. Virtualization and cloud computing methods have shown their potential in IT industry, like data center applications and have potential to be utilized to functionalities for mobile networks. Potential use cases are in mobile networks virtual operator concepts, network sharing principles and core network user plane distribution related virtualization and centralizing of the NW control overlay functions. This could enable better resource utilization, E2E QoS policy control and Heterogeneous Network (HetNet) control e.g. for load balancing.

5.5.2 Network Energy Efficiency improvements by efficient capability utilization

There is a need for optimizing the efficient usage of the mobile networks, because the number and the capacity of the network elements increase due to high capacity demand.

Depending on the changed user activity (like time of date, weekend etc.) some redundant capacity of the network can be switched off (like hot spot layers). The control and optimisation of this functionality need some more study, where also the network control virtualization could help.

In core network energy efficiency could be achieved by flattening the network protocol stack i.e. removing some of the network layers and use transport layer directly (i.e. Ethernet), to reduce the processing per packet. An improved traffic engineering to reduce the traffic to flood the network unnecessarily also would help in the energy efficiency.

The mobility management could be optimized, by reducing paging areas according to the terminal based parameters, such as mobility profile. Analyze the tradeoffs between location update and paging based on network hierarchy structure and where to store the paging information is needed.

RAN energy efficiency is a more important issue due to high amount of nodes, but this is considered in other projects than MEVICO.

The overhead when considering small cell scenarios cannot be ignored since it will increase heavily compared to the current overhead due to handover in current cell sizes. The overhead in small cell with high mobility will have bigger impact based on analysis from EARTH [9] project. In that study they consider mainly the radio aspect but the same signaling overhead will span over the fixed network up to the eNodeBand beyond. Quote from EARTH deliverable

D2.3 [9] “From an energy efficiency perspective, we are reminded that, according to our estimates, subframes without scheduled data are in the order of 20 times more frequent than subframes with data. This makes it important to carefully consider the fixed non-transmission dependent overhead....”

5.5.3 Network efficiency improvement for Video/Multimedia Applications

Video content delivery will set the biggest challenges to the mobile network scalability due to its high demand for bandwidth.

The further evolution need is studied for the mobile network architecture to support CDN, torrent-type of delivery and other intelligent data caching methods. It is important to identify the optimum location of CDN servers within the EPC network elements and possible impact to the architecture. These together with converging networks (fixed, mobile, home, enterprise) might indicate evolving sources of the data and thus needs for changes in the network topology and thus for architecture optimization.

It is necessary to look at transfer optimization protocols that are specific to information centric applications to enable controlling of traffic flows/content delivery according to content and access availability. One of these is the IETF Application Layer Transport Optimization (ALTO), which needs some further adaptation to wireless networks, where maintaining the QoE is even more challenging. There is also related work ongoing in EU Future Internet (FI) projects (like SAIL[11]).

5.5.4 Network improvement for M2M Applications

The foreseen high increase of the Machine-to-Machine (M2M) type of devices and applications might cause impacts or improvement needs to mobile network architecture or functional requirements. Most of the identified challenges are investigated in 3GPP Releases 11 and 12 work items NIMTC and SIMTC as well as ETSI Technical Committee M2M work.

The M2M application special characteristic issues to be studied for network architecture evolution:

- Improvements on network/interface selection mechanisms, location tracking, steering of roaming for MTC devices – Possibilities for signaling optimization
- The network impacts of the M2M specific properties: MTC group concept (MTC Devices that are co-located with other MTC Devices), MTC monitoring, MTC time controlled and time tolerant functions, MTC low mobility, MTC small data transmission.

5.5.5 Application based network traffic analysis and engineering

In order to improve network QoS and application QoE, there is need for modeling the selected Internet applications traffic characteristics and their adaptive behavior at times of congestion. Based on this input there are two types of traffic engineering mechanisms to be investigated:

- The macroscopic traffic engineering relates to adaptive routing, gateway selection mechanisms, multi-path transmission and mobility support, described in more details in Mobility section 5.1.
- The microscopic traffic engineering relates to mechanisms for rate reduction of active traffic flows under QoE constraints, based on the behavioral models of application operation, QoS mechanisms, measurement and control functions.

The application generated traffic mix observation at the short time scales is needed to test Traffic Engineering solutions, as well as for admission control of traffic for short term network dimensioning. Target is to utilize method to derive short term traffic mix estimations from long term ones given as input. The proposed method is non-linear unlike usual methods and thus more reliable.

5.6 Network Topology

From the perspective of operators, the challenges in the coming years are to be able to apply the required network evolutions for becoming mobile broadband integrated providers.

The current networks topology is centralized which creates a set of bottlenecks in the communication with the servers that handle mobility, service provisioning, etc. The objective is to propose and describe different architecture scenarios over the same network topology model and to study different protocol scenarios to be compared under different angles, for examples:

- Handover performances (establishment, delays)
- Flexibility in integrating and making use of various access technologies (typically LTE/Wi-Fi),
- Packet transport,
- Economical comparison according to traffic forecasts.

Different options for enabling multipath through Heterogeneous accesses and optimize the multipath management, should be analyzed. Moreover, the various transport protocols (GTP, MIP based, other ...) and the load balancing with heterogeneous access should be studied.

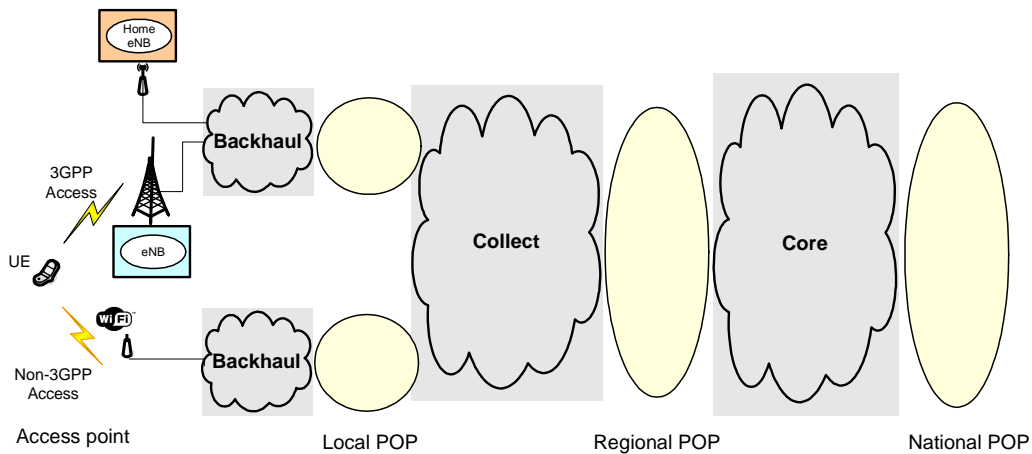


Figure 4 Network topology model

The architecture scenarios are divided into three families, depending on the distribution level of EPC and IMS nodes in network topology model presented in Figure 4.

- Scenario 1: progressive distribution from the national PoPs to the regional PoPs
- Scenario 2: progressive distribution from the regional PoPs to the local PoPs
- Scenario 3: maximum distribution to the Access points on the antenna sites

6. Architecture Approach

The results of the studies done in MEVICO related to mobile scenarios and traffic demands show that there are several parameters that will cause network scalability and performance problems. Some of the most relevant parameters are related to changes in traffic, i.e., growth in data volume (i.e. increase by 3-10 times by year 2020 at busy hour) and number of subscribers (i.e. increase of mobile broadband subscribers by 8-12 times by year 2020), and heterogeneity aspects including the handling of multiple traffic patterns (e.g. MTM), and smart and seamless support of multiple technologies, e.g., in the field of mobility management.

In order to simplify and narrow down the high number of usage scenarios, derived challenges and proposed technologies, it was inevitable to work out a new research process. The proposed technologies will cover different but probably overlapping challenges and functionalities, thus generating multiple architecture options. The whole process may guarantee that the proposed architecture provides coherent functionalities and tackles most of the challenges raised by usage trends.

There are several steps concerning the architecture options creating:

- Firstly, key performance indicators (KPIs) described in Section 6.2, have been defined for the validation of architecture proposals. The challenges described in Section 4 have been mapped to the key performance indicators based on the knowledge that whether the solution of a given challenge can be measured by the key performance indicator.
- Then the usage cases described in “IR1.1 Network usages and scenarios” have been prioritized and summarized in Section 2.2. We analyzed the number of challenges raised by the usage cases and selected the top three scenarios having the most challenges connected and targeted by the most of the proposed technologies.
- Then, technologies have been mapped with challenges based on the rationales of the technologies. Knowing which challenges are covered by which system validation KPIs, we can also enumerate the technologies which are relevant under a given system validation KPI. An initial ranking of the technologies have been made according to the KPIs, basically representing that the technology contributes to a given KPI or not.
- Top ranking technologies are considered with high importance of the focused architecture. The reasoning behind this is the intention to cover most of the challenges by fewest technologies. The aim of system validation is to prove that the specific technologies perform well under the system validation KPIs.
- At the same time, issue of co-existence and performance of top ranking technologies has been analyzed. Technologies covering similar functionality resulted in the creation of several architecture options.
- Finally, the system validation alternatives are architecture options, i.e., combinations of proposed technologies integrated within the EPC.

MEVICO proposes different architecture options - a set of technology solutions listed. It aims to address the traffic and user growth demands as well as the requirements and challenges identified in Section 3 and 4. This section proposes a set of architecture approaches to address the above mentioned high level requirements and challenges. Finally the new technology solutions proposed in Section 5 are mapped to the appropriate topology models, indicating the location of the functionality enhancement in the existing network elements or the need for the new elements. Rationales about the benefits and performance impacts as well as possible co-existence of the technologies are also discussed.

6.1 Topological models

This section defines the following concepts of centralised, distributed and flat architectures as baselines for the different topologies that MEVICO architecture approaches will consider when addressing the high level requirements and challenges.

6.1.1 Centralized architecture

The centralized architecture is considered as the current 3GPP Rel.10 architecture with the enabled functionalities (S-GW localization, SIPTO offload...). In the centralized architecture S/P-GW and IMS components are located in the National PoP.

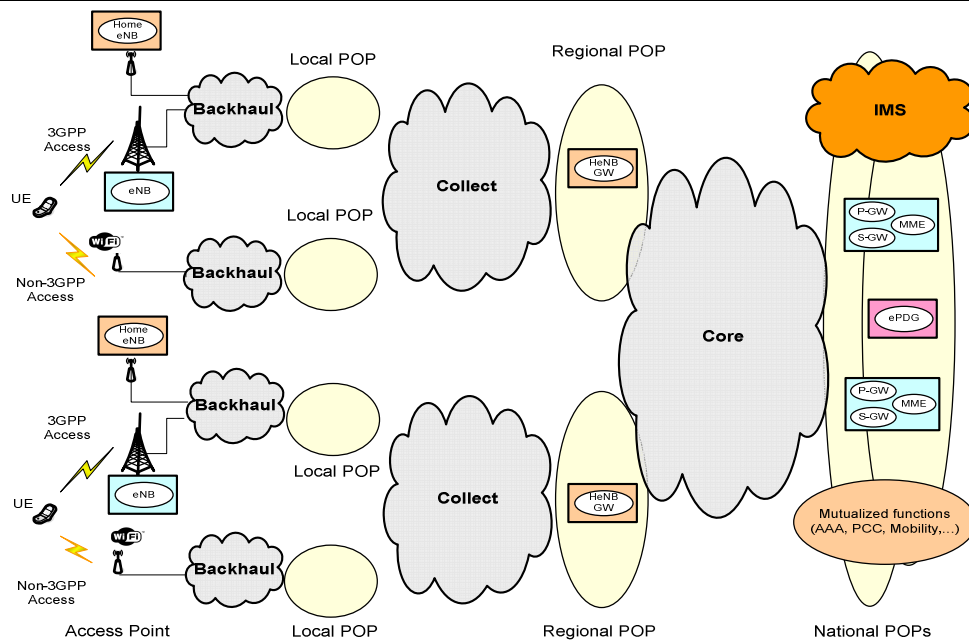


Figure 5 Centralized architecture model

6.1.2 Distributed architecture

The distributed architecture will include multiple gateways (S/P-GW functionality) located in the regional POPs. The distributed architecture is assuming that the functionalities are enabled to be distributable (optimized, and other EPC functionalities might still be centralized).

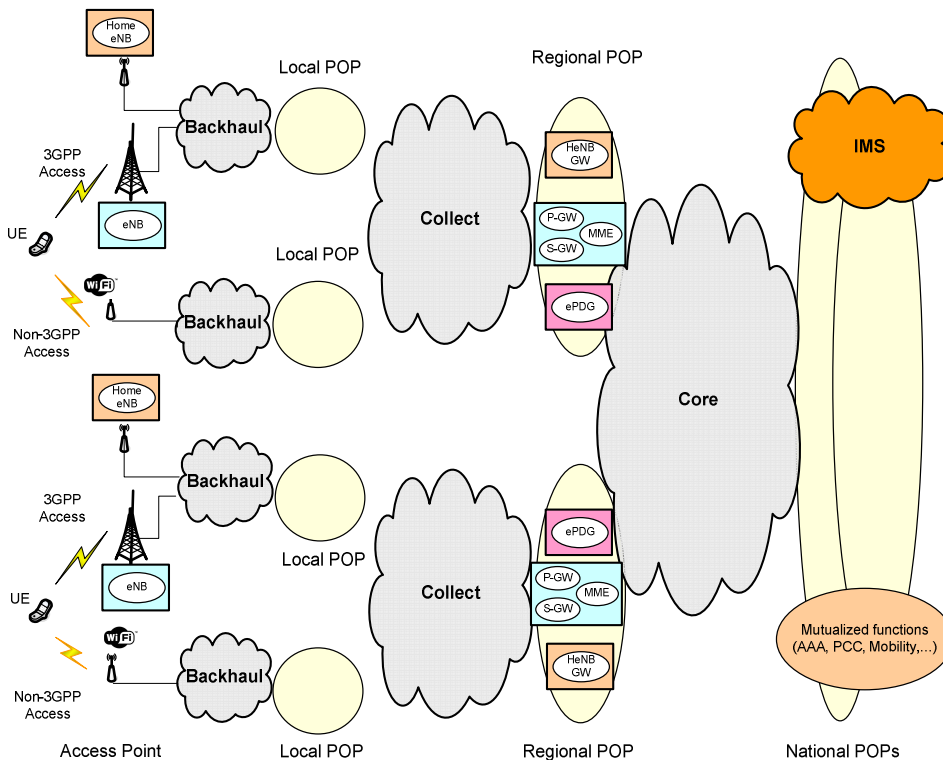


Figure 6 Distributed architecture model

6.1.3 Flat architecture

The flat architecture also referred as ultra flat architecture consists of the architecture where S/P-GW, MME and possibly (part of) IMS functionalities are in the local PoP. The legacy IP routed network is expected up to the eNodeB side (it will be operator controllable).

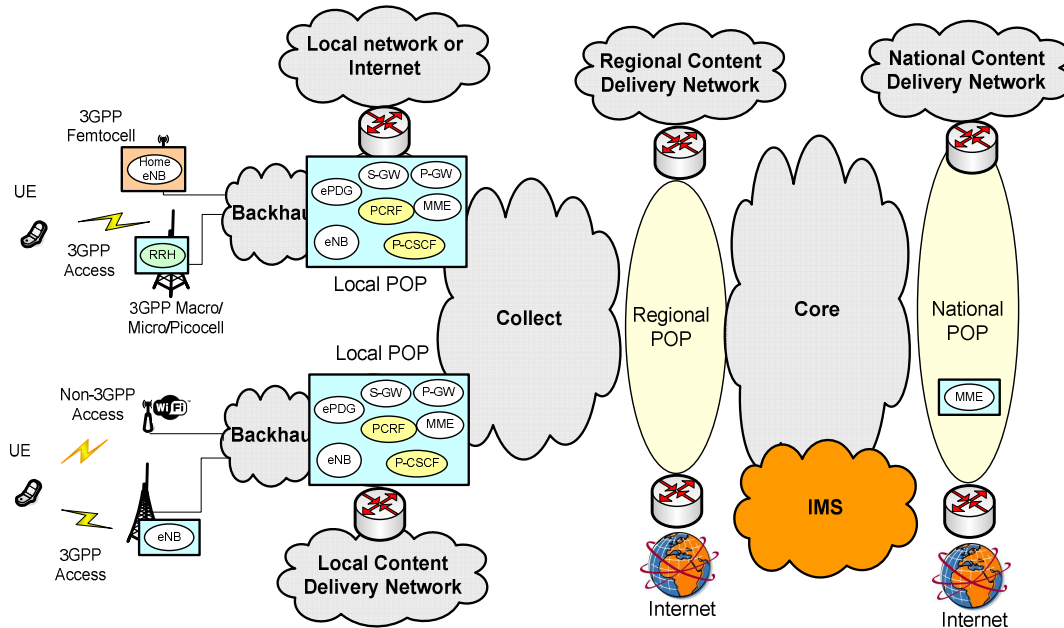


Figure 7 Flat architecture model

6.2 KPIs

This section describes the system validation criteria, i.e., specifies the criteria and the related performance metrics (i.e., system validation KPI). System validation KPIs in general should be considered as important questions/aspects to investigate in order to evaluate and rank the architecture options. For the assessment of the metrics, results from validations should be applied whenever possible.

In the following subsections, under each system validation KPI we enumerate the list of related challenges, and the technologies that try to address those challenges. Criteria definitions have been formulated for each challenge based on the challenge descriptions. These criteria together with the list of technologies and the set of functionalities in focus of the top ranking usage scenarios lead the specification of how to assess a system validation KPI.

Besides the specification of the evaluation of system validation criteria we also have defined satisfying performance values for KPIs based on specifications, or using some derivation based on the traffic trends. The performance of an architecture option should be estimated, and compared with these values when proposing the ranking. If it can be seen that the expectations will not be met, then integration issues are raised, and those questions should be addressed.

The system validation criteria can be split to three groups: performance criteria, deployment criteria, and validation/technology maturity criteria.

6.2.1 Performance criteria

Performance criteria are related to three important topics, i.e., improving backhaul, improving the usage of heterogeneous network resources and improving the core.

6.2.1.1 Throughput gain in 3GPP access and backhaul

The proposed technology will increase the network throughput and will be measured in terms of increase of number of packets and packets size in the access and backhaul. For example, expected throughput based on modest estimation for a large European country is 300 Gbs. The requirements in 3GPP defined for classes of traffic should be considered when measuring this KPI.

6.2.1.2 Backhaul and RAN influence on E-E delay

Requirements for delays in the backhaul and last mile come from the requirements from the end user applications. Typical delay budget for backhaul part in LTE case is 10 ms, reducing to 1ms for LTE-A and subms for beyond 4G mobile technologies. BS synchronization transfer and new mobile system features like 3GPP rel 11 CoMP may pose stricter delay and delay variation requirements for backhaul connections.

6.2.1.3 Reliability, recovery time from link failures, congestions and OPEX reduction

This KPI should consider the 3GPP requirements defined for link failure recovery. The KPI could measure the route establishment in switches when link break happens. The 50ms delay for link failure recovery is the starting point and it should be reduced. This KPI should calculate mean time between failures and recovery.

Other technologies such as SON can provide rough figures in terms of OPEX to compare the labor effort to be done against not having such technology in place.

6.2.1.4 Efficient load distribution in the backhaul and in the core

This KPI should show that the application of load balancing mechanism contributes to the non-congested states of the network in case of high traffic demands. The traffic load, the inter-arrival time and transmission delay should be measured either on end points or in the routers/switches if possible. The KPI could also use global packet loss ratio to measure the congestion of the end to end network

6.2.1.5 Offload gain due to the usage of multi-access capabilities

The KPI can measure the user and operator point of view.

- User point of view. The KPI should measure the traffic that goes on each interface of the UE in case of simultaneous use of radio interfaces or it should measure the end to end delay of transmission.
- Operator point of view. The KPI should measure the load on different elements of the network. It could measure the proportion of load in different access i.e. WiFi access versus LTE access.

6.2.1.6 Capacity aggregation and E2E QoE provision

This KPI should measure the throughput gain due to multipath communication, including goodput. The KPI will also measure QoS packet delay jitter packet loss plus any additional QoE measurements.

6.2.1.7 Service interruption delay due to handover

This KPI should measure the packet transmission additional delay due to flow mobility/handover. The KPI may measure the service interruption delay and jitter due to HO, as well. Packet delay budgets for guaranteed bitrate real-time services can be considered as hard constraints for induced E-E service interruption delay.

6.2.1.8 Handover related signaling load on the network

The handover procedures together with handover initialization, preparation, completion phases should be analyzed. Show that compared to state-of-the-art handover the new technologies provide reduced signaling load on different parts of the network.

This KPI may measure transmitted data overhead for HO process or the number of HO messages and their size.

6.2.1.9 E-E delay between UE and content

This KPI may measure RTT (on UE or server). The 3GPP Requirements by application type in terms of E2E delay budget should be considered. This KPI could also measure the path lengths in terms of number of L2/L3 hops.

6.2.1.10 Offload gains for core network equipments

This KPI may measure the throughput (i.e. number of data flows) on network elements such as S/P-GW, furthermore, user signalling reduction (i.e. number of signalling messages, the number of active user contexts per network equipment) on network elements such MME or S/P-GW. It can also measure goodput values on the specific network element.

6.2.2 Deployment criteria

6.2.2.1 Impact on UE

This KPI indicates whether the technology needs changes in the UEs. No changes are preferred over any change due to the wide range of differences and the high number of affected UEs.

6.2.2.2 Impact on existing network elements

This KPI indicates if a technology has impact on any of the existing network elements. No changes are preferred; however changes are better manageable than in case of the UEs.

6.2.2.3 Impact on new network elements

This KPI indicates if a technology requires new network elements. No additional elements are preferred, but addition of new network elements is better manageable than deployment of changes in the UEs.

6.2.3 Validation/technology maturity criteria

6.2.3.1 Standalone validation for each technology

This KPI indicates whether the technology will be validated in a standalone fashion.

6.2.3.2 Integrated validation of some technologies

This KPI indicates whether a technology can be validated with other technologies, and whether integration with other technologies is planned.

6.2.3.3 Validation maturity

This KPI assesses the maturity level of the technology. Validation with prototype, with simulation and no validation reflect the maturity level in decreasing order.

6.3 Architecture options

The above-mentioned KPIs have been part of the most important tool the project has used to remain focused on its objectives and to derive high quality technologies. The result is a wide set of technologies that, each considered independently, are able to bring important improvements over specific challenges. On the other hand, it was also important to provide a coherent set of architectures where selected technologies would be able to operate and cooperate together to cover the widest set of challenges but most importantly to achieve the highest level of efficiency.

To that end, the number of covered KPIs and the level of maturity have been the two main aspects that help introduced a ranking between the technologies and that have influenced the definition of the MEVICO architectures.

6.3.1 Technologies

The technologies selected to address the above mentioned challenges and scenarios are described in this section.

6.3.1.1 Wireless Mesh Network (WMN)

WMN is a high bandwidth communications network made up of point-to-point communications links organized in a mesh topology providing a virtual transport service for a set of eNBs. The technology is not sensitive to the EPC topology scenarios presented in this document. It is compliant with lots of transport technologies and combinations of technologies. It is a complementary solution to the existing wireless and wireline backhaul access solutions for LTE and LTE-A.

WMN technology provides many economic and technical advantages for backhauling LTE and LTE-A base stations. The level of utilization of the transport resources can be greatly improved within the mesh coverage area. Overall data throughput and transport connectivity is increased by sharing transport capacity flexibly between the client nodes in the

mesh network. Other advantages include operational easiness, high reliability and flexibility/scalability to adapt to traffic fluctuations and network changes since the network throughput can be dynamically and autonomously optimized. The technology enables a flexible way to enlarge and build the network according to the transport capacity need. It also enables horizontal connections between base stations for fast X2 connections. The in-built SON features simplify deployment & installation, maintenance and network management processes.

6.3.1.2 Transparent Interconnection of Lots of Links (TRILL)

From a transport point of view, Ethernet-based technologies have several appealing features, notably they allow increasing the available throughput by moving towards lower layer switching and minimizing processing per packet. TRILL combines the advantages of bridging and routing and fully distributed mobility mechanism implemented in the Link layer (i.e. Ethernet). TRILL enables handling mobility in Ethernet when nodes are moving between eNodeBs, which reduces the amount of mobility requests that have to be handled in upper layers (i.e. IP, Transport or Session layers).

TRILL reduces the signaling traffic and reduces the latency to manage mobility, thus increasing the overall capacity of the backhaul network.

6.3.1.3 Ultra Flat Architecture (UFA)

The fully distributed architecture scenario suggests PCC and IMS nodes are distributed in addition to S/P GW and MME. The mobility implies the change of all these nodes. In order to have low impact on handover performance, the UFA concept simplifies the concatenated equipments into a single GW, and adds a proactive step and a network controlled handover execution.

UFA is flat and introduces distributed signaling and data anchors, which are the UFA Gateways (UFA_GWs) and the SIPcrossSCTP GWs (SxS_GWs). This enables to better distribute the traffic load, contrary to centralized anchors. UFA_GWs distribution enables to distribute the S-CSCF and the Application Servers, which enhances their scalability and reduces the delay for accessing Application Servers content.

UFA is a flat architecture based on SIP. As for current mobile networks, it implements IMS and policy control functions. However, it is constituted of a single layer implementing these functions. It reduces the number of node types and interfaces, and only requires distributed and temporary anchors, instead of centralized ones.

The main idea of UFA is to gather as much information as possible into one Gateway and exchange information with another Gateway. UFA contains the I-CSCF, S-CSCF and the HSS nodes and two new nodes that are the UFA_GW and the SxS_GW. The basic UFA_GW function is providing physical connectivity to users (capacity, coverage).

The UFA_GW is the main UFA node; it gathers classical IP-AN nodes functions (e.g. NB, RNC, SGSN and GGSN functions for UMTS), policy control functions, P-CSCF functions, SCC AS functions and new functions. The SxS_GW is in some cases necessary to handle the case of non-SIP native services. UFA performances are measured for services transported over SCTP. When data is lost due to handover, SCTP considers that these losses are due to congestion and retransmits them after a timeout, causing high handover delay and resource use degradation. UFA solves these issues. Its performances are compared to the most known solution handling the mobility of these applications.

In case of data growth, UFA_GWs will be duplicated to satisfy the connectivity criteria.

Most technologies should be applicable on UFA, especially PMIP.

6.3.1.4 Self Organizing Network (SON) solutions in EPC

SON collects automated network management solutions that are operating autonomously in the network. This includes Self-Healing, Self-Configuration and Self-Optimization features. The goal of these automated management solutions is to decrease the costs related to operating of the mobile network. Definition of SON solutions for LTE radio access network is considered as a major issue in 3GPP since Release 8. However, the transport impact of the radio SON solutions and SON solutions for mobile backhaul/transport networks is out of scope in those investigations.

Transport SON solutions and solutions complementary to the radio SON features enables the efficient utilization of the mobile backhaul resources. For example, an additional feature making the Mobility Load Balancing algorithm's operation transport aware can avoid radio SON actions being optimal for the radio part but at the same time being non-optimal from the backhaul point of view. The self configuring and self optimization algorithms operating on the EPC nodes and/or in the mobile backhaul aim at improving resource utilization and shall cope with issues like:

- Multitude of alternative transport solutions and options
- High number of transport and radio parameters
- Dimensioning/planning based on measured or predicted traffic/load
- Parameter configuration based on guidelines and recommendations

- Static parameters not capable to adapt to the changing conditions

These listed examples result in non-optimal system operation, inefficient resource usage and difficult management. To overcome these issues the SON features defined for transport networks shall provide solutions for the automated tuning of transport related parameters and automated connection setups in the radio access and transport devices. The scope of the automated self-configuration and optimization algorithms in centralized and distributed SON operation scenarios is to reconfigure or adapt the system configuration in order to follow the changes in traffic/load. These algorithms and solutions should ensure the consistent, efficient, adaptive and optimized configuration of mobile backhaul.

Different SON architecture variants

Centralized SON architecture

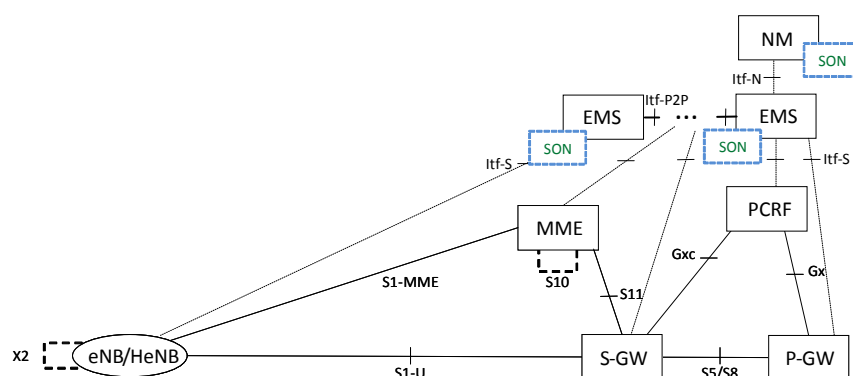


Figure 8 Centralized SON architecture

Distributed SON architecture

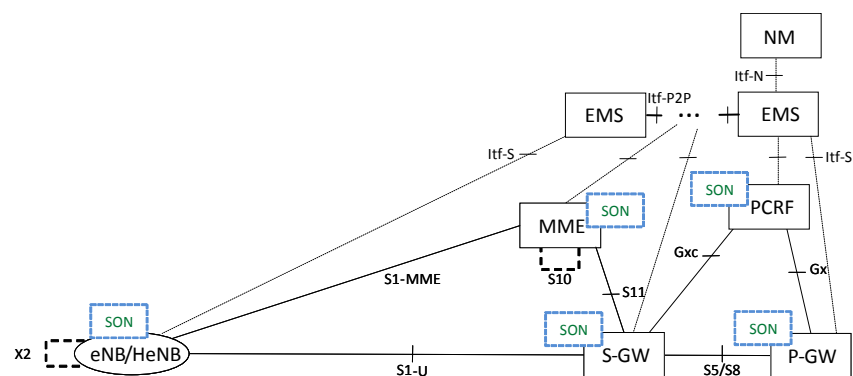


Figure 9 Distributed SON architecture

Hybrid SON architecture

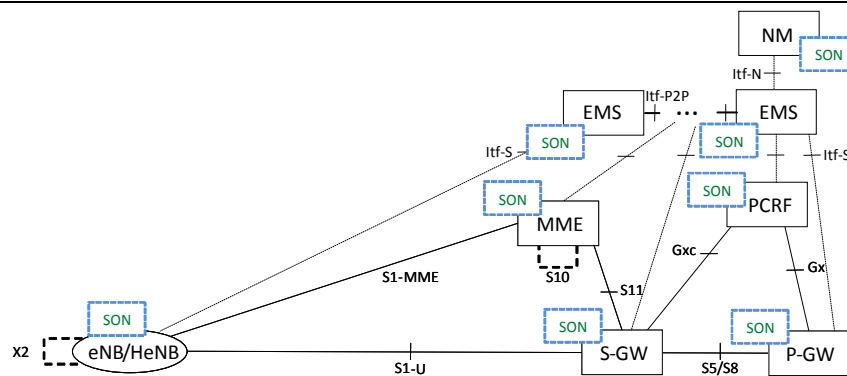


Figure 10 Hybrid SON architecture

6.3.1.5 Wi-Fi

Wi-Fi represents the most appropriate technology for offloading the wide area radio network towards the fixed wireless network. The technology is not topology sensitive in the sense that as long as there is a defined anchor point it will work. Operator can provide personal connectivity services for devices in residential network, e.g. firewall, content filtering, secure authentication. Operator partner services tied to mobile subscription can be provided also over WLAN behind RGW, e.g. Spotify. Operator (Fixed, Mobile or ISP) manages the Wi-Fi AP. Mobile Operator can provide better indoor coverage for Wi-Fi enabled devices.

6.3.1.6 Stream Control Transmission Protocol (SCTP)

SCTP provides a connection oriented reliable service and congestion control services, like TCP. Additionally it provides multistreaming and multi-homing that provides resiliency in case of path failure. It could be useful in case of data losses due to a mobility handled by EPC.

An extension m-SCTP could replace the PMIP, MIP or DSMIP protocols used in the EPC architecture, to handle user mobility, to bring a gain on the signalling plane.

It does not use any centralized server, and has few impacts from the topological choices but works better for flat architecture.

The main rationale for using SCTP as the transport protocol is the main features of the protocol. SCTP inherits most of its features from the most predominant reliable transport protocol on the Internet: the Transmission Control Protocol (TCP). Like TCP, SCTP provides a reliable, ordered transport service ensuring that data is transmitted across a network without error and in sequence. Furthermore, like TCP, SCTP provides connection-oriented communication and mechanisms to control network congestion. Prior to data transmission, a connection or, as it is called in SCTP parlance, association, is setup between the two communicating endpoints, and it is maintained during their entire communication.

One of SCTP's novel features is multi-homing. Multi-homing enables the endpoints of a single association to support multiple IP addresses. Each IP address is equivalent to a different network path towards the communicating peer, for sending and receiving data through the network. Currently, SCTP uses multi-homing as a means for path-level redundancy to provide uninterrupted service during resource failures, and not for load balancing.

Another novel feature that SCTP provides is the Dynamic Address Reconfiguration extension which leverages SCTP with mobility support. The Dynamic Address Reconfiguration extension enables SCTP to dynamically add an IP address to an already existing association, dynamically remove an IP address from an association, or dynamically request to change the primary destination address of the peer endpoint during an active SCTP association. Moreover, a local SCTP endpoint can influence its incoming network interface, by advising the peer endpoint about the destination IP address it should use for data transmission.

SCTP is to our knowledge the only transport layer protocol that actually supports a layer other than the application on top of it, and has a special field in its header that indicates the next protocol to receive the data. Other transport protocols like TCP and UDP do not provide this option and always assume that the next layer is the application layer.

SCTP requires support on the UE or the application, and does not need modifications on the network side. It can provide end-to-end anchorless mobility. Its performance is independent of the architecture.

6.3.1.7 Access Network Discovery and Selection Function (ANDSF)

ANDSF can be used when either 3GPP and non-3GPP accesses are available or when multiple non-3GPP accesses are available. The usage of ANDSF capabilities is intended for scenarios where access network level solutions are not sufficient for the UE to perform Network Discovery and selection of non-3GPP technologies according to operator policies.

The ANDSF contains data management and control functionality necessary to provide network discovery and selection assistance data as per operators' policy. It consists in sending information to UEs about available access networks and inter-system mobility policy.

Access Network Discovery and Selection Function features help to optimize the selection of the various radio accesses. This can be either 3GPP or non 3GPP (mainly Wi-Fi) accesses. The scan which is the process to determine the available radio accesses and the quality of radio links is costly in time and energy. The radio accesses knowledge may be used to shorten the scan process and prioritized which radio accesses should be scanned first. As an example, it is useless to scan Wi-Fi Access Point where the user has no right to access.

In the 3GPP ANDSF version the operator provides rules, mainly dependent on the location of the UE, giving the radio accesses preferences of the operator. These rules are intended to be merely static.

The IEEE ANDSF version named MIH is based on an information service that allows the same kind of features. Furthermore the event service permits to have a network controlled HO.

This is a key feature for offloading techniques to be used by connection manager.

6.3.1.8 Not Mobile IP (NMIP)

NMIP links terminals such as mobile phones directly with servers, cutting out the need for tunnelling and reducing the network itself to simple switches. NMIP is designed as a light mechanism to provide an anchorless mobility management. It does not use any centralized server, and has few impacts from the topological choices. NMIP implementation will use a rules database that we have put on the MME.

NMIP is an extension of the TCP protocol that solves the problem of the connection break when the IP address of one of the correspondents changes. In a mobility context, it allows to manage the radio interface change. As the reconnection is a fast process, the HO is realized seamlessly. Events such as interface up/down are caught and may trigger change of TCP connections. A rule system is implemented to determine which TCP connections may migrate. This technique may be used to realize automatic offloading when one radio interface is a 3GPP one and the other is Wi-Fi.

6.3.1.9 Multipath TCP (MPTCP)

MPTCP is a modified version of the TCP protocol that supports the simultaneous use of multiple paths between endpoints and no centralized anchor is needed. As a consequence of the multipath the traffic is balanced on the available paths. Fairness is ensured on each path to avoid any starvation on one link. The throughput of the connection is then improved as transmitted data is sent simultaneously on several links. The reliability of the connection is increased as even if one radio link fails, the other links can cope with the data transmission.

6.3.1.10 Gateway selection

The optimum selection and reselection of core network elements might be dependent on the currently selected access network. A function which coordinates the different selection procedures might improve the overall system performance. To support traffic management also in distributed gateway scenarios it could be beneficial to include additional criteria into the GW selection process like:

Load of the transport network, mobility behavior of users (e.g. low mobile, high mobile), access networks supported by the GWs and the UEs...

6.3.1.11 Network-based IP Flow Mobility (NB-IFOM)

The currently standardized IFOM (IP Flow Mobility) solution in 3GPP is strictly UE centric as the operator must firstly deliver the flow routing policies to the UE, and then the UE must provide these policies to the PDN Gateway. Also the ANDSF has no interface to the PCC system, therefore requires other ways to get informed about the updated flow routing policy for a particular UE. NB-IFOM (Network-based IP Flow Mobility) tries to eliminate the above limitations and create an operator centric flow management framework. The advantages of NB-IFOM enable operators to enforce IP flow routing policies without involving the UE first, such making able the PCRF (the central policy control entity) to decide on the flow routing policy based on e.g., the available resources in the network, before signalling the policies to

the UE. The network-based solution is more efficient than the ones that rely on the UE to perform policy acquisition and enforcement: in the current, UE centric standard it is possible that the network context and resource availability may have changed by the time the UE provides the routing policies to the network; therefore the PCRF will not be able to authorize the new flow policies anymore. Such situations can be avoided if NB-IFOM is applied in the architecture.

6.3.1.12 Application Layer Transport Optimization (ALTO)

The IETF ALTO protocol provides guidance to content delivery applications in networks such as P2P or Content Delivery Networks (CDNs), which have to select one or several hosts or endpoints from a set of candidates that are able to provide a desired data resource. This guidance shall be based on parameters that affect performance and efficiency of the data transmission between the hosts, e.g., the topological distance. The ultimate goal is to improve QoE of the application while reducing resource consumption in the underlying network infrastructure.

While flow movements within the EPS can have an impact on the e2e path and its performance, there is no current way for decision elements within an EPS to anticipate it. Therefore it is necessary to find a way to integrate decision functions in the EPS with knowledge at the e2e scope. To improve its QoE for applications such as video download or streaming, the UE may use the ALTO protocol to jointly optimize the user QoE and the usage of EPS resources by providing the UE with information helping it to choose the best possible location from which to download the whole or piece of content while considering path changes within the EPS.

6.3.1.13 Host Identity Protocol – Ultra Flat Architecture (HIP-UFA)

HIP-UFA refers to HIP-delegation service based IP mobility management for HIP/IPsec based data traffic tunneling. The technology proposes a new secure tunneling option for the 3GPP EPC. It provides secure inter-GW mobility and mechanism for network-based GW relocation (similar to P-GW relocation with DMA solutions).

The HIP-UFA technology fits well the use cases where mobility between distributed ePDGs (distributed architecture), X-GWs (flat architecture) must be supported. It proposes HIP-based network access service. This technology is expected to decrease the number of HIP Diet Exchanges or HIP Base Exchanges in case of frequent inter-X-GW handovers when using HIP for user access authorization. It also removes data traffic anchors, the only anchors are the distributed GWs (first IP hop) of the UEs/MRs.

It could also be introduced in the centralized and distributed architecture on the P-GW. However in those cases L3 access authorization and the security overhead caused by HIP Base Exchange between the UE and the P-GW is unnecessary. Indeed, in that case there is no need for L3 authentication and IPsec SAs between the UE and the P-GW, since 3GPP EPC already has its AKA / SIM protocol to authenticate the user through by the MME.

The coexistence of PMIP and HIP-UFA has no benefits, but also has no technological constraints. The same UE could support both technologies in different network domains.

HIP-UFA and DMA technologies are alternative solutions for optimal GW locations but can't be applied simultaneously.

6.3.1.14 Host Identity Protocol – Network Mobility (HIP-NEMO)

HIP-NEMO provides mobility management for moving networks with reduced signaling compared to the case when each LFN should update their IP address. It provides similar functionality to PMIP-NEMO.

HIP-NEMO introduces a HIP extension with the concept of mobile Rendezvous Server (mRVS). The mRVS is required in order to handle moving networks. It is expected to reduce HO related signaling for LFNs in the moving network. This technology should be used when HIP-UFA provides inter-GW mobility management (see section on HIP-UFA) or when the MR is HIP-enabled and uses HIP for user access authorization (see section HIP-auth).

The coexistence of PMIP-NEMO and HIP-NEMO has no benefits, but also has no technological constraints. The same MR could support both technologies in different network domains.

6.3.1.15 Proxy Mobile IP – Route Optimisation (PMIP-RO)

The route optimization solution addresses the problem of centralized mobility anchoring in PMIPv6 to reduce the impact of triangular routing by using intermediate anchors closer to UEs. The objectives are twofold: reduce unnecessary load at the LMA and provide a set of methods that allows transferring the data anchoring role from the LMA to distributed servers. The transfer of role would allow having a moving functionality that would optimize routing within a PMIPv6 domain.

From a centralized to a distributed architecture, the number of heterogeneous RANs should increase leading to a more intensive use of PMIPv6 in the core network, i.e., S5, S8, S2a, S2b, interfaces. Optimized communications between RANs will be improved by using this technology.

6.3.1.16 Proxy Mobile IP – Network Mobility (PMIP-NEMO)

PMIP-NEMO is an extension to PMIP to support the movement of prefixes (not just single addresses) allocated to MRs (and subsequently used for configuring addresses on LFNs) at MR's request. The current procedure of PMIP allows to allocate a "Home Network Prefix" (HNP) to a MH directly connected to the infrastructure. This HNP has to be used on conceptual "home link", i.e., the network link between the MH and the MAG. The typical implementation of NEMO assumes that the MR requests a specific (set of) prefix(es) to be announced to the LFNs. Hence, the requested prefixes are not directly used on the "home link". The support of such procedure is not specified in PMIP, which thus requires an extension.

Besides PMIP-NEMO, HIP-NEMO is another alternative to support moving networks in MEVICO. HIP-NEMO makes use of the infrastructure and procedures of HIP to support LFNs mobility. Here, the modification of the IPv6 address is a trigger to update the mapping between the locator and the identifier of the considered UE. Generally speaking, a PMIP vs HIP comparison can be made: HIP (with or without NEMO extension) can be considered as a solution supporting mobility by involving both ends of a communication flow (MN and CN needs to be modified for HIP), whereas PMIP does not require modifications on CN to support mobility.

PMIP-NEMO and HIP-NEMO can be used on the same network infrastructure but the coexistence may not show any benefit. The support of NEMO by PMIP will hide modification of LFNs' IPv6 prefixes needed by HIP to update its mapping. However, both solutions are strongly tied to the mobility management protocol they inherit. Therefore, a core mobile infrastructure running PMIP will find advantages relying on PMIP-NEMO in the same way than an infrastructure supporting HIP will find benefit relying on HIP-NEMO.

6.3.1.17 Distributed Mobility Anchoring (DMA)

DMA with GTP technology in here intends to optimize the EPC based on the ideas of the IETF DMA, but utilizing existing 3GPP protocols like GTP with as less as possible changes in distributed architecture. The technology principle allows the GWs to optimize the data routing in case, when the existing connections have faced user mobility, over the multiple distributed GW serving areas. This procedure is improved by including the GWs itself in the decision on GW relocation, what is currently done in the control plane node (MME) only. Furthermore the relocation for active mode devices can be allowed (by user activity detection, e.g. DPI). A new IP address and service interruption can be acceptable from application point of view during the inactivity and a reconnection can be forced, that allocates a new more optimal PGW and new IP Address.

In a centralized only architecture these optimizations are not needed. In the flat architecture the UFA GW can be seen as a central GW. If applying GW functions in the flat architecture in the eNodeB the proposed solutions may result in unproportional high signalling overhead. Hence it works best in a distributed architecture with distributed GWs.

Further benefits can be achieved if the network control functions are even more centralized and the distributed GWs contain less functionality and can contribute to further savings in HW spending.

6.3.1.18 DPI

Deep Packet Inspection (DPI) is a networking technology that involves the process of examining the header and payload content of a packet. Most DPI systems identify communication streams and maintain state information for large numbers of concurrent packet flows. DPI monitors traffic and is able to provide application, transport or network flow level measurements in a non-intrusive way for technologies enforcing flow mapping and requiring, for instance, probing the transport links in the backhaul and core (i.e., the S1-U, S2, S5/S8, Sgi interfaces). DPI enables diverse operations including: advanced network management, improving network security functions and monitoring customers' data traffic in order to mediate, e.g., its speed. Initially, DPI (combining stateful intrusion detection and prevention) was used to help tackle harmful traffic and security threats and to throttle or block undesired or "bandwidth hog" applications. This role has evolved very fast, including in the mobile sector, where DPI can be deployed for a wide range of use cases aimed at helping to assure and improve the performance of individual customer services and improve customer quality of experience. Based on its potentials, DPI has become a key component in modern network monitoring systems.

Different network elements such as PGW, MME, SGW and eNB all require or can profit from DPI to ensure policy enforcement, lawful interception, QoS, billing, network management, security control and introduce new functionality such as management of differentiated services through fine grained real time bandwidth analysis and content classification, fine-grained diagnosis of network bottlenecks, application-based billing, QoE analysis. In LTE, some of

these functionalities, such as policy control, are mandatory and require DPI functionalities for performing traffic classification.

Network data collection can be carried out with passive network probes (i.e., non-intrusive) that are connected to specific connection points in different interfaces. As data collection usually includes user plane traffic, network probes must be able to handle heavy traffic loads and perform smart data filtering. The DPI based monitoring solutions can adapt to both centralised and distributed architectures and be integrated to network management systems, SON functionality, CES... Limitations in the use of DPI are mainly due to: heavy traffic analysis costs; encrypted traffic allowing only statistical payload analysis; and, legal issues (e.g., network neutrality, differing regulations).

6.3.2 Architectures

As it is mentioned above, the technologies are selected based on the KPIs. And there are three topologies proposed in the section 6.1 Topology models, centralised, distributed and flat. This section introduces the mapping of different selected technologies on three topologies. Each diagram contains the selected technologies which can be implemented on the correspondent architecture topology. However, the technologies in red cannot work simultaneously and the coexistence issue is addressed in the following section.

Centralized Architecture

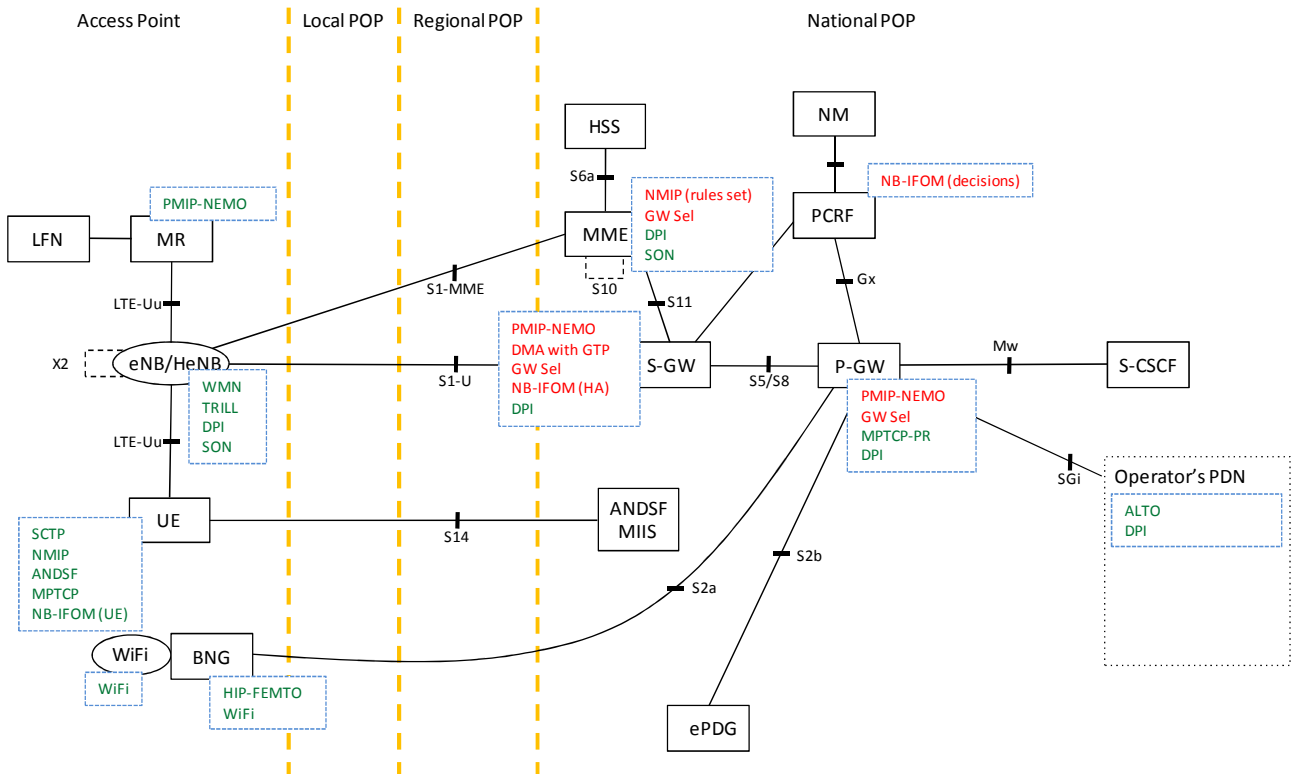


Figure 11 Centralized architecture with selected technologies

Figure 12 Distributed architecture with selected technologies

Figure 13 Flat architecture with selected technologies

6.3.3 Technology Coexistence analysis

This section analyzes the coexistence problems that might appear when deploying some of the listed technologies simultaneously in the same architecture. These technologies are bound to a specific IP tunnelling options, and provide different ways for IP mobility management. These technologies hence mutually exclude each other and provide different options for the MNO's network service layer. The maximum granularity of co-existence of these technologies is UE-level, but often it is an MNO-level choice.

6.3.3.1 UFA-SIP

Key features – radio independent architecture, QoS-integrated/very efficient service establishment and mobility procedures, SIP-based but support any application type

In UFA, almost all network functions/layers are within the same node (UFA_GW). Thus all user related contexts are within the same node, making decisions easy to perform (example adapt QoS needs to available resources during service establishment or mobility) and context transfer during mobility very fast.

UFA-SIP is based on SIP protocol and is compliant with IMS. SIP allows, among other things, to give information about the requested application in a simple way. This allows to get rid from complex mechanisms like DPI.

UFA-SIP applies to any application (SIP native and non-SIP native). Non-SIP native applications have to implement specific APIs, making them convertible/manageable with SIP. The current solution covers support for SIP-based and SCTP-based applications, and not TCP-based applications.

UFA-SIP is complete network architecture. It is radio-independent as its procedures are based on L3 and above protocols. In terms of AAA and security, it implements the same features as a 3GPP network. Better than that, it optimises current 3GPP attachment and authentication procedures.

Deployment requirements

UFA-SIP is a new architecture, constituted of the UE, the UFA_GW, and the SxS_GW. The last node is necessary for managing non-SIP native applications towards Corresponding Nodes not able to convert SIP-native applications to non-SIP native ones. Non-SIP native applications have to implement APIs making them convertible/manageable with SIP. Note that this "constraint" is similar for the ANDSF solution, where applications within the UE have to implement APIs allowing them to interact with the local connection manager.

Preferred topology

As already said UFA-SIP is a new architecture. The UFA_GWs can be in the local or regional PoPs (meaning the distributed/flat topologies). The appropriate topology will depend on traffic previsions, equipment capacities and tech eco aspects.

Co-existence issues

In general, we recommend using only one technology at the same time. The choice of the technology depends on the expected objectives. In case it is needed to activate other technologies with UFA-SIP, a special attention should be taken with regards to the following aspects: conflict in terms of handover decisions, security threats of a given solution regarding UFA-SIP, etc.

6.3.3.2 UFA-HIP

Key features – load distribution, mobility management, security, support of any application type

Key features in focus of the MEVICO project of UFA-HIP technology are seamless intra and inter-GW handover on service data flow level using the Host Identity Protocol (HIP), IEEE 802.21 Media Independent Handover protocol in network-controlled mode, HIP signalling delegation services and context transfer protocol. Depending on the distribution level of the GWs, the load distribution can be ameliorated in the network.

Unified security service is provided in the entire network using IPsec, independently from the access network. IPsec ESP tunnel between the UE and the GW and between the GWs shall provide integrity protection, message origin authentication, confidentiality, anti-replay protection on L3. Note: integrity protection, message-origin authentication and anti-replay protection is not currently provided in 3GPP-access for user plane traffic. HIP-DEX AKA hence adds this security service in 3GPP-access. In Untrusted non-3GPP access the security level remains the same as it is currently in the 3GPP standard. In Trusted Non-3GPP IP Access and in 3GPP-access L2 security services are provided as described by the 3GPP standard.

Other benefits from HIP/IPsec tunneling are the support for legacy application that do not implement mobility nor security, and support of coexistence of IPv4 and IPv6 network segments, transparent for UEs and applications.

Any application can use the HIP/IPsec transport. Good for applications requiring uniform security, mobility management from the mobile service layer. HIP in general has one important restriction, i.e., non HIP-enabled UEs/CNs should not be connected to HIP-enabled UEs/CNs due to security reasons. However, should it still be the case, there exist HIP proxy solutions, but raised security threats must be analyzed.

Deployment requirements

The deployment of the technology requires changes in existing network elements.

HIP daemon extended with support for signalling delegation services must be added in the UE and GW. HIP-enabled DNS service may be added for name resolution to Host Identity Tags of the application servers. AAA server functionality should be added depending on the preferred network access service protocol. In case of resource-constrained UEs HIP DEX-AKA might be a good solution, which requires AAA functionality located in the GW. HIP Rendez-vous service may be added to provide initial reachability.

For security reasons, HIP-enabled UEs and GWs should not run applications that use other tunneling options letting some applications to bypass the IPsec firewall. I.e., UFA-HIP technology should be used by specific applications requiring IPsec protection, with UEs/MRs that only use HIP/IPsec tunnelling option.

Preferred topology

The distributed architecture is the preferred topology because of the trade-off between load distribution and CAPEX/OPEX. Flat network topology is also supported. In case of centralized topology, the technology could be beneficial, if there are multiple S/P-GWs or ePDGs.

Co-existence issues

UFA-HIP architecture proposes a new HIP/IPsec based tunneling option for the EPC, hence it can not be used with other tunneling options, such as GTP, DSMIP or PMIP-based tunnels.

6.3.3.3 NB-IFOM

Key feature – network-controlled IP flow mobility

The main goal of NB-IFOM technology is to provide load balancing and to improve the congestion less state of the network. The decision is based on information, such as bandwidth, packet loss and latency of service data flows or network links. The flow mobility control is initiated by the Home Agent near/in the P-GW. Frequency of decisions can vary from a couple of seconds up to several minutes. This might be defined by the operator or manufacturer. The PCRF makes the decision based on some PCC rules defined for service data flows, or based on subscription data. The Home Agent in the P-GW executes the flow mapping decision.

NB-IFOM is applicable to DSMIPv6 based tunnelling option, i.e. when UEs get IP connectivity through the S2c interface. Hence it cannot be applied when other tunnelling options are used by the UE, such as GTP on S1 and S5/S8, GTP on S1 extended with PMIP/IP GRE on S5/S8, GTP on S2a/S2b, PMIP/IP GRE on S2a/S2b.

The information for the decision making on the mapping of service data flow to a new UE interface, hence to a new path towards the Home Agent (HA), is collected from the following services. The network management system shall provide performance measures with network management granularity, such as transport link utilization. DPI or BAT shall provide performance measures with service data flow granularity.

Deployment requirements

In order to perform network based and traffic management oriented flow-mobility operations the existing Mobile IPv6 (NEMO/MCoA) architecture must be extended with the following special nodes:

- **Measurement Unit:** One or more DPI capable (Deep Packet Inspection) devices throughout the core network. They passively monitor the overall and flow based network usage statistics for a given link. When DPI is not available, i.e., when the tunnelled traffic is encrypted, it reports only aggregated statistics on a given link.
- **Mobile Node/Router (MN/MR):** Mobile IPv6 node with extended functionalities. Perform policy routing and flow binding based on network events. Such policies received from the Mobile IPv6 network management entity (Home Agent) always overrule the local decisions and predefined settings.
- **Home Agent (HA):** Mobile IPv6 central management entity with extended functionality. Relays and enforces network-based policies received from the Policy Server. Synchronizes its Binding Cache to the Policy Server.

Policy Server (PS): A single central entity which performs flow binding based on overall network parameters. It receives link and flow usage information from multiple Measurement Units and maintains an aggregated Binding Cache from multiple Home Agents. Knowing the actual flow binding usage on the network it activates policies when trigger conditions are met.

Preferred topology

The most preferred topology for DSMIPv6 based NB-IFOM technology is the centralized topology. Still in a distributed topology, the technology can have benefits by balancing the load between different paths within the domain of the same

distributed GW. However note that the technology does not deal with inter-GW flow mobility and service continuity for inter-GW mobility scenarios.

Co-existence issues with other technologies

This solution is associated with the IP tunnelling option that uses DSMIPv6 extended with multiple care-of address support for IP mobility management, and IPsec tunnels or IPv6-in-IPv6 between the UE and the Home Agent. Hence it provides an alternative to the other technologies described in this section.

Note that DSMIPv6 establishes IPsec tunnel between the UE and the HA on the S2c interface. Hence a dependency issue is that DPI cannot get service data flow level information, except the rare case when null-encryption algorithm is selected for data protection.

GW selection is a higher layer decision system and it could be mapped to any tunnelling solution (e.g., GTP, PMIP, DSMIPv6). GW selection provides triggers to tunnelling protocols about where to map a given traffic flow. Since GW selection is located in the MME, a network-initiated IP-connectivity tunnel update procedures fit better to it than UE-initiated procedures.

In case of PMIP or DSMIPv6-based NB-IFOM, the GW selection could take effect only during the initial session establishment procedure, e.g., selection of LMA/HA for the UE, because these NB-IFOM technologies do not support GW change for on-going sessions. E.g., the NB-IFOM provides network-controlled IP flow mobility using DSMIPv6 based tunnelling (which is orthogonal to the other tunnelling solutions used in MEVICO), GW Selection comes into picture during HA selection for a UE (based on GW load, and probably other parameters of the network). After selecting the HA, NB-IFOM can make more fine grained distribution of service data flows, and may not change then the GW. For In order to support distributed architecture, inter-GW mobility extension should be added (e.g., Global Home Agent to Home Agent protocol [GHAHA]).

6.3.3.4 DMA with GTP

Key features – Distributed Mobility Anchoring principles with GTP tunneling

DMA with GTP technology optimizes the data routing, when the existing connections have faced user mobility over the multiple distributed GW serving areas.

Deployment requirements

Technology utilizes the already defined 3GPP procedures, like re-establishment of the PDN connection request. The network is in full control over the usage of the local PDN connection and explicitly triggers the UE, when to request a new PDN connection and IP address. The impact on standardization and further on the deployment is low: A new cause codes for PDN connection release messages have to be defined for GW to MME interface.

Preferred topology

Technology benefits rely on the multiple P-GWs in the area, so it works best in a distributed network topology with distributed GWs. In centralized GW case the scenario of the un-optimized PGW routing is not very relevant. Further benefits can be potentially achieved, if the GW control functions would be centralized and the distributed GWs (with pure user plane functionality) are controlled with the Open Flow principles.

The functional overlappings and possible co-existence issues with similar functionality technologies UFA-SIP, HIP-UFA, PMIP-RO and NB-IFOM (DSMIPv6 based) are likely and most probably it doesn't make sense to include these functions simultaneously in the network.

Co-existence issues with other technologies

In general 3GPP has assured the coexistence of PMIP and GTP. The question of co-existence has to be looked more for the suggested technologies and concepts rather than looking for PMIP-GTP co-existence only.

PMIP-RO proposes tunnelling the traffic between MAGs (SGWs), bypassing the LMAs (PGWs) and traversing over an intermediate anchor (IA) developed for the PMIP protocol. DMA with GTP proposes to change PGWs using intelligence in the PGW or changing SGW for routing optimization.

The DMA proposals with PGW relocation may not coexist with PMIP-RO as they provide different solutions for the routing problem: The PMIP-RO solution provides tunnel modification while keeping the UE IP address, the other solution is to select IP (PDN) connections in the PGW for what a new IP address and service interruption may be acceptable from application point of view and force a reconnection that allocates a new more optimal PGW and new IP Address. It is clear that these are alternative solutions for optimal GW locations but can't be applied simultaneously.

The proposal to relocate the SGW to achieve maximal SGW-PGW collocation and optimal routing (DMA) could also coexist with PMIP-RO, if MAG changes are possible and may also result in MAG-LMA collocations.

In principle the NB-IFOM would not directly conflict the functionality of DMA with GTP, because NB-IFOM operates on the finer granularity (IP flow level) inside the single Packet Data Network (PDN) connection. It should be ensured that the potential anchoring point change (with DMA) is conformant with the all potentially related IFOM connections.

6.3.3.5 PMIP RO

Key features – Routing optimization, localized routing, offloading

PMIP-RO is an extension to current PMIPv6 procedure to control communications data paths within and/or outside the EPC, i.e., between MAGs and within the LMA's realm. This extension relies on the concept of intermediate data anchors (IAs) located throughout the EPC. In a network setup where MAGs are located in local PoPs and the LMA in a national PoP, the IA function could be located between (or inside) local or regional PoPs. The role of IA could be played by MAGs or intermediate LMAs or other specific hardware having routing capability. Knowing that the P-GW (where the LMA is generally located) has specific treatments to perform on flows (such as charging, lawful interception, or content filtering), it is expected that IAs are able to perform a subset, all, or additional services of what the P-GW is normally expected to be capable of.

The LMA through new signaling messages and for a given traffic characteristic is now able to change, update, or generate a specific data path after selection of one or several IAs. Because traffics are tunneled in the PMIPv6 domain, the resulting data path will be a succession of tunnels between MAGs and IAs. For example, the operator may want to redirect data traffic coming from sensors connected to specific MAG(s) to a specific IA for data aggregation reducing the treatment load at the LMA. In a vehicular scenario, two communicating vehicles along a highway could have their communications redirected to closer IA(s) to gain better jitter performance. One IA could be used temporarily for a UE as data buffering close to the attached MAG in case of radio link disruptions.

Deployment requirements

The deployment of the technology requires modification of existing network elements and protocols. The PMIP's LMA daemon must be updated on P-GW(s) as well as PMIP's MAGs on RAN GWs (S-GW, ePDG, etc.). New network elements may need to be deployed as Intermediate Anchors. Furthermore, current operation of BBERF and PCRF may be extended to handle localized and optimized routing.

Preferred topology

PMIP-RO relies on the distribution of data anchors throughout the network to localize and optimize data traffics. Hence, it is not best fitted for centralized deployments. PMIP-RO will benefit from distributed topologies, though there is a tradeoff to consider with the amount of signaling messages to maintain optimized routing paths during mobility.

Co-existence issues with other technologies

PMIP-RO enables localized routing and traffic optimization within and/or outside of the EPC while keeping the LMA (located on the P-GW) as signaling anchor. This means that any protocols that may change or relocate the P-GW would affect negatively the performance of PMIP-RO. Hence, DMA with GTP should not be applied simultaneously to the same traffics. On the other hand, NB-IFOM is compliant with DSMIPv6, which is not compatible with PMIPv6. Therefore, PMIPv6 and DSMIPv6 (and by extension PMIP-RO and NB-IFOM) may co-exist if the network selects which of the two mobility management protocols would handle the PDN connection or the UE.

6.3.4 Roaming

International roaming is a cornerstone of mobile networks. MEVICO new architecture proposals, specifically the distributed/flat approaches, raise some questions about it. In the following diagram, you will find a state of the art of the 3GPP roaming reference points depending on the various potential cases, knowing that the standardization is not complete so far, and neither the operators' usages. For instance, while 3GPP has specified a number of technical scenarios for both 3GPP and non-3GPP roaming GSMA [10] has defined recommendations for roaming cases involving S6a, S8, S9 interfaces but has not yet defined LTE specific guidelines for non-3GPP roaming (case B1, B2 and D below).

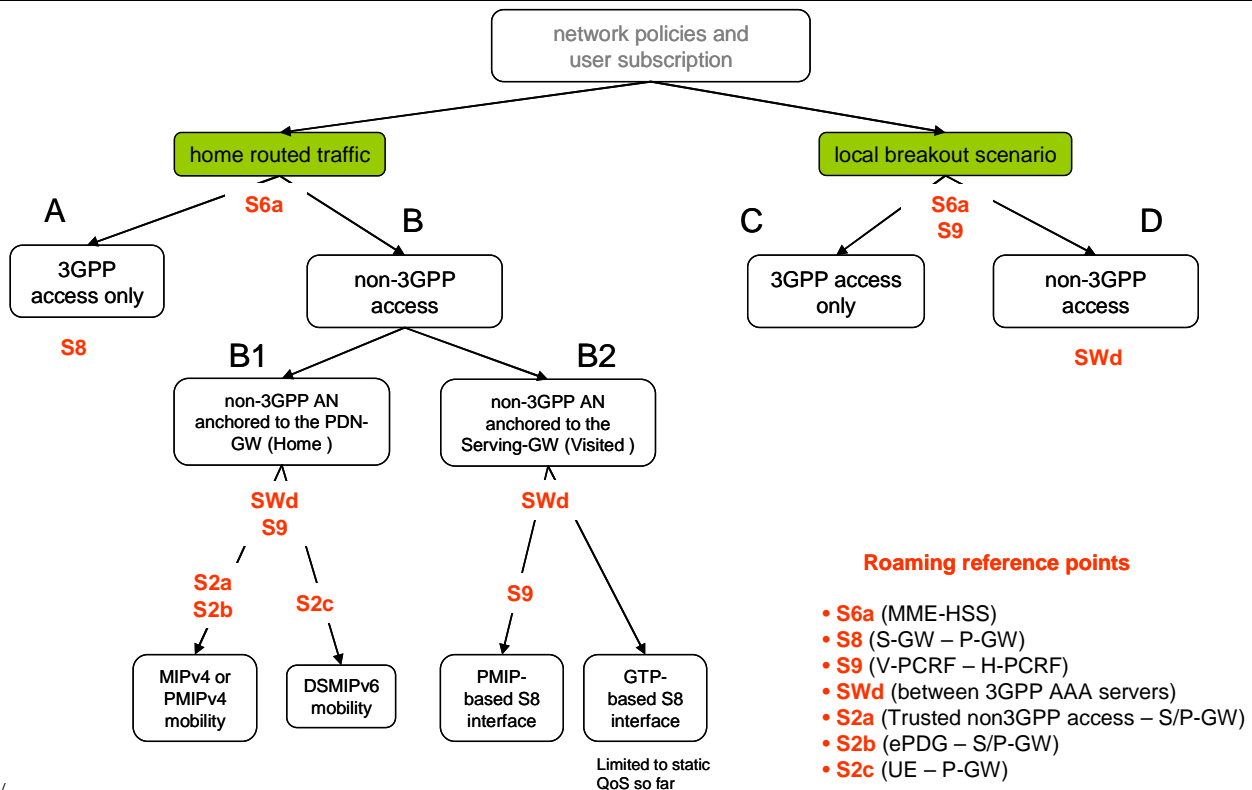


Figure 14 Roaming reference points

For a certain number of cases, including the abroad local calls and the web browsing, the local breakout scenario suits better for a question of optimized routing. Nevertheless, as long as the operators will provide walled garden services, the home routed scenario will be needed, to the cost of a suboptimal routing. A mix of both scenarios could apply depending on subscriptions, services, involved operators' policies and agreements between the operators.

In addition, 3GPP has defined deployment of S14 reference point [5] for both across the home and visited network between UE and H-ANDSF and within visited network between UE and V-ANDSF. For the home routed scenario, ANDSF is deployed at the HPLMN. For local breakout scenario, V-ANDSF and S14 reference point is topology agnostic, but most likely deployed at the national POP.

As a first approach it seems the main point is to minimize the number of roaming reference points. That is why case B1 above is not recommended at all.

In all the cases, the MME distribution appears to be an issue.

In addition for 3GPP accesses, in case A, the serving-GW distribution is an issue. In case C, it is the PCRF distribution that is a challenge.

For non-3GPP accesses, both cases B2 and D show that PCRF distribution is an issue. 3GPP AAA servers are not concerned by distribution.

Both S6a and S9 interfaces are based on Diameter. In order to support scalability, resilience and maintainability, and to reduce the export of network topologies, GSMA has recommended [10] deployment of a Diameter Edge agent at the operator network edge. The Diameter Edge agent is considered as the only point of contact into and out of an operator's network at the Diameter application level. GSMA also recommends to deploy Diameter proxies to provide functionalities such as admission control, policy control, add special information elements (AVP) handling for each application (such as MME) supported by the operator. A Diameter proxy may reside within the Diameter Edge agent. Diameter routing and discovery of the next-hop agent is based on realms. GSMA recommends [10] that in the search order the Diameter Edge agent first consults its list of manually configured Diameter agent locations. Diameter realms can be optionally resolved using DNS.

This means deployment of Diameter based interfaces i.e. S6a interface for MME distribution and S9 interface for various cases for visited PCRF distribution requires quite a lot of effort by the operator for distributed and flat scenarios as pointed above for cases C, B2 and D. In both home routed and local breakout scenarios MME needs to interact with HSS at the HPLMN, thus centralized scenario is best suited from MME point of view in roaming case.

For the home routed scenario, VPLMN operator may wish to deploy user plane anchor point (in general GW, or more specifically S-GW) for S8 interface as close as possible to the network egress point (case A above) to avoid frequent

anchor changes that would be visible over potentially long data path (consider data traffic across a home operator in Europe and visited operator in Asia) across inter-operator network (GRX/IPX) at the HPLMN operator (P-GW). This may even be part of the roaming agreement between operators. The centralized scenario would be better suited than distributed or flat scenario in this case.

In the local breakout scenario, the choice of architecture and topology scenario (centralized, distributed or flat) for deployment of the user plane anchor points would be up to the VPLMN operator decision taking account to the discussion above about Diameter based interfaces, especially if a service requires support for S9 interface.

For all the scenarios discussed above, choice of technology options for particular scenario in roaming case, in addition to technology co-existence discussed in other parts of this document, will depend on subscription specific details (such as roaming restrictions in subscription records), services, involved operators' policies and roaming agreements between the operators.

6.4 Other Architecture options

This section includes the architecture diagrams including other technologies being analyzed in MEVICO that have lower priority but still are under consideration to address some of the challenges and requirements. Those technologies are; Customer Edge Switching (CES), Ethernet Operation and Management (O&M), HIP Authentication, Mobile Relaying and mobile Peer4Peer.

Centralized Architecture

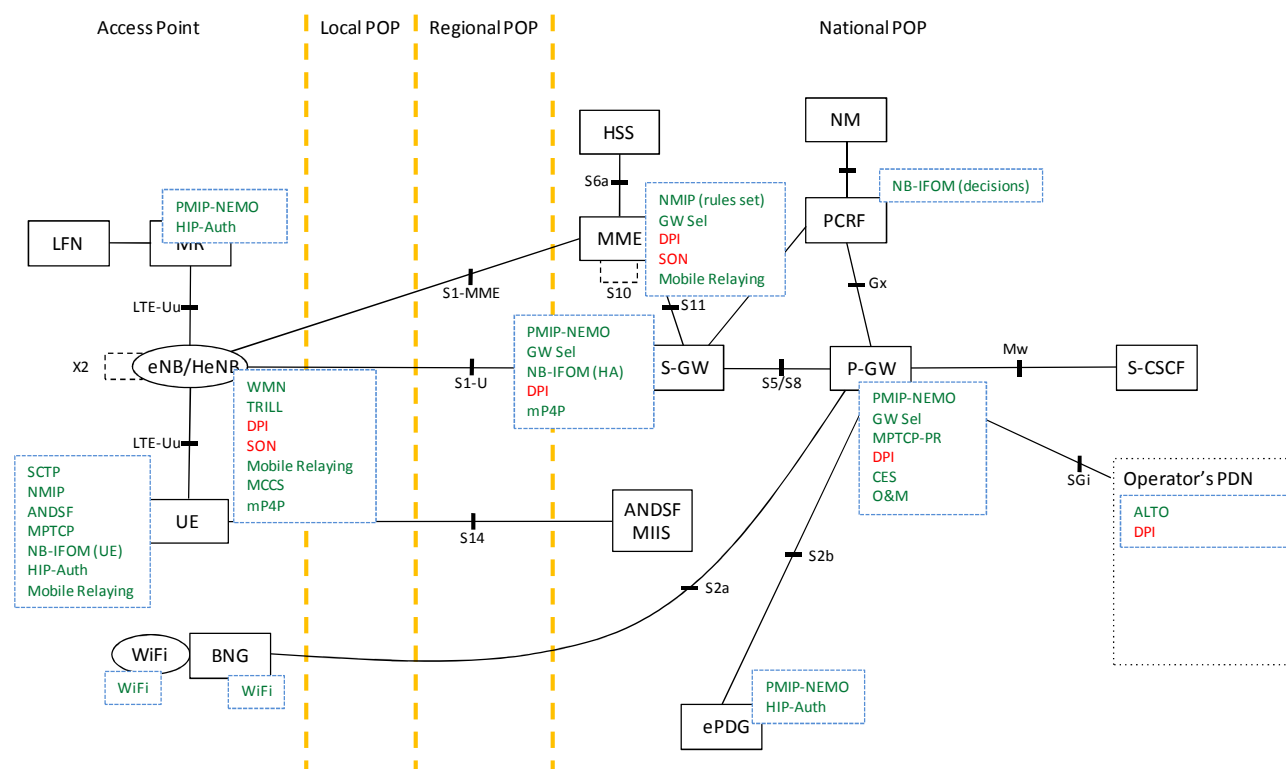


Figure 15. Centralized Architecture with all MEVICO technologies.

Distributed Architecture

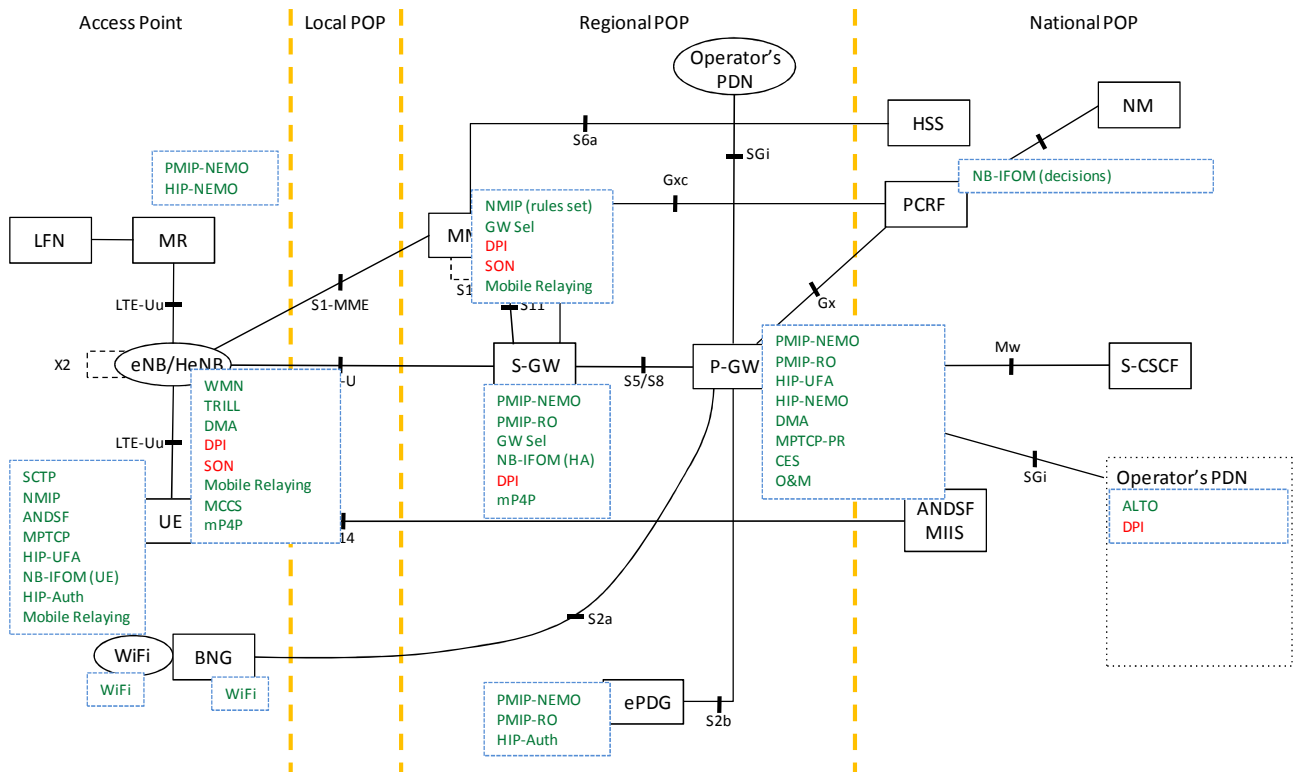


Figure 16. Distributed Architecture with all MEVICO technologies.

Flat Architecture

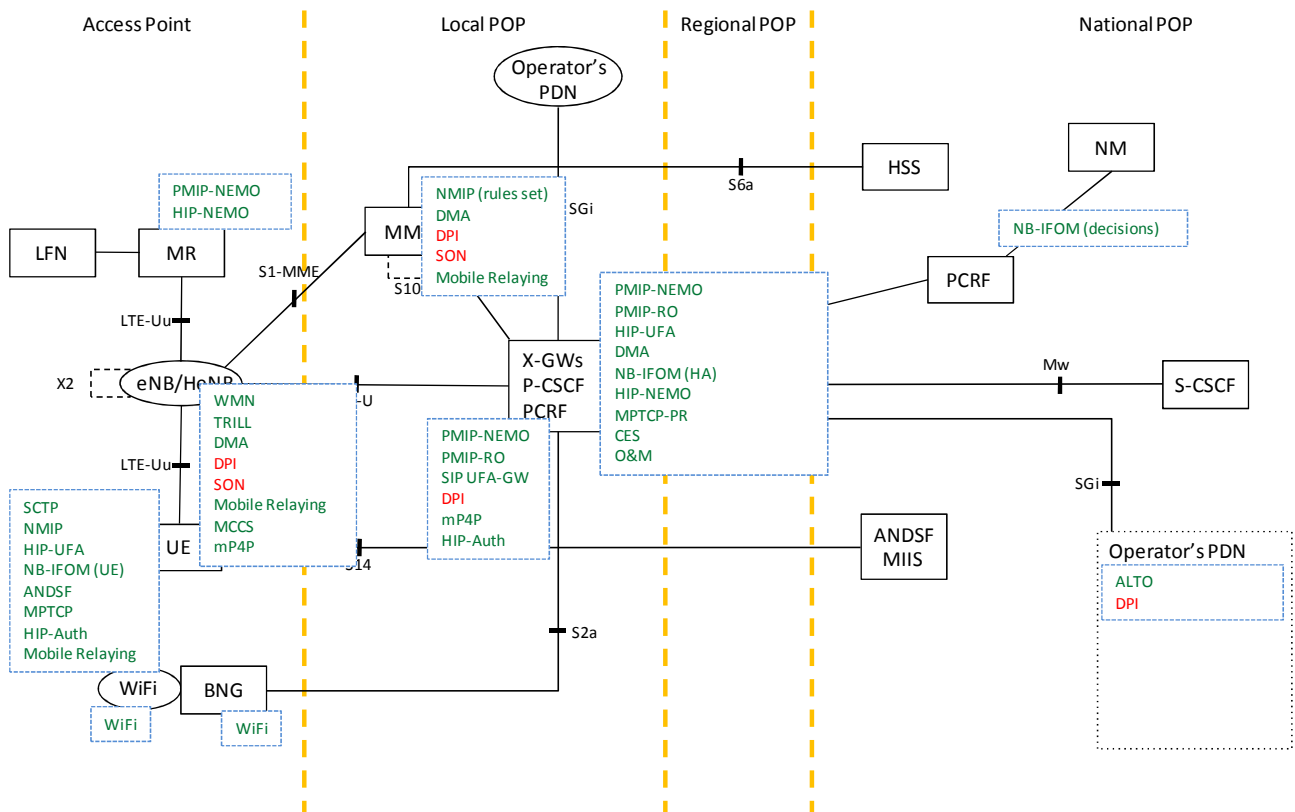


Figure 17. Flat Architecture with all MEVICO technologies.

7 OPEX and CAPEX analysis

This section describes the CAPEX/OPEX analysis for the mobile packet core network for the LTE and LTE-Advanced of 3GPP. The main purpose of the analysis is to compare and evaluate the three main topology evolution scenarios: centralized, distributed and flat. The work focuses onto the following items:

- Centralized vs. Distributed network deployment
- Localization of Internet Exchange Points
- Sensitivity analysis of input parameters within the above models

7.1 Model description

The model proposed for the analysis is inspired from the traffic flow model described in the Traffic Description document:

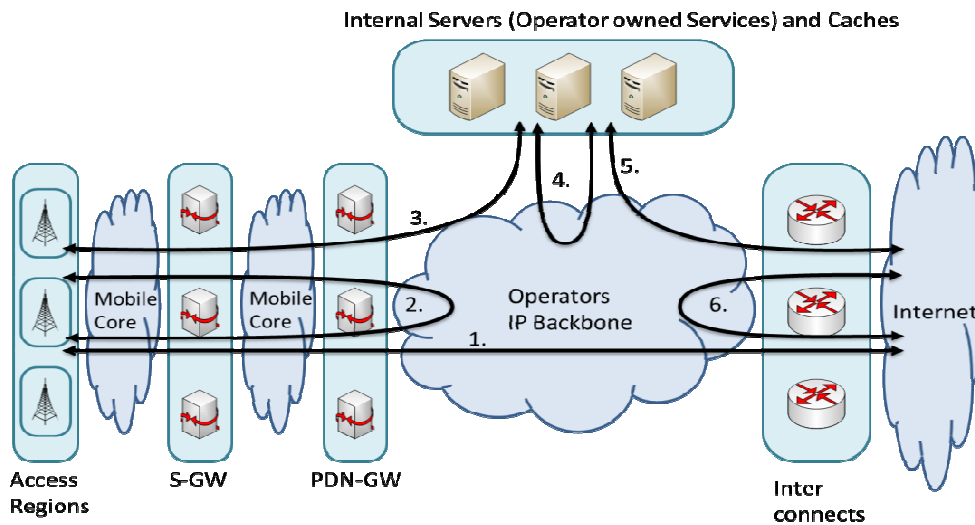


Figure 18: Traffic flow model

The network is composed by 60 POPs for a large size country. Three different types of equipment can be installed on each POP:

- LTE equipment (S-GW, P-WG, Interconnects and Caches)
- Aggregation equipment (interfaces, routers and switches)
- Internal Servers (operator specific applications or services).

An amount of internet traffic is generated at each POP of the network. The total traffic forecast for 2015 (280.52 Gbps) has been proportionally distributed for each POP according to the number of inhabitants of the region. French population data has been used to distribute the locations of the POPs over the country and to determine the traffic generated at each POP. Depending on the architecture a different number of POPs can be equipped with LTE equipment:

- Flat architecture, a minimum of 21 POPs and a maximum of 60 POPs can be LTE equipped.
- Distributed architecture, a minimum of 6 POPs and a maximum of 20 POPs can be LTE equipped.
- Centralized architecture, a maximum of 5 POPs can be LTE equipped.

Two different scenarios are tested considering different architecture network topologies for the Internal Servers. The number of POPs equipped with Internal Servers is limited to 5 and 20. The amount of traffic processed by the Internal Servers is fixed to 20% according to the application share forecast for 2015.

The equipment, installation, and transport costs between the base antennas and the backhaul network are not considered in this analysis. However, it is worth remarking these costs are constant over the three network architectures.

Different transport and equipment capacities are considered (1 Gb, 2.5 Gb, 10 Gb, 40 Gb, and 100 Gb).

In order to satisfy the end-to-end delay time, the average number of intermediate POPs before reaching a LTE equipped POP has been fixed to 2.5. The schema in next figure illustrates the costs considered in the analysis:

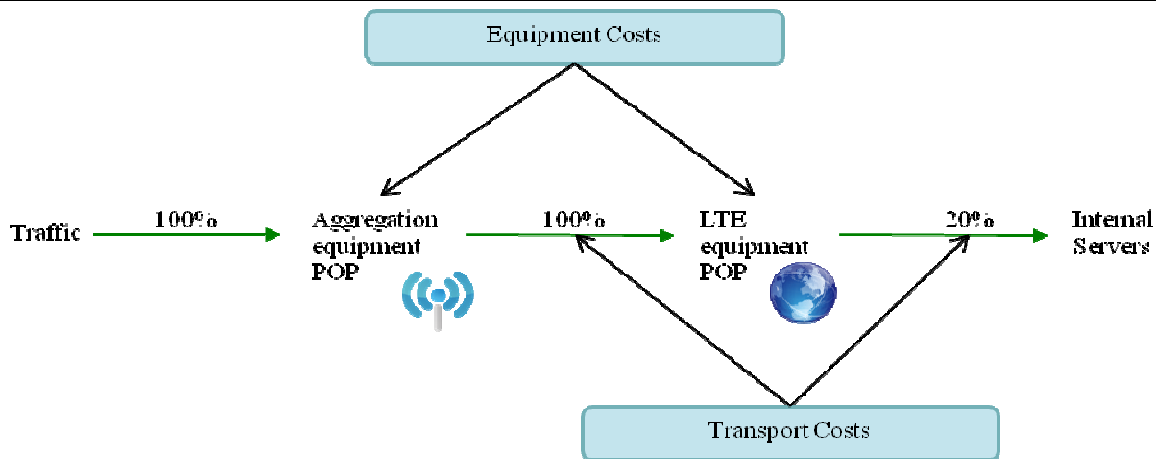


Figure 19: Schema of the costs considered in the analysis

Table 1 and Table 2 display the transport and equipment costs for the different components of the network.

In this CAPEX/OPEX analysis we consider the cost factor as the relationship between the required network capacity and the cost of the LTE, aggregation equipment and transport equipment. Therefore, a cost factor of 10:3 means that increasing the network capacity by 10 times requires an increase by 3 times the cost of the LTE, aggregation and transport equipment.

k€	1Gb	2.5Gb	10Gb	40Gb	100Gb
Internal Servers	0	0	0	0	0
LTE equipment	3000	4645	9000	17438	27000
Aggregation equipment	100	154	300	581	900

Table 1: Equipment costs

k€	1Gb	2.5Gb	10Gb	40Gb	100Gb
New Connection cost (cost/km)	70	70	70	70	70
Cable cost (cost/km)	5	7.74	15	29	45

Table 2: Transport costs

The model determines the best positioning for the LTE equipment and Internal servers by minimizing the costs.

In order to obtain a more realistic network, the current France Telecom network has been considered as connections already installed. It means that the installation cost ("New connection cost") for those lines is 0.

7.2 OPEX and CAPEX results

The analysis of OPEX and CAPEX performed for the different architecture and topologies described in previous Section 6.3 provide the following results for the two evaluated scenarios (5 and 20 Internal Servers).

	5 Internal Servers			20 Internal Servers		
k€	Flat	Distributed	Centralized	Flat	Distributed	Centralized
LTE equipment costs	155438	105521	90000	146456	115314	---
Aggregation equip. costs	19405	21055	22480	18395	19985	---
Connections costs	305196	310669	311589	296820	294866	---
Total cost	480039	437245	424069	461671	430165	---

Table 3: Costs for each type of network topology architecture

Centralized Architecture

The centralized architecture costs have been calculated only for the first scenario where 5 Internal Servers are available. Indeed, the 20 Internal Servers scenario has no sense here since the 100% of the traffic has to be processed by the S-GW and/or PDN-GW and only 20% of this traffic is then directed to the Internal Servers. Thus, the optimal solution here is to locate the Internal Servers at the same POPs where LTE equipment is installed. In the optimal solution only four of the five LTE equipments available have been installed. The four POPs have been spread over the country and the

amount of traffic allocated to these POPs is very high: three LTE equipments with a capacity of 100 Gbps and one LTE equipment with a capacity of 10 Gbps.

Distributed Architecture

The results show that only six of the available LTE equipment has been installed (the minimum number authorized). It means that the cost will increase if a larger number of LTE equipments are installed.

Flat Architecture

The flat architecture has LTE equipments where 21 and 24 over 60 LTE equipments are available. Based on the costs proposed in this analysis, the flat architecture is then more expensive than the other two architectures.

7.3 Sensitivity Analysis

A sensitivity analysis has been performed in order to evaluate the robustness of the solution. The idea is to evaluate how the optimal network architecture evolves when some input parameters are modified. These parameters are mainly related to the costs. From the Table 3 the most suitable network architecture is the centralized architecture, so the sensitive analysis is performed for the distributed architecture (a maximal of 20 LTE equipments are authorized) for the 5 Internal Servers scenario.

Costs

Decreasing the equipment costs

In this first analysis, the analysis focus on difference between the equipment costs and the transport costs independently on the equipment capacity. The LTE equipment costs have been divided by two. The new LTE equipment costs are displayed in Table 4. The rest of costs are not modified.

k€	1Gb	2.5Gb	10Gb	40Gb	100Gb
Internal Servers	0	0	0	0	0
LTE equipment	1500	2322	4500	8719	13500

Table 4: LTE - Equipment costs

The results show that the optimal architecture evolves to a more distributed architecture with more LTE equipments installed (11 LTE equipments are installed) when the ratio between LTE equipment costs and transport costs (aggregation equipment, connections) decreases.

Increasing the equipment and transport cost

For this analysis, the cost factor when the network capacity increase is 10:6 which means that when the the network capacity has to increase by 10 times then the equipment and transport cost increase 6 times. Table 4 and Table 5 display the new transport and equipment costs for the different components of the network.

k€	1Gb	2.5Gb	10Gb	40Gb	100Gb
Internal Servers	0	0	0	0	0
LTE equipment	3000	6120	18000	52937	108000
Aggregation equipment	100	204	600	1764	3600

Table 5: Equipment costs

k€	1Gb	2.5Gb	10Gb	40Gb	100Gb
New Connection cost (cost/km)	70	70	70	70	70
Cable cost (cost/km)	5	10.2	30	88	180

Table 6: Transport costs

The results show that a centralized/distributed architecture with some extra LTE equipped POPs keep being the optimal architecture when the costs of the equipment with higher capacity increased (6 LTE equipments are installed). However, if this cost factor is increased to 10:8 which means that if the network capacity has to increase by 10 times and the equipment and transport costs increased 8 times, the optimal architecture is a hardly distributed or flat architecture with almost 30 LTE equipped POPs.

Increasing the transport cost

Finally, in this third analysis only the transport costs is increased. Table 1 and Table 6 display the equipment and transport costs, respectively.

In that context, the optimal architecture evolves to a more distributed architecture with more POPs equipped with the LTE technology (15 LTE equipments are installed).

7.4 Conclusions

From the CAPEX/OPEX analysis, we can conclude that the results are dependent on the costs. The costs sensitivity analysis demonstrates that if the ratio between LTE equipment costs and transport costs (aggregation equipments and connection costs) decreases (i.e. network capacity increases but equipment and transport cost cannot be kept competitive), then the optimal architecture evolves to a more distributed architecture.

The second interesting point is that if the cost of the equipment with higher capacity increases, then the optimal network architecture also evolves to a distributed or flat architecture with a higher number of LTE equipments with lower capacity. Contrarily, if the cost of equipment with higher capacity is cost effective then it will be then more interesting to use a centralized architecture with a few number of LTE equipments with higher capacity.

The traffic forecast in the scenario based on mobile network in France the traffic is expected to grow in 2020 about 2300 Gbps, so the capacity of the equipments installed at the POPs for the centralized architecture will be huge (500 Gbps and 1000 Gbps). In that context, if the equipment with these capacities is not technologically feasible by the time, the solution is then to double the equipment or connections in order to get the necessary capacities. In that case, the cost will be considerably increased for the centralized architecture network since the cost factor (10:3) cannot be maintained. Besides, new challenging technology problems such as energy consumption or heat dissipation may appear increasing also the final cost.

Another factor not considered is the network deployment since networks cannot be installed over night when capacity increase is required. When operator deploys a system it has to consider the expected traffic forecast so the equipment installed should be able to handle that traffic in the next years. However, that means the equipment is underutilized in the initial years. This supports the idea of having more distributed system that can be deployed, extended with the traffic demand. However, adding new equipment in the POPs might require having an Internet Exchange point which has some additional cost due to contractual and legal negotiations with Internet service providers and other exchange carriers in the scenario where the mobile operator is not owning the fixed Internet network infrastructure.

The following table summarizes the advantages and drawbacks of the different topologies depending on different techno-economic factors.

Main Factor	Centralized Architecture	Distributed Architecture	Flat Architecture
Cost scaling factor when capacities increases	<i>If factor is kept competitive so the LTE technology can handle the required capacity maintaining a low cost scaling factor (10:2 to 10:6), which means the network capacity increases by 10 times but the cost of the equipment/transport increases by 2-6 times.</i>	<i>If the transport network is not able to keep up with the required capacity with a cost effective solution with cost ratio (10:7 to 10:8)</i>	<i>If there exists a technologic limitation: it is not feasible to have very high capacity components (cost factor 10:9 to 10:10)</i>
OPEX costs (maintenance, power supply)	<i>Lower OPEX costs to maintain few POPs and the associated real estate</i>	<i>Higher OPEX costs to maintain a larger number of POPs spread over the country (mobility)</i>	
CAPEX extra costs (cooling equipment, renting contracts)	<i>Possible additional infra requirements (even not technologically feasible for very high capacity equipment) concerning cooling limitations</i>	<i>This requires additional footprint in the current POPs which might lead to renegotiation of the renting contracts with the associated cost in order to install new equipment which require additional space.</i> <i>Additional infra requirements concerning cooling limitations in the LTE equipped POPs.</i> <i>The POPs might require Internet Exchange points with the required technical and contractual/legal cost</i>	

Network deployment	<i>Network deployment not optimized based on the expected traffic forecast (ressources underutilization/network saturation)</i>	<i>Optimized network deployment adapted to the real evolution of the traffic demand based on the geographical location. Thus, new equipment can be deployed only in selected areas with higher demand.</i>
--------------------	---	--

8 System validation

8.1 Validation Usage scenarios

The usage scenarios allow discussing the most relevant parts of signaling and data communication. Moreover, this enables the possibility of investigating the dependencies and the collaborative effects of proposed technologies on the system validation KPIs.

After investigating the challenges covered by different usage scenarios, and the challenges covered by MEVICO contributions, three main usage scenarios have been selected for the decision on the final architecture option.

1. Increasing number of mobile users that will use VoIP or other multimedia services over mobile internet. The MNO should guarantee E2E QoE for user traffic while accessing the service over residential Wi-Fi, home Wi-Fi or LTE wide area network.
2. Increasing number of mobile users that will use Premium VoD with guaranteed QoS. The network should support offload to Wi-Fi provided by Fixed broadband provider, QoS update based on available bandwidth, user-specific network policy enforcement.
3. Improve the scalability of the MNO in terms of better resource utilization and cost-effective network. Particularly, provide seamless user experience of mobile Internet over multiple GWs and multiple interfaces.

The first two usage scenarios are about the most important end-user services, i.e., the subscriber accesses Voice/Multimedia communication and Video on Demand services. The proposed architecture must meet the requirements of fixed-mobile convergence where the MNO provides the same services to the user over fixed or mobile networks regardless of the access and location. The access network could be the home mobile network, the roaming mobile network, home Wi-Fi network, residential Wi-Fi network. Both usage scenarios describe the mobility of a user between residential WiFi, LTE wide area network and home WiFi. In both scenarios the most important requirement is the support of E2E QoE with suitable QoS adaptation.

The third usage scenario highlights the mobile network operator aspect and architecture evolution. Seamless user experience of mobile Internet should be provided for the users, even if the architecture is changed from centralized to distributed. The user traffic can be conveyed over multiple GWs and multiple UE interfaces. Transport, traffic and network management should be able to support smart traffic steering and automatized functions. The most important challenge is to improve user experience by taking advantages of multiple interfaces and multiple paths to the servers, application-level awareness of endpoints and network utilization awareness. Provide a better user experience of mobile Internet by managing connectivity towards multiple access of the end device. Upon events, the system may decide to move some flow between accesses in order to increase QoE or for load sharing purpose. There are a lot of possible events, hereafter is a non exhaustive list for example: UE detects the coverage of a WiFi AP or a femtocell access and loss of access link, QoS of current access degrades, QoS of preferred access improves, new application starts, etc.

8.2 Validation results

This part discusses the performance gains of the MEVICO technologies in terms of the system validation KPIs.

First, the logical structure of the proposed technologies is described. Technologies are positioned in terms of their place and influence in the network segments and layers of the MNO. Technology usage alternatives are presented for situations where the supplied functionalities are overlapping, and hence there are multiple options to convey the user traffic.

Then the functionalities and the performance gains achieved by the MEVICO technologies are explained based on the KPI assessment results. In the validation multi-access environment is assumed where UEs are moving among hierarchically structured LTE networks and between LTE/WiFi access networks meanwhile users are connecting to different applications including videos, P2P, SCTP-based applications, SIP-based applications. The business benefits of the technologies that are close to market are also mentioned.

Finally, we discuss how to avoid conflicts between the decisions of different technologies that enforce traffic steering policies for group of users, individual user, group of flows, or an individual flow. Figure 20 shows the logical structure of technologies.

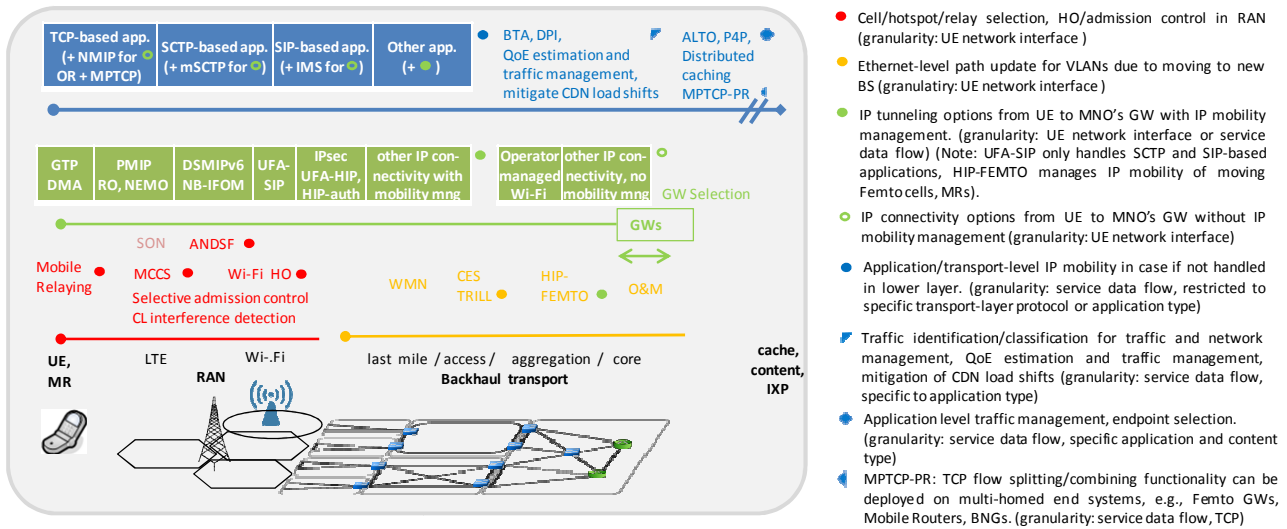




Figure 20. Logical structure of MEVICO technologies.

The technologies can be grouped into the following categories regarding their position in the network layers and segments, and their influence on network resource utilization; RAN layer optimization, Transport layer optimization, Mobile service core optimization and application level optimization.

- **RAN layer optimization** (combined with backhaul): red technologies influence radio access network selection with user-level granularity, and include RAN-layer traffic management, network management. These technologies determine 3GPP cell / Wi-Fi hotspot selection for the UE's network interfaces (as indicated by ●) hence optimize RAN and backhaul network resource utilization.
- **Transport-layer optimization** (on different backhaul network segments or core transport): yellow technologies improve the transport network layer's capacity, increase the transport connectivity possibilities, provide capacity improvements, fault-tolerant transport. RAD distributed policy control enables MNO-controlled traffic shaping and QoS enforcement in the transport network layer even in case of distributed/flat EPC architecture. Transport-layer technologies reduce deployment, operation and maintenance costs per transferred traffic volume. As Ethernet is the emerging technology in the backhaul access part, TRILL provides micro-mobility management for UE's moving between different eNodeBs/hotspots/base stations, i.e., updates the Ethernet VLAN endpoints and enforces traffic steering with UE network interface granularity (as indicated by ●). Among other benefits, HIP-FEMTO could manage IP mobility of moving femtocells/Wi-Fi hotspots with cell/MR level granularity of traffic steering (as indicated by ●).
- **Mobile service core optimization**: green technologies influence UE interface selection, GW selection, the path between UE/MR and GW by route optimization. These technologies assign traffic flows to a given UE network interface on IP-level. Part of them provide IP-level handover execution for the UE/MR (as indicated by ● and ●). The traffic steering to new interface can be user or flow level at these technologies. Gateway selection algorithm determines the other end of the path; it works on user-level and is restricted to S-GW, P-GW selection. The optimal placement of multiple S-GW/P-GWs is influenced by many factors such as demand traffic matrix, breakout possibilities given by transport network layer and tunneling options, content location (caches, CDN). The CAPEX/OPEX analysis described in previous section indicates which topology is the best to use.
- **Application-level optimization**: blue technologies are providing traffic/mobility management functionalities on application-level. The traffic treatment granularity of the technologies in this group is flow-level, they are restricted to specific transport-layer protocol or application type.
 - Some of the technologies (NMIP, SCTP) provide IP-level terminal-based handover execution if not handled by the mobile service layer (as indicated by ●), e.g., DSMIP is not deployed or activated on the UE, PMIP is not available in an access network. Hence these enlarge the space of seamless service continuity in heterogeneous access networks, and reduce the load of the IP mobility management function in the mobile service layer.
 - Technologies, such as ALTO and P4P influence the endpoint selection and load scheduling for specific applications (as indicated by ●), hence they provide better spatial and temporal distribution of these traffic demands, considering network operator aspects, such as network utilization.

- It is possible to deploy MPTCP-PR (TCP flow splitting/combining functionality) on multi-homed end systems (the GWs), such as multi-homed Femto GWs, Mobile Routers, BNGs, enabling capacity aggregation for specific TCP flows going through the end-system (as indicated by .
- Traffic identification and classification technologies are used by traffic and network management. QoE estimation and traffic management technology identifies QoS degradations for video streams and updates the stream bandwidth and other parameters accordingly. Mitigation of CDN load shifts is a technology influencing the resource utilization of the network segments between MNO's GWs and other IXPs, accordingly (as indicated by .

Considering the set of technologies in Figure 18 we can see that similar technologies can provide IP mobility management and/or certain technologies are bound to a given IP tunnelling option. These technologies hence mutually exclude each other on the level of specific UE, group of UEs. SON is a special case that influences both RAN and transport layers. Its functions automate network configuration and optimization; need to interact with O&M; and, sometimes have impact on both layers, e.g., triggering load balancing to optimize radio and transport.

Following, the functionalities and the performance gains achieved by the MEVICO technologies are explained based on the KPI assessment results.

RAN layer optimization

ANDSF

Performance gains in terms of KPIs:

KPI 1.1, KPI 2.1: The impact of the use of ANDSF is indirect: by improving the access to alternative radio access it allows to maximize the use of Wi-Fi AP. If we consider the mobility ratio 60% of user on the move vs 40% non mobile user (at home or at the office) , it can be assumed than at least 40% will be the base ratio of the traffic that will be offloaded.

KPI 2.3, KPI 3.1: ANDSF is using rules to inform the UE on which radio access should be preferred, hence link statistics such as service interruption delay due to handover or end to end delay may be taken into account to optimize these properties.

MCCS

Multi-criteria cell selection considers hierarchical 3GPP RAN cell structure. The cell selection for a given UE has influence on the 3GPP RAN throughput and radio transmission delay. The objective of the validation was the comparison of the performances of distance based cell selection, SINR based cell selection and global cell selection algorithm. Furthermore PF schedulers have been compared to RR schedulers. The validation scenario included 1 macro-, 0...2 pico-, 0...4 femtocells and 20...100 UEs. The analysis of MCCS methods was restricted to 3GPP-access networks but the topic can be extended to multi-access environments. The granularity of the traffic steering by this technology is device level (3GPP RAN network interface level).

Performance gains in terms of KPIs:

KPI 1.1: SINR-based provides higher throughput then distance based cell selection on the average (16-75% gain in terms of average total throughput depending on the number of UEs). On the average, SINR-based UEs have 11dB higher SNR then in case of distance-based cell selection. The PF scheduling algorithm gives ~10% better result than the RR scheduling algorithm in terms of average throughput in case of both cell selection algorithms.

KPI 1.2: maximum transmission delay is reduced by ~15% when distance based approach is used instead of SINR based approach

Mobile Relaying

The objective of this technology is to increase the RAN throughput by connecting macro-celledge users through mobile relays and/or Wi-Fi access network. The validation included the following main scenarios: macrocell without relaying as the reference scenario, macrocell with relaying, macrocell and celledge users connecting to Wi-Fi, macrocell with mobile relaying and Wi-Fi. Mobile relaying has impact on the UE and on existing network elements, provides low cost LTE RAN infrastructure, due to less micro-cells and same coverage with fewer base stations, furthermore provides one aspect for the selection of UEs to be offloaded through Wi-Fi. The traffic steering granularity is user-level or UE network interface level. The results are achieved for LTE, but can be extended to any OFDMA-based RAN, and IEEE 802.11n.

Performance gains in terms of KPIs:

KPI 1.1: In cellular communication systems, the users located at the edge of the cell generally receive the lowest throughput. Mobile relaying helps these users to increase their throughput via exploiting cooperation between the users. As a result of the throughput gain at the cell edge, overall system throughput increases as well. Simulation results have validated the stated increase in throughput of the celledge users (2-20% depending on the scenario).

KPI 2.1: In order to evaluate the mobile relaying system we used a heterogeneous network where Wi-Fi exists at the cell edge. A user at the cell edge can offload via Wi-Fi or a user-mobile relay. In either case the direct traffic to eNB is offloaded to another entity in the system. The simulation results have shown that, using mobile relaying together with Wi-Fi, 10% offload gain is achieved.

Selection of HO candidates in Wi-Fi hotspots

This technology focuses on decision algorithms for Wi-Fi offload, i.e., user selection in order to increase the efficiency of Wi-Fi resource utilization and maximize the offload to Wi-Fi access. It assumes operator managed Wi-Fi hotspots, or public hotspots. This network based access selection algorithm can be implemented in Wi-Fi APs or to separate network entities. The decision algorithm assumes multi-access scenario: users not connecting to Wi-Fi are assumed to access through 3GPP RAN.

Wi-Fi capacity is filled up with randomly selected terminals (as users/operator are willing to connect the user to Wi-Fi), however, when the QoS degradation of certain flow-types connecting through Wi-Fi reach some threshold value, one STA is selected and disconnected from Wi-Fi. QoS monitoring is application-aware, meanwhile the enforcement is user-level. Three selection schemes of STAs have been compared to decide which STA should be disconnected from Wi-Fi in such situation: 1) Random selection, which lowers the load in the network by randomly selecting a terminal, has very low complexity, but is RAN-layer technology-agnostic; 2) RSSI-based selection, where high received signal strength indicates good wireless channel, and accordingly we select the terminal with lowest RSSI value; 3) a novel Cost Function Approach (CFA), which selects inefficient terminals based on explicit evaluation of WiFi transmission parameters. Within CFA two approaches are considered a) one based on inefficiency metric, b) the other based on weighted aggregation of inefficiency and current network load of a STA.

Performance gains in terms of KPIs:

KPI 2.1: CFA and RSSI perform the same, and provide 15% gain in terms of number of supported VoIP flows with sustained flows compared to random selection. This has been achieved in a 802.11n hotspot where 200 STAs are uniformly distributed and the maximum distance of the STAs from the AP is 100m each generating one VoIP flow. For traffic-mixes containing both VoIP and FTP DL flows, the RSSI provides 15% gain, CFA provides 20% (a) or 50% (b) gain in terms of maximum supported VoIP flows, compared to random selection. Meanwhile the supported number of FTP DL flows is higher for RSSI based method then for CFA methods. Compared to "random selection" the gains for RSSI and CFA are 38% and 19% (a), -38% (b) in a 802.11n hotspot with 100m radius.

Selective admission control

With increasing traffic demands and the diverse application types carried by the LTE/EPC network, a novel admission control mechanisms should guarantee the QoS of different application types. Proposed mechanism, selective admission control, preserves the rule known from PSTN networks that an already admitted user is a satisfied user (QoS requirements are fulfilled). New flows with an explicit session start (e.g., TCP SYN packets) are rejected in case the measurements indicate congestion. Test scenarios will evaluate from the perspective of individual TCP flows this mechanism. The mechanism has flow-level granularity, and restricted to TCP traffic.

Performance gains in terms of KPIs:

KPI 1.1: Selective admission control will help to improve throughput in WiFi / LTE access

KPI 1.2: avoiding, mitigating congestion situations in the RAN improves QoS sustainment for other flows, including radio transmission delay

Cross-layer interference detection

The objective of this research is to find the exact behavior of TCP and ARQ in case of radio interference situations and propose interference detection algorithms in the network (e.g. to be implemented in the DPI). The mechanism can be deployed in any network element that is able to get RAN parameters and inspect TCP flows.

Performance gains in terms of KPIs:

KPI 1.1: by reducing interference the goodput of MAC frames will be increased.

KPI 1.2: by reducing interference radio transmission delay will be reduced.

SON

The SON concept and algorithms have been designed to improve the management plane via automated network configuration and operation. This has direct impacts on the costs (OPEX and CAPEX) but also on the transport and end-to-end quality and performance of the network. SON is necessary to be able to cope with the increased demands in network capacity and coverage, and the increased network complexity (coexistence of multiple technologies: Mobile-Fixed, Multi-radio and Multi-layer). Currently, SON focuses mainly on radio driven use cases that improve the Radio Network Layer efficiency. However, the reconfiguration of the radio parameters can have negative impact on the Transport Network Layer performance and thus on the overall quality of service (QoS). This makes it necessary to consider the status of the transport network and its potential QoS and user experience impacts during SON the decision process. Only Mobility Load Balancing optimisation (MLB) has been used to show performance gains when introducing eNB transport status evaluation and inter-rat MLB. Other algorithms were analysed to determine if they have impact on end-to-end performance. These are Automatic Neighbour Relation Function (ANRF), Energy Savings (ES) and Mobility Load balancing (MLB). On the other hand, the following have little or no impact: Coverage and Capacity Optimization (CCO), Interference Reduction (IR), Automated Configuration of Physical Cell Identity (PCI), Mobility Robustness Optimisation (MRO), RACH Optimisation (RO) and Inter-Cell Interference Coordination (ICIC).

MLB

The validation activities in the SON management solutions focused on two topics.

The first topic considered multiple packet data technologies, such as WLAN, and how they can be used to improve the efficiency of load balancing. The main goal is to minimize the number of unsatisfied users in the network by distributing the load from heavily loaded cells to less burdened ones. This allowed obtaining significant performance boost in network resource utilization and in the average data rate per user.

The second topic considered evaluating the benefits of introducing transport aware operation in the radio SON solutions. The radio SON feature selected was intra-LTE mobility load balancing (MLB) that was extended to integrate both radio and transport network in its decisions. The goal of the validation was to show that the enhanced solution is able to precisely detect radio overload and accurately evaluate the transport status of eNBs and, thus, avoid QoS deterioration observed when using only transport agnostic load balancing algorithms.

Performance gains in terms of KPIs:

KPI 1.1: LB allows integrating multiple packet data technologies, such as WLAN. Distributed intra-frequency load balancing algorithm based on automatic adjustments of handover thresholds reduces call blocking rate and increases cell-edge throughput. Number of unsatisfied users in the network is reduced by half or more (up to 5 times reduction were measured). Resources of the network are used more extensively with the LB algorithm, which means that more users can be satisfied with the same available number of PRBs and the average PRB number per user decreases constantly with increasing total number of users.

KPI 1.3: Congestion control is another aspect of load balancing where dynamic adjustment of frequency is used to minimize the inter-cell interference. Proper allocation of resources allows reducing the backhaul cost by redirecting some traffic into non-3GPP and global internet.

KPI 1.4: LB allows transferring users between base stations for more balanced load distribution in order to maintain appropriate end-user experience and network performance. Efficient load distribution makes sure the resources are utilized in the best way possible. The number of satisfied network users is an indication to the effectiveness of the load distribution algorithm.

The LB algorithms, designed for air interface and radio network layer, are improved by integrating knowledge of the status of the transport network (Load Balancing and Traffic Steering Entity). This helps avoid shifting traffic from overloaded cells to cells with sufficient radio resources but having only limited transport capacity. Results showed that congestion was eliminated that normally would not be eliminated. Furthermore, by considering the transport status, transport congestion can trigger LB and prevent QoE deterioration by unloading transport congested cells.

KPI 2.1: Multi-access can be used to reduce core network traffic, congestions and minimize the overall cost of the network. Mobile users can be served with the small cells in crowded places and WLAN inside airports, shopping streets and public places where high dynamicity is signified. The LB algorithm measures the node utilizations and decides how the users can be allocated so as not to overload a single node.

KPI 2.2: Integration of several technologies, with load balancing among them, improves the overall capacity of the network. The LB algorithm allows load balancing between different technologies, WLAN and cellular. The results present a significant real time improvement in overall capacity in terms of PRBs.

KPI 2.4: In intra E-UTRAN handover and E-UTRAN to WLAN handover scenarios, more efficient signaling is achieved when based on optimized LB algorithms.

Transport layer optimization

WMN

Performance gains in terms of KPIs:

KPI 1.1: The maximum possible throughput of the validation system (1 Gbps) was verified through one WMN node/eNB with 1000 byte packet length. With smaller packets the performance of the demo software implementation will deteriorate slightly. In traffic congestion situations lower QoS classes are dropped in favor of guaranteeing maximum throughput for high priority traffic classes.

KPI 1.2: WMN Only delay measurements with relative values were able to measure due to the restrictions of the validation system. Measured value 1,5 – 2,5 ms translates in theory to around 500 us with 200 us slot. Note this delay is for access network only. Aggregation and core transport delays needs to be added to traffic which is not handled locally within the WMN access network.

KPI 1.3: WMN Only recovery time with Ethernet cable connections was able to be measured. Ethernet port failure detection time was noted to be the dominating factor. On average 0,45 s failure detection time was measured, the protection switching or rerouting time being negligible in comparison (few us). With radio links, the targeted hitless protection switching and μ s level of the total recovery time inside the WMN system is expected to be reached.

KPI 2.2: In WMN system the throughput gain under varying network load situations is achieved through balancing and optimizing the available transport resources (transport link and routes) between the client systems by using the developed traffic management (congestion control, traffic balancing and route self-optimization) schemes. All mechanisms are QoS aware meaning that higher priority traffic flows are always favored over lower priority traffic flows. The implications are that basically any real time or delay constrained traffic can be forwarded with quite deterministic delay bounds and sustainable QoS through multi-hop WMN networks.

CES

Performance gains in terms of KPIs:

KPI 1.1: CES filters out unwanted traffic before it enters the backhaul, therefore the throughput increases. The gain depends on the amount of unwanted traffic and the used policies. CES also increases the throughput by removing the need for relaying packets, which typically is performed if the operator uses NAT. CES provides the same benefits as a NAT but completely removes the relaying and signalling overhead occurring when a NAT is present.

TRILL

Performance gains in terms of KPIs:

KPI 1.1: Fast and seamless mobility at Ethernet layer between eNodeB will reduce the amount of packets that need to be buffered or lost during the handover process.

The primary focus of our design is to reduce the S1 related signaling in the transport architecture. As such, our DHT extension to RBriges brings minimal benefits to throughput in the access network. However, RBriges have several significant throughput benefits over STP-based switches that are supported by our extension without any modifications.

KPI 1.2: TRILL The reduction in the handover delay and not having the need to utilize require S1 signaling for interdomain handover will reduce the E-E delay since packets do not need to be buffered during the mobility process.

The end-to-end delay measurements for the KPIs depend on the delays we choose for the simulated "Core - Internet", "Aggregation - Core", and "STPRoot – Aggregation" links. The baseline without any delays on the links is expected to be in the single digit milliseconds or lower.

In the centralized transport architecture, the end-to-end connection will suffer from all three modeled delays on the path, as public Internet access is through the core network. In the flat transport architecture, public Internet access is moved to the access network, so the end-to-end connection will suffer from only Internet related delays.

The testing was performed by taking the average of 1000 ping RTT results from the network, using the UE as the source of the ping, and the End Point as the destination. For the centralized transport architecture, the packet passes through the access, aggregation, and the core networks before travelling through the Internet to the end point. This results in 135ms of one-way delay, for a round-trip time of approximately 270ms. For the flat transport architecture, packets sent and received by the UE pass through the Access network directly to the public Internet, and to the End Point. This results in 30ms of one-way delay, for a total round-trip time of approximately 60ms

KPI 1.3: TRILL is a newly standardized link layer protocol, and as such does not provide rapid link failure recovery as part of the standard. TRILL uses a modified IS-IS as the link state protocol, which is used to respond to link and node failures in the network.

The IS-IS specification reacts to link failures upon not receiving a message from the neighbor in three consecutive heartbeat intervals. At minimum, the specification allows the heartbeat transmit interval to be set to one second, thus the minimum response time to non-local link failures is approximately 3 seconds.

Our DHT extension to TRILL responds to link and node failures in conjunction with the link state protocol operation. The information content held in TRILL nodes by our DHT extension is automatically repaired during link state protocol convergence when TRILL nodes fail.

Bootstrap time of TRILL nodes and our DHT extension is likely to be of minor consequence, as forwarding nodes in the access network rarely break. The typical bootstrap phase (ready to receive and transmit user traffic) of an RBridge when powered on is roughly 30 seconds. TRILL OAM and improved link failure detection features in general.

KPI 2.3: By implementing TRILL and our DHT extension in either transport architecture, we bypass the S1 signaling completely, and keep the mobility signaling in the access network. Expected value for the handover delay in this case is a sum of the normative handover delay and a single digit milliseconds delay caused by our extension updating the DHT location information for the UE, and propagating it to the necessary entities in the access network.

KPI 2.4: TRILL solution creates signaling load in the network whenever a UE moves between two eNBs. Attachment to the new eNB generates a signaling message (e.g., gratuitous ARP message), informing the TRILL/DHT node where the new eNB is connected, that an end-host has arrived behind the node.

The signaling message is intercepted, and location (and layer 3 addressing) information is updated in the access network via one (or two) DHT signaling primitive(s) messages used by our DHT extension. The destination of the update information is the TRILL/DHT server responsible for storing the information. If the server notices a change in the information, (e.g., the location has changed), the updated information is propagated in another DHT signaling primitive message to the TRILL/DHT node that the old eNB is connected to.

After receiving the updated information, the old eNB will forward all frames received for the UE to the new eNB, and in turn update the information on the TRILL/DHT node that originated the frame with the incorrect location information.

In summary, a handover by a UE causes at minimum two unicast signaling messages in the access network, typically at least three. Additional signaling is also required to reactively propagate the information to network nodes that are communicating with the UE.

Handover delay measurement test case:

The handover delay measurements were performed by moving the UE in the network between the two eNBs (a mobility event) in while a stream of ICMP echo packets (ping) were sent from the End Point in the Internet to the UE and back. The interval of the mobility events was set to 5 seconds, and 100 mobility events were recorded for each test. The performed handover was seamless, i.e., no lost packets were observed during the mobility event, however packet reordering was observed.

The centralized transport architecture modeled the simulated handover delay as 350ms, with a standard deviation of 50ms. In addition, the baseline test case added a single RTT of delay (70ms with a standard deviation of 7ms) to the handover delay to model the signaling delay related to the operation of the S1 protocol between the eNBs and the Core network. For our TRILL/DHT extensions, we added an additional delay of 50ms with a standard deviation of 5ms to model link latencies between the two separate Access networks where the eNBs are located and the Aggregation network.

The reported average handover delay of the mobility events was calculated by collecting the round-trip times of the first buffered ICMP echo packet, with the average end-to-end delay removed from each round-trip time. During the testing, the Internet End Point was sending ICMP echo packets with an interval of 275ms to the UE in the Access network.

The flat transport architecture tests model the X2 interface to signal handovers directly between the eNBs, bypassing the Aggregation and Core network links completely. Thus, in both the baseline, and our TRILL/DHT extensions case, the handover was modeled by a 85ms delay, with a standard deviation of 5ms. No additional delays were added to the handover. The reported average handover delay of the mobility events was calculated as with the centralized transport architecture, however during the testing, ICMP echo packets were sent with an interval of 80ms from the End Point in the Internet to the UE in the Access Network.

HIP-FEMTO

This technology improves the initial attachment phase of femtocell GWs to security GWs, plus, provides IP multihoming and mobility management for moving femtocell GWs. The security GW remains a centralized traffic anchor.

Performance gains in terms of KPIs:

KPI 1.1: It is demonstrated that significant throughput gains can be achieved through femtocell deployment, regardless of whether closed-access or open-access is considered. IP multihoming is a promising solution to achieve throughput increment. Speed of mobile stations has an impact on the throughput as well. We also measure the file transfer throughput of mobile stations with respect to their speed. We noticed a significant difference in throughput (134.6% improvement) between the singlehomed and multihomed approaches. Moreover, with multihoming a stable throughput with a less packet drop (178.6% improvement) was observed. The measurements are obtained over the mobile stations moving in an average speed of 72kmph.

KPI 2.3: In case of support of multihoming is enabled the service interruption time reduces to the routing update of flow. In case of break-before-make handovers HIP has low handover signaling latency compared to MIPv6 (200% reduction) due to the fact that the three way handshake defined in HIP enables faster reestablishment of layer 3 associations.

KPI 2.4: In this topic, an approach to reduce the handover signaling load was proposed and evaluated by means of the simulations. The other hand, it results faster recovery of the associations upon handover. HIP update procedure defines three messages and the first one is initiated from the UE. The average time to generate, send and process the first update packet is around 20ms. The first packet itself delivers the new location of the mobile station. The first, second and the third messages altogether consumes about 628 bytes which is comparatively less compared to mobile IP or IKE. Also, signaling delegation extension to HIP can immensely reduce the signal load on the network taking over the signaling rights behalf of mobile stations.

KPI 3.1: End to end delay has a significant impact on the real time mobile services such as VoIP, VoD and video calls. In fact, it is an aggregated delay that includes transmission delay, propagation delay and processing delay. Any small cell technology can contribute to reduce this aggregated delay. They can minimize the unnecessary collisions in transmission since there are only few subscribers connected to a femtocells or similar small cell at a time. Also, low probability of congestions is expected since they do not utilize the same core network infrastructure. The other hand, processing delay may be reduced at least at the access networks since a femtocell serves only few subscribers those who are registered for the services. In that sense, end to end delay between the UE and the content can be reduced. This topic measures the round trip time of different protocols with different architectural options defined in this particular scope and identify the proposed HIP solution has a significant improvement (almost 110%) over MIPv6.

Mobile service core optimization

Gateway Selection

GW selection influences which GW must be selected for a given tunneling option/IP connectivity, and is not restricted to any tunnelling option. Its usage is appropriate in any network architecture with multiple GWs. In 3GPP standard the following parameters influence the selection: used service class (APN), area served by S-GW, distance between S-GW and eNB (ordered list in DNS response), distance between PGW and service domain (ordered P-GW list in the DNS response), load of the GW, topological proximity of S-GW and P-GW (knowing from the structure of DNS name of the GWs). The added feature is to consider the following parameters as well: load of the transport network, access network supported by the GW and the UE. Regarding scenarios, GW selection is restricted to S-GW, P-GW selection executed by the MME. In case of IFOM, GW selection is activated before IFOM operations in order to select the GW for the lifetime of the IFOM connection. In case of MAPCON, GW selection could be used for selection of second GW as well, furthermore to decide where to assign a given flow.

Performance gains in terms of KPIs:

KPI 1.4: GW selection algorithm is hence concerned with GW and transport utilization and distance reduction from UE via S-GW, P-GW. However if a transport link is congested it avoids selecting such GWs that utilize that link.

KPI 2.2: delay and jitter reduction for different packets of the UE going through different interfaces towards the GW: such GW is selected which provides the shortest paths towards the known interfaces of the UE.

KPI 3.1: such GW is selected that minimizes the average delay from the UE to the content towards the different interfaces.

Operator managed Wi-Fi

Operator managed WiFi provides easier user access authorization management, new opportunities to offload traffic, increase in a cheap way indoor coverage, balance load between 3GPP and non-3GPP RAN and backhaul network, provide local breakout to local services, and meanwhile keep the provision of MNO services over Wi-Fi to the user, increase E2E user throughput, etc., enables new MNO business cases. At the end of the project this technique is already a product offering from several vendors and future work will be to further develop functionality so it will be a full-fledged Access Point in a HetNet solution.

Performance gains in terms of KPIs:

KPI 1.4: As for the prototype it is implemented in a way that makes it possible for an operator to use it efficiently for load distribution and load balancing. It opens opportunities for better scalability, i.e., it could help in adding and removing gateways when needed, shorter paths between the endpoints to meet the real time constraints, better fault tolerance, improvements in user experience as the data is transmitted through several paths or through the best, e.g. least loaded path in terms of QoS, network resource optimization by the use of multiple interfaces.

KPI 2.1: By using operator managed Wi-Fi, users can get operator partner services tied to mobile subscription also over WLAN behind RGW, e.g. Spotify, better indoor coverage and Busy Hour throughput in a home network could be considered much better than that of a macro cell that is shared between mobile devices in the cell.

From the operators' point of view, they manage the Wi-Fi access point and they can provide personal connectivity services for devices in residential network. They can provide better indoor coverage and also offload broadband traffic from wide area radio network to Wi-Fi.

The traffic evolution estimates done in MEVICO indicate that the traffic at home is 40%, the traffic On-the-move is 60%, but 25% of that is at workplace. With the operator managed Wi-Fi the operators have the possibility to offload the mobile based traffic onto a fixed network, either by enabling an easy subscriber handover to/from Wi-Fi or a default handover where the terminal automatically selects and switches to Wi-Fi.

KPI 3.2: this technology also provides offload gains of core network elements. Firstly, it offloads the mobile access network by routing the IP traffic via the fixed network to the S/P-GW. In that case the traffic load in the mobile access network and backhaul is significantly reduced while the core network traffic load is not affected at all. Secondly, IP traffic may be routed directly to the fixed broadband network and in that case the mobile core network traffic load is significantly reduced.

DMA with GTP

DMA with GTP improves QoS, IP mobility, session continuity in EPC containing multiple PGW. The technology has gains in scenarios with fast moving UEs.

Performance gains in terms of KPIs:

KPI 1.4: DMA for GTP allows for more flexible GW Selection in core. In particular it allows handling existing data connections. This way load distribution could be achieved more quickly as with standard mechanisms that work only when assigning new UE/PDN connections to the network.

The OpenFlow controller has the network wide status view of traffic load and is able to load balancing accordingly. More optimum selection of the offloading GW can be made based on the application needs.

KPI 2.2: Selecting combined S-GW/P-GW allows to nearly half the core network delay for packet processing what contributes to better QoS/QoE. Providing P-GW changes in an application dependent manner allows minimizing the impact on the user and QoE.

KPI 2.4: Avoiding significant changes and implementation effort to existing procedures to ensure backwards compatibility towards network deployments already in the field for selection of optimal P-GW location (preferably collocated with the S-GW) in order to have more optimal routing as the tunnels are terminated more close to the base station can be introduced by GTP enhancements with addition of new cause values (S11).

KPI 3.1: Provides the possibility to do the GW reselection for the existing flows to optimize the routing and to increase the usability of localized GWs. This way the mobile network topology can be adapted to the CDN topology. A main use case is to break out traffic to local content sources for minimal delay.

NB-IFOM

The objective is to implement different flow-level load-balancing strategies enforced by Network-based DSMIPv6. The decision and enforcement is flow-level, and the result is that flows are allocated to given UE interface, RAN and backhaul towards the HA. Consequently NB-IFOM has gains in multi-access scenarios. NB-IFOM is restricted to UEs that register to the given HA (during GW selection). Currently, without global HA-HA mobility extension, it supports only centralized mobility management.

Different load balancing strategies are implemented which specify where a flow is mapped, until there is no problem: 1) round-robin enforcing uniform distribution of SDFs among available uplinks, 2) least used enforcing the mapping of traffic to least used RAN and backhaul segment, depending current BW on UL, 3) lowest latency depending on current latency on UL, 4) overflow which waits while a link becomes full, does not utilize all accesses.

As the QoS of a flow measured by DPI or a network parameter provided by network management becomes suboptimal, policy change events are triggered, to reallocate flows to other accesses. The implementation of these policies is still future work.

Performance gains in terms of KPIs:

KPI 1.4: this solution explicitly enforces load-balancing with flow-level granularity.

KPI 2.1: feedback on user's flows by DPI, and Policy rules for service types together could control the PCEF (NB-DSMIPv6).

KPI 2.3: MCoA enables parallel establishment of tunnels from UE's given network interface to the HA, whenever an IP connection is ready. In case of overlapping RANs, to hand off a flow, the routing policy update must only be triggered in UL and DL direction for the given flows. Packet losses may happen when routing updates are not well synchronized, or the overlapping of different RANs is not enough high resulting in waiting times for IP connection establishment. Comparing DSMIPv6 with DSMIPv6+MCoA in multi-access scenario with overlapping WiFi and 3G RAN, 0% of flows fulfills less than 300ms HO delay. However with MCoA support, 30% of flows has <50ms, 50% has <100ms, 72% has < 150ms, 100% has <300ms service interruption delay.

KPI 2.4: compared to DSMIPv6, NB-DSMIPv6 reduces the signaling overhead by 40% if no policy change is required.

KPI 3.1: E2E delay between UE and CN is not the most optimal in NB-IFOM, IPinIP/IPsec tunnels are anchored to the HA that is allocated at the P-GW. However, the lowest latency-based flow distribution policy results in assigning flows to the interface with lowest transmission delay towards the HA.

PMIP-RO

The objective of PMIP-RO is to provide better E2E QoS and better traffic distribution due to route optimization in the network.

KPI 1.1: depending on the location of MAGs and transport network connectivity among MAGs and IAs, the PMIP-RO could enhance transport network utilization and increase the available capacity.

KPI 3.1: the path between UEs using PMIP is changed from [UE,MAG,LMA,MAG,UE] to [UE,MAG,IA,MAG,UE]. The transmission delay of the path depends on the placement of the IA, MAG, and IA selection policy.

KPI 3.2: One of the objectives of MPI-RO is to offload the LMAs. On the other hand, IAs should inherit some functions from P-GWs (charging, legal interception, etc). Assuming certain transmission delays between the UE, LMA, MAGs, IAs, the E-E delay reduction can be between 66% and -16.6% (i.e., there are situations where a centralized path could induce less transmission delay.)

PMIP-NEMO

The objective of PMIP-NEMO is to reduce the signalling overhead in the RAN, backhaul and core network and the load of the PMIP-based mobility management system in case of network mobility scenarios, e.g., for vehicular networks.

Performance gains in terms of KPIs:

KPI 2.4: PMIP-NEMO reduces handover signaling by 1 divided by the number of UEs attaching to the MR.

UFA-SIP

The objective of UFA-SIP is to improve the scalability of mobile service layer, in terms of mobility management, traffic management, network resource awareness and network-control. It also aims to improve the E2E QoE of applications by re-designing SIP-signaling and QoS enforcement procedures. The technology is restricted to SIP-based applications and SCTP-based applications.

Performance gains in terms of KPIs:

KPI 1.4: Since UFA-SIP realizes the flat architecture scenario, it can provide the gains in terms of total network load and CAPEX investments compared to the centralized EPC, as given by the CAPEX analysis. The most gain is achieved when two UEs are communicating through the same UFA GW or neighboring UFA GWs because the data traffic avoids the core and aggregation transport network. It depends on the handover decision mechanism implemented in source UFA GWs, that how the different service data flows are handed off among UFA GWs. An UFA GW is generic in terms of RAN type, hence the decision policies can also support traffic offload through Wi-Fi, etc.

KPI 2.3: For a SIP-based VoIP call the application handover delay is 80ms based on the number of lost packets. Regarding non-SIP based applications, the specific case of applications based on SCTP in the user plane has been considered. In terms of mobility, these applications can be managed either by mobile SCTP protocol or by UFA. Therefore, to show UFA benefits, a comparison with m-SCTP has been performed. UFA enables a gain ranging from 0.4% to 7.8% compared to m-SCTP in terms of downloaded data volume in a scenario where 1 to 13 handovers are made during 500 seconds.

KPI 2.4: the gain in terms of signaling messages of UFA-SIP compared to 802.21 network-controlled HO using PMIP-based handover execution procedure is around 43.75% and 26.3% for SIP and legacy internet applications, respectively.

KPI 3.1: by assuming the following transmission delay components the E-E delay reduction compared to centralized network falls in the interval [66%,-16.6%]. The transmission delay components are as follows: 1) UE to PGW: 30ms, 2) UE to UFA-GW: 10ms, inter UFA-GW: [0,40] ms.

UFA-HIP

UFA-HIP offers a new IP tunnelling, mobility management option for distributed or flat EPC architecture. It is convenient for scenarios where applications require uniform security, mobility management from the lower layer. Requires changes in the UEs and existing network elements, hence it should be introduced for new services e.g., when the MNO provides secure mobile internet service platform for M2M applications. UFA-HIP has the following restriction in terms of usage scenarios: non HIP-enabled UEs/CNs should not be connected to HIP-enabled UEs/CNs due to security reasons, or if necessary then the raised security threats should be analyzed.

This technology provides: (1) uniform security over any access network, (2) seamless service continuity in case of intra and inter-GW handovers over heterogeneous access networks, (3) support of legacy application that do not implement mobility nor security, (4) support of coexistence of IPv4 and IPv6 network segments, transparent for UEs and applications becomes possible due to UFA-HIP GWs, (5) better traffic distribution in the network, better E2E QoS.

The traffic steering granularity of this technology can be flow-, network interface- or user-level. The validation scenario demonstrates the mobility management functionality of the technology using simulations, and focuses on the measurement of HO performance gains.

Performance gains in terms of KPIs:

KPI 1.4: it is possible to involve different load distribution mechanisms based on current network status when deciding the candidate UFA-GW for a UFA-HIP aware UE.

KPI 2.1: scenario: overlapping Wi-Fi networks with different capacities. Offload policy: ongoing FTP applications were set up to always use the access network with the highest bandwidth among the available connections, while ongoing VoIP applications were specified to run on the access network with the lowest number of possible handovers. UFA-HIP scheme could result in more than 19% average throughput gain on the user side over the standard HIP multihoming solution (i.e., over the legacy case where only one interface is chosen for every application based on static priority). It also means significant load reduction on the 3G network elements as bandwidth hungry applications (FTP in our measurements) will be immediately offloaded from 3G access when an alternative radio network with higher potential bandwidth appears.

KPI 2.3: The service interruption delay was defined as the time elapsed between losing the connection at the old Access Point and the UE gets connected to the new Access Point. Measurements show that the service interruption delay of UFA-HIP is decreased by 72% with respect to the MIPv6 case and by 71% compared to the HIP case in average.

KPI 2.4: The handover related signaling load on the network was measured as the number of required messages for one successful run of the whole handover procedure consisting of initialization, preparation and execution. We used PMIP (RFC 5213) as a basis for our comparisons. UFA-HIP requires an average 55% increase in the number of signaling messages compared to the PMIP case.

KPI 3.1: UFA-HIP provides route optimization between UEs or UEs and CNs depending on the distribution of the UFA-HIP GWs. Depending on the placement of UFA-GW delay, assuming certain transmission delay values for the network segments, the delay reduction is between 66.6% and -16.6% compared to a centralized network.

KPI 3.2: UFA-HIP GWs take the role from P-GW, S-GW in IP connection provision. Data traffic of HIP-aware UEs hence is offloaded from centralized network elements.

HIP-Auth

This technology (HIP DEX-AKA) provides a new option for IP-level user authentication and key agreement procedure that could be used by MNOs. It could substitute IKEv2 in untrusted non-3GPP access for IPsec connection establishment. It also provides a possible option for fast initial authentication procedure in UFA-HIP. It supports seamless intra-GW IP mobility, and reduces the re-authentication time to new GWs in case of inter-GW handovers.

HIP-auth is appropriate for scenarios with highly-resource constrained UEs requiring uniform security services independently from the access network. The aim of the validations was to compare HIP DEX-AKA authentication procedure with IKEv2 EAP-AKA authentication and other methods, in case of different network topologies.

Performance gains in terms of KPIs:

KPI 2.3: HIP DEX-AKA provides 53%, 54% and 52% gains in the distributed, flat and centralized reference scenarios, respectively, compared to IKEv2 EAP-AKA in terms of authentication delay. IKEv2 EAP-AKA always hinders seamless handover for GBR services. HIP DEX-AKA over Wi-Fi access enables seamless inter-GW (ePDG) handover for certain GBR services, i.e., VoD, video streaming which have 300 ms Packet Delay Budget (PDB), but not for interactive videos, voice and gaming having less than 150, 100 or 50 ms PDBs.

KPI 2.4: The ratio of HIP DEX-AKA and IKEv2 EAP-AKA in terms of the number and total size of messages is 56% (9:16) and 37% (2054 bytes: 5512 bytes), respectively. Another important aspect is the number of control messages charging the aggregation and core network. HIP DEX-AKA requires on average two messages less per re-authentication than IKEv2 EAP-AKA, i.e., it provides 40% gain in these network segments.

KPI 3.2: In terms of computational requirements, HIP DEX-AKA has 88% gain on the UEs, and 98% gain on the GW. In terms of memory requirements, it has 80% gains on the UE and the GW.

Application-layer optimization

SCTP

The main functionality provided by the session layer implementation over SCTP is ‘session continuity’. The inherent support of SCTP for multi-homed endpoints (at either or both ends of an association) as well as its dynamic address reconfiguration extension makes SCTP quite attractive as an Internet mobility solution at the transport layer.

SCTP requires support on the UE or the application, and do not need modifications on the network side. It can provide end-to-end anchorless mobility. SCTP performance is independent of the architecture selected. It works equally well with all the architectures (centralized, distributed or flat). SCTP protocol can co-exist with other technologies as well. From the operators’ point of view, charging policies, gateway selection etc. is out of scope in this project. There is one issue with SCTP which is that, not all firewalls allow SCTP packets through. The firewalls have to be modified and new rules have to be added in order to use SCTP protocol.

Performance gains in terms of KPIs:

KPI 1.3: SCTP is designed to manage link failure and session resumption.

The resumption delay in an application-initiated session suspension will be equal to the round-trip-time (RTT) between the client and the server.

	Min(ms)	Max(ms)	Average(ms)	Std.Dev.(ms)
3GResumptiondelay	211,212	308,624	259,976	32,193
Wi-FiResumption Delay	10,498	27,527	17,317	4,280
3GRTT	205,749	335,571	227,014	23,340
Wi-FiRTT	4,394	85,663	16,635	15,985

Measurement result of session resumption delay in application-initiated suspended session

An optimization is introduced to reduce this resumption delay. Once the client’s network interface recovers, the session layer at the client will initiate a new transport level connection with the server. After the new transport level connection has been achieved, the session layer then pushes session resumption request to the transport layer. This session resumption request with a valid session ID is then sent to the server.

Interface recovery delay	Transport reconnection delay(ms)				Resumption procedure delay(ms)				Total delay(ms)			
	Min	Max	Avg.	S.Dev.	Min	Max	Avg.	S.Dev.	Min	Max	Avg.	S.Dev.
30s	36,19	135,62	92,62	39,18	6,35	23,69	11,71	5,13	43,85	159,31	104,33	42,13
60s	36,59	144,29	81,91	29,97	7,51	27,71	14,65	6,12	46,30	159,45	96,56	30,74

Measurement result of improved session resumption delay in application-initiated suspended session

The resumption delay consists of the time required forestablishing a new transport connection with the server and the time duration from sending the session resumption request until the session resumption acknowledgement is received from the server.

KPI 3.2: As for the other protocols (NMIP and MPTCP), SCTP does not need any anchor point, so it allows to minimize the number of hop between the end points.

NMIP

Performance gains in terms of KPIs:

KPI 2.1: The use of NMIP allows having fast vertical handover between interfaces and then to maximize the offload gain that have been estimated for ANDSF to 40%.

KPI 2.4: The service interruption delay during the handover is comparable to a RTT duration and its value is around 150ms.

MPTCP

Performance gains in terms of KPIs:

KPI 1.3: When the use of multi-access is available, the response time to link failure response time is negligible as the other access(es) will be used to manage all the communication. When the communication will be reestablished on the failed link, the transmission will be done transparently (no visible bootstrap time).

KPI 1.4: MPTCP with its multi-access property allow to distribute the load among the available interfaces. The packet distribution is done fairly an algorithm is used to avoid a MPTCP connection to get all the resources of an interface, it ensure that it will not get more resources than a usual TCP connection.

KPI 2.1, KPI 2.2: In our experiments, the capacity aggregation works well when the throughput of the different interfaces are close enough from each other, that is the case with Wi-Fi and LTE one example: MPTCP only Wi-Fi: 6.5Mbits/s MPTCP only LTE: 8.5 Mbits/s, MPTCP Wi-Fi+LTE: 10Mbits/s, this is a clear gain but it is not equivalent to the theoretical maximum (14Mbits/s)

KPI 2.3, KPI 2.4: The use of multi access allows to use the alternate link to continue the traffic so handover delay and service interruption delay due to handover are not anymore an issue.

KPI 2.5: When a new interface is available a TCP like connection is created (3 packets SYN SYN/ACK, ACK) and for the disconnection there are 4 packets on both side (FIN, FIN/ACK) this is negligible for a video session where thousands packets may be exchanged.

Distributed caching

Distributed caching improves resource utilization between the cache and content server (GWs to ISPs), the E2E QoE for the users. The caches must be placed behind the GWs at Gi interface or the breakout points. Supported applications depend on the cache-type, performance gains also highly depend on the cache hit rate.

Efficiency of small caches has been confirmed via evaluation of traces for user requests to YouTube videos. As a result, 0.5% of the videos accounted for 32% of the requests and 10% of the requests were addressed to 0.03% of most popular videos, which corresponds to around 500 videos in the trace.

Performance gains in terms of KPIs:

KPI 1.1: distributed caching at a local level (e.g., BTS) will directly reduce backhaul and transport network load.

KPI 1.2: distributed caching will partly eliminate backhaul delay.

KPI 3.1: distributed caching will efficiently and directly reduce E2E delay between user and content.

mP4P

This technology brings QoE improvements for users connecting with P2P applications, furthermore reduces the network utilization mainly on the network segment from GWs to IXPs.

The aim of the validation was to compare ALTO-server aided P4P video delivery with normal P2P video delivery using simulations. A topology with 3 ISPs in a fixed network has been used containing in total 750 peers assigned with random bandwidth, 30 seeds with 16 MB content, one AppTracker, one ALTO server for each ISP.

Performance gains in terms of KPIs:

KPI 3.1: e2e delay is reduced significantly due to ALTO-server knowing the cost map

KPI 2.2: using the P4P system, Download/upload rate increases by two-fold.

KPI 1.4: the inter-ISP traffic reduces significantly from 40% to 5%

Seeing these results, one can argue that similar results can be reached in LTE/EPC environments, however this is future work.

ALTO

ALTO influences the endpoint selection and load scheduling for ALTO-aware applications. It improves the spatial and temporal distribution of the traffic demands of these applications. Since ALTO server is operated by the MNO, network operator aspects, such as network utilization, can be considered.

ALTO Cost map and Endpoint costs help in the selection of Endpoint for the content. ALTO Cost schedule extension will help application to schedule their connection at time favorable wrt network usage. ALTO P4P reduces inter-domain traffic.

ALTO may have some impact on the UE on the application-level (ALTO client is necessary). On the network side ALTO server is required. It influences the path from UE to MNO-operated content server.

Performance gains in terms of KPIs:

KPI 1.1: throughput gain in access and backhaul is ameliorated. ALTO minimizes the needless use of network links, controls the utilization of network resources in terms of endpoints and time-schedule, hence it provides better throughput for others.

KPI 1.2: better delay for others due to balanced resource utilization of ALTO applications.

KPI 1.4: ALTO provides guidance to select application endpoint with respect to metrics including hop count and path bandwidth. Hence it has influence on load distribution of the backhaul and core transport.

KPI 2.1: ALTO associated with MAPCON influences offload decisions, i.e., it chooses endpoint considering path and resources. Furthermore, the ALTO server configuration could consider the same parameters as GW selection.

KPI: 3.1 minimization of E-E delay between UE and content is a core objective of ALTO.

BTA

MNOs would like to understand the usage of the network in detail. This technology provides a cost-efficient solution to get full picture on network usage. The usage will be categorized into P2P, Conversational (VoIP, IM, Video chat), Video, Web browsing. In addition to this the detailed reports will be displayed for the following application types: 1) P2P protocol type, 2) conversational: application-type like Skype, talk, 3) Video: distribution into youtube, Facebook, dailymotion, etc., 4) Web browsing: distribution of usage into URLs. This will enable a network operator to understand usage to make additional settings on PCRF, provide more efficient campaigns.

Reporting can also be done by DPI, however in order to achieve that, detailed configurations and settings are necessary in the network. This would require an immense amount of investment and operational costs would be high.

Performance gains in terms of KPIs:

KPI 1.1: Using BTA, analysis of network usage will enable the operator to prioritize traffic better and develop campaigns to move some traffic such as P2P to non-busy hours, hence improving usage distribution and in effect capacity of the network.

DPI

Deep packet inspection enables to get detailed view on traffic flows and their QoS/QoE, which can be used as input for network management and traffic management. It enables policy enforcement from low to very high granularity.

Performance gains in terms of KPIs:

KPI 1.1: The DPI by itself will not increase the overall throughput. Rather it is a mechanism that can be used to increase the throughput of selected application classes by prioritizing their traffic (if the legislation allows such prioritization) or reducing the throughput of undesired non vital services.

KPI 2.2: E-E QoS sustainment: DPI is a core technology to classify traffic based on the application type and to measure the QoS metrics at the application level. It can also provide input for application QoS/QoE measurement mechanisms.

MPTCP-Pr

MPTCP-PR provides multipath support for TCP-based applications without implications on the UE. TCP flow splitting/combining functionality could be deployed on multi-homed end systems, Femto GWs, MRs in case of NEMO scenarios. By offloading some of the TCP traffic via breakouts, the MNO's access and backhaul network will be prevented from excessive data traffic. The granularity of this traffic management solution is under TCP flow-level.

Performance gains in terms of KPIs:

KPI1.1: load reduction on specific access and backhaul network parts.

KPI 2.1: Offload gain due to usage of multi-access capacities. MPTCP-Pr provides additional flexibility for offloading also for the single-homed scenarios (no multihoming feature is necessary for the UEs,).

KPI 2.2: Capacity aggregation due to multipath, for a specific TCP flow.

QoE estimation and traffic manipulation

QoE estimation has no impact on existing network elements. Similar to DPI, it doesn't increase the overall throughput or performance. It can be seen as a mechanism that can be used to evaluate other optimization techniques to improve their effectiveness with respect to the user's QoE. This technique is particularly useful for improving overall user experience when using real-time video streaming based on, for instance, RTSP and RTP protocols.

Performance gains in terms of KPIs:

KPI 1.1: The QoE estimation will not increase the overall throughput, but it will allow increasing the throughput of selected services. The QoE of selected service flows is estimated and, if there is a low QoE detected, the priority of this flow will be increased which leads to a higher throughput for the specific service flow. The gain in throughput is achieved by lowering the throughput of a non-vital service.

KPI 2.2: At a certain set of observation points within the network, the QoE of specific services (e.g., IPTV, YouTube) is estimated using QoS measures mapped to Quality Indicators using fuzzy logic techniques. The QoE estimation method used was shown to be highly correlated to both the participants' subjective QoE scores as well as to the estimated Mean opinion Score (MOS) obtained by other techniques, e.g., Video Quality Model (VQM). VQM measures the perceptual effects of video impairments, combining them into a single metric. The estimation accuracy emphasizes the ability of the proposed system to measure the impact of the network conditions on the user satisfaction.

In addition to the previous work, YouTube video QoE estimation has also been performed by TCP flow observation within the network. Therefore, no access to the user's end device or the YouTube server is needed. TCP/ACK timestamps are compared with the video timestamps which are contained in the payload of the corresponding TCP segment. Out of this comparison the fill level of the video play out buffer is calculated and video stall events are identified. To increase the processing speed two variants of the QoE estimation algorithm based on throughput measurements have been developed.

8.3 Validation Conclusions

The technologies considered can interwork and the validations results on the different layers show performance improvements in several parts of the network as described next.

In the RAN-layer, the MNO should prioritize among the decisions made by the different technologies in terms of Wi-Fi hotspot, macro-, micro-, femtocell selection. RAN-layer technologies are probably the most relevant in terms of bringing gains to the network resource utilization. While typically the influence of these technologies is user-level, UE-network interface level, ANDSF decisions could also influence flow mapping to access network based on fivetuples.

The technologies proposed in the RAN-layer provide following improvements. With ANDSF up to 40% of the traffic can be offloaded to WiFi connection. MCCS provides throughput improvements in the order of 16%-75% and maximum transmission delay is reduced by 15%. The Mobile relaying increase the throughput between 2-20% and used together with WiFi 10% of the traffic can be offloaded. Including RSSI-based or CFA logic to perform HO with WiFi hotspots can provide increase 15% in the number of supported VoIP flows. The selective admission control together with cross-layer interference detection, SON functionality and MLB provide guaranteed QoS for different applications.

The transport layer technologies provide improvements that do not interfere with the RAN-layer, mobile service network layer or application. WMN provides a throughput of up to 1Gbps with a maximum delay of 1.5-2.5ms and a failure detection of 0.45s which results in reliable mechanism to extend the network capacity in the edge. TRILL provides handover in access network in the order of few ms, thus reducing the overall end to end delay of ongoing connections. A single digit ms is required by TRILL+DHT to update the MAC address of the UE after moving to new eNodeB, thus reducing signalling and HO delay from hundred of ms to few ms. HIP-FEMTO provides throughput

increase in the order of 134.6-178.6% and handover latency reduced 200% compared to MIPv6. Other technologies such as CES, distributed PCRF and MACEth provide improvements in the throughput by reducing unwanted traffic and optimizing the security and QoS setup.

As soon as traffic steering decisions between multiple accesses are made by mobile service layer, the RAN-layer and mobile service layer policies should be harmonized. Flow-level traffic steering is probably needless for better load distribution and offload in the network because RAN-layer optimizations, user-level GW selection can achieve the objectives related to network utilization efficiency, energy consumption etc by these technologies.

On the other hand mobile service network layer technologies provide good options to enforce subscription, service, application-based traffic steering policies, or move specific flow types to another access/backhaul. MNOs should prioritize network and user aspects represented by the policies that influence RAN-layer optimizations and enforce user SLAs.

The mobile service layer technologies provide the following improvements. UFA-SIP provides handover reduction normally in the order of 80ms by 0.4-7.8% compared to m-SCTP. Moreover, the gain in signalling compared to network controlled HO using PMIP is around 43.75-26.3% and E-E delay reduction about 66-16%. UFA-HIP decreases service interruption during HO an average of 55% and 72% compared to PMIP. It also provides route optimization, thus reducing the E-E delay between 66.6-16.6%. HIP-Auth provides 52-54% gains in authentication delay compared to IKEv2. Moreover, gain of 40% in reduction of messages compared to IKEv2 in addition to 80% gain in memory requirements. Other technologies such PMIP-NEMO, PMIP-RO, NB-IFOM, DMA with GTP, Gateway Selection and network managed WiFi provide improvements in the area of load balancing and traffic offload that improves the resource allocation, QoS and traffic distribution.

Finally application layer technologies provide optimizations described as follows which then complement the above listed technologies integrated in several parts of the network and at different layers.

SCTP contrary to other protocols such as NMIP and MPTCP do not need anchor points so mobility is optimized. NMIP provides optimized service interruption of 150ms. MPTCP provides improved aggregated throughput with multiple interfaces so MTCP with Wi-Fi provides a throughput of 6.5Mbits/s, MPTCP with only LTE provides 8.5 Mbits/s throughput but MPTCP Wi-Fi+LTE provides a throughput of 10Mbits/s (close to the 14Mbs theoretical throughput). Distributed caching shows improvement in serving content as the results show that 0.5% of the video accounts for 32% of the requests. mp4P shows improvements in the reduction of ISP to ISP traffic by 40-5%. ALTO together with MAPCON minimize the E-E delay and optimize the offload decisions considering path and resources. MTCP-Pr provides load reduction on certain part of the network and improves the capacity aggregation due to the multipath support for TCP flows. BTA, DPI, QoE estimation and traffic manipulation improve the knowledge of the network and the performance bottlenecks, thus improving in the selection of other technologies to overcome traffic loads and increase the overall QoE.

Summarizing, the technologies analyzed in MEVICO provide performance improvements according to the listed KPIs. A remaining issue is to analyze deeper the interworking of the policies that influence the access network selection, handover, admission control, traffic steering. The proposed architecture includes several technologies with their corresponding performance improvements. However, network manufacturers and operators would decide the ultimate architecture based on the specific needs and required improvements on selected parts of the end to end mobile network.

9 Conclusion

The digital lifestyle goes mainstream and mobile broadband traffic volume increase is inevitable and the future network architecture evolution has to be optimized to cope with this situation. Moreover, the evolving technologies in connectivity, end devices User Interfaces, and end user applications such as Video on Demand, Multimedia Streaming, Home networking, Remote Monitoring, Tele-Health and M2M applications will change the traffic patterns as they are known today. Therefore, mobile networks have to support not only rapid traffic demand but also variation in the traffic profiles. In order to manage the increased traffic and new applications with new requirements, LTE-EPC technologies have adopted an all-IP architecture that integrates a more distributed management and QoS strategy. This architecture simplifies the network stack, but can make the efficient operability more complex. Operators seek to successfully deliver robust rich media data, voice and video services. There is a need to measure and assure QoS in the all-IP network. This requires not only proper planning and network engineering, but also a system that is robust, optimized and designed to handle future mobile data demand. In MEVICO project the target was to specify a network model optimized to maximize the end-user mobile broadband experience and ensure efficient congestion-free network performance.

This MEVICO architecture document firstly describes the problem statement that mobile networks will face in the future. The challenges and requirements for the next generation of mobile networks are identified. The challenges are mapped into different KPIs. Secondly, MEVICO proposes a set of selected technologies to address those KPIs. Thirdly, MEVICO architecture approaches describe how those technologies are planned to be deployed in the current Evolved Packet Core (EPC) network and how they could co-operate to provide efficient architecture evolution. Finally, the coexistence issues during the deployment of the different technologies are indicated.

Operator's choice for future business strategy and technology approach especially in terms of mixture of evolution opportunities for operator's current technology assets and choice of new technologies is a challenging task. For each operator the choice is different. The validation results in section 8 show there are indeed many possibilities to consider on the main layers of an architecture, namely RAN layer optimization, Transport layer optimization, Mobile service core optimization and Application level optimization, for instance in terms of performance improvement for a particular technology of a choice.

Technologies for RAN layer optimization can influence radio access network selection with user-level granularity, and include RAN-layer traffic management, network management. These technologies determine 3GPP cell / Wi-Fi hotspot selection for the UE's network interfaces, hence optimize RAN and backhaul network resource utilization.

Technologies for Transport-layer optimization can improve the transport network layer's capacity, increase the transport connectivity possibilities, provide capacity improvements, fault-tolerant transport. Distributed policy control enables MNO-controlled traffic shaping and QoS enforcement in the transport network layer even in case of distributed/flat EPC architecture. Transport-layer technologies reduce deployment, operation and maintenance costs per transferred traffic volume. As Ethernet is the emerging technology in the backhaul access part, TRILL provides micro-mobility management for UE's moving between different eNodeBs/hotspots/base stations. Among other benefits, HIP-FEMTO could manage IP mobility of moving femtocells/Wi-Fi hotspots with cell/MR level granularity of traffic steering.

Technologies for Mobile service core optimization can influence UE interface selection, GW selection, the path between UE/MR and GW by route optimization. These technologies assign traffic flows to a given UE network interface on IP-level. Part of them provide IP-level handover execution for the UE/MR. The traffic steering to new interface can be user or flow level at these technologies. Gateway selection algorithm determines the other end of the path; it works on user-level and is restricted to S-GW, P-GW selection. The optimal placement of multiple S-GW/P-GWs is influenced by many factors such as demand traffic matrix, breakout possibilities given by transport network layer and tunnelling options, content location (caches, CDN). The CAPEX/OPEX analysis described in section 7 indicates which topology is the best to use.

Technologies for Application-level optimization can provide traffic/mobility management functionalities on application-level. The traffic treatment granularity of the technologies in this group is flow-level, they are restricted to specific transport-layer protocol or application type.

All technologies this technologies described above as part of the MEVICO architecture support all the different topology options. However, following are some technologies that need to be considered by the operator based on their specific needs

- UFA-HIP, UFA-SIP, DMA with GTP, Gateway selection technologies are based on the assumptions that there are multiple GWs. The UFA technologies were originally designed and validated in flat architecture options. If the techno-economical results indicate the distributed or centralized architecture options to be the most cost efficient, they also might work in a distributed or centralized architecture without bringing the performance gains in terms of better load distribution in the transport and core network.
- NB-IFOM is a centralized mobility management solution keeping the user plane anchored to the home agent. It enforces load balancing strategies in backhaul segments and implements network-controlled access interface selection based on the fivetuple containing source and destination IP addresses, and transport protocol type.
- PMIP-RO and PMIP-NEMO protocols could support centralized and distributed architecture options, but the LMA is centralized, and the performance gains in terms of path reduction, load distribution (in case of PMIP-RO) depend on the area covered by LMA and the MAG distribution.
- Distributed caching is meaningful in distributed architecture option only.

The results of the CAPEX/OPEX analysis shown in section 7.4 indicate that the ratio of network equipment/transport capacity cost determines what would be the preferred architecture topology. The results indicate that the technology evolution that commoditizes either the network elements capacity or the transport network will affect the above mentioned ratio, thus being decisive in identifying the optimal topology.

For the MNO, with the legacy network implemented, the deployment cost/effort of the new technology might be requiring quite significant improvement in the performance and benefits, in order to have enough rationale to deploy the technology. Also the general industry acceptance and the possibly needed UE vendor commitment are crucial to make the technology attractive. Standardization in the industry forums (like 3GPP) is the way to ensure the commitments from the key players.

Overall, MEVICO work did not identify clear bottlenecks in the LTE architecture, but there are needs for high scalability and flexibility of the network capabilities due to potentially quickly evolving demand. Network needs to adapt and optimize itself to meet the changing needs of subscribers, the services they use and the operational state of the network itself. Network shall become continuously aware of user traffic demands and the network resources that are available to serve those demands dynamically. Thus, improved architectural optimizations (both with respect to CAPEX and OPEX) have to be identified in order to ensure the sustainability of future mobile networks.

There are certain challenges pointed out in the extensive GW distribution, like economical sustainability and created complexity compared to potential benefits. For the future network capability enhancements some new evolving networking trends might be reasonable to consider for further study.

Mobile networks are constructed with utilization of IT and networking technologies. Both technology areas are currently experiencing fast transformation, mainly due to virtualization. Clouds are changing the way software systems are deployed and managed, and network virtualization enables dynamic provisioning of virtual network slices to different users of the network. One promising solution to this dilemma is to apply software defined networking, or programmatic control of network resources, to decouple the innovation cycles in software and hardware systems.

The advent of mobile ubiquity is pushing the growth of traffic to unforeseeable levels. To obtain sustainable true mobile broadband, the increase of revenue needs to follow this increase of traffic since, if it does not, then the user experience will be negatively impacted. LTE brings the means to obtain bigger, smarter and cheaper pipes, cheaper network operation and high performance; but needs have arisen to introduce new service and revenue models that integrate service differentiation, cost and optimizations, virtualization, energy efficiency and self organizing networks.

On one hand, Cloud Computing (CC) is being adopted in enterprise IT systems to achieve virtualization of IT infrastructure to optimize resources through sharing, outsourcing operation and service architectures to obtain the necessary CAPEX/OPEX improvements. On the other hand, LTE flat network architectures variants and network convergence scenarios proposed in the MEVICO project need to be introduced to obtain the required improvements in mobile communications. Network virtualization, CC, Software Defined Networking, as well as other technologies, need to be studied and developed for achieving the needed e2e QoE and the CAPEX/OPEX improvements.

In all, security, flexibility, maintainability, interoperability, performance, scaling issues still need to be further addressed before widespread use of these technologies is possible. A common solution based on carrier grade CC will allow achieving complete convergence that better integrates the telecom and IT worlds to obtain the benefits each can offer.

The use of CC in telecom networks

The different NE and functions used to run only on dedicated servers. The elimination of hierarchical network layers (flat architectures) allows using generic servers hosting the network functions and locating them in the cloud so that they can be dynamically created, allocated, de-allocated, moved and removed from virtual machines. This will bring several benefits that include scaling telecom services on demand, improving reliability and availability and thus make the use of the telecommunication infrastructure more efficient.

The use of telecom networks in CC

Cloud computing usually requires large amounts of processing power, storage and bandwidth. Managing these resources becomes critical to support the promised quality and service mobility to the users. Furthermore, many future cloud devices will be wireless, requiring ubiquitous high speed, wireless internet support that can most effectively be provided by mobile networks.

CC will have a big impact on telecom networks due to increased demand for traffic but also, for instance, making it necessary to define new business models that deal with SLA responsibility and assure the migration of current data centres.

The introduction of new services and technologies also has high impact on networks due to the increased complexity that is required to manage them. This implies the need to introduce new techniques that allow automating the deployment, optimisation and operation of network virtual or non-virtual resources and services. Self-Optimising Networks, Software Defined Networking are expected to become key elements that allow shaping and managing bandwidth, fine grained interaction with applications to deliver the quality of service required by customers. These require the introduction of technologies that provide the capability to recognize different traffic flows, such as deep packet inspection, service steering, and intelligent traffic control to dynamically monitor and control sessions on a per-subscriber/per-flow basis.

Power consumption in the networks has not been considered in MEVICO but will be a main requirement in future network solutions, we need to add comments in this respect to the different architectures and technologies if only as a reference to conclusions and findings in other projects e.g. EARTH [9] and DC2F .