Project Number:	**CELTIC / CP7-011**
Project Title:	Mobile Networks Evolution for Individual Communications Experience – MEVICO
Document Type:	PU (Public)

Document Identifier:	D 4.2.1
Document Title:	**Traffic management building blocks in next generation mobile telecommunication systems**
Source Activity:	WP4
Main Editor:	Jochen Eisl
Authors:	Jochen Eisl, László Bokor, Zoltán Faigl, Zoltán Németh, Thomas Bauschert, Gerd Windisch, Gerhard Hasslinger, Florian Metzger, Albert Rafetseder, Sabine Randriamasy, Ece Saygun, Salih Ergüt, Alper Kirbiyik, Tapio Suihko, Łukasz Budzisz, Berthold Rathke, Edgardo Montes de Oca
Status / Version:	1.0
Date Last changes:	10.06.2011
File Name:	D4.2.1

Abstract:	This document describes the different categories of traffic management scenarios including problem statements. Subsequently the requirements are derived for a traffic engineering architecture within the evolved packet system.

Keywords:	building blocks, QoS, requirements, traffic management

Document History:	
22.03.2011	Initial version based on IR-4.2.1
09.05.2011	Version to be submitted for QEG review
08.06.2011	Final updates after review

Table of Contents

Authors

Partner	Name	Phone / Fax / e-mail
Nokia Siemens Networks		
	Jochen Eisl	Phone: + 49 89 5159 35125 e-mail: jochen.eisl@nsn.com
	Wolfgang Hahn	Phone : e-mail: wolfgang.hahn@nsn.com
Technical University of Chemnitz		
	Thomas Bauschert	Phone: e-mail: thomas.bauschert@etit.tu-chemnitz.de
	Gerd Windisch	Phone: e-mail: gerd.windisch@etit.tu-chemnitz.de
Deutsche Telekom		
	Gerhard Hasslinger	Phone: e-mail: Gerhard.Hasslinger@telekom.de
Technical University Berlin		
	Łukasz Budzisz	Phone: e-mail: lukasz.budzisz@tu-berlin.de
	Berthold Rathke	Phone: e-mail: berthold.rathke@tu-berlin.de
Budapest University of Technology and Economics – Mobile Innovation Center		
	László Bokor	Phone: e-mail: bokorl@hit.bme.hu
	Zoltán Faigl	Phone: e-mail: zfaigl@mik.bme.hu
	Zoltán Németh	Phone: e-mail: znemeth@mik.bme.hu

University of Vienna

Florian Metzger

Phone:

e-mail: florian.metzger@univie.ac.at

Albert Rafetseder

Phone:

e-mail: albert.rafetseder@univie.ac.at

Ericsson Turkey

Ece Saygun

Phone:

e-mail: ece.saygun@ericsson.com

Turk Telekom

Salih Ergüt

Phone:

e-mail: salih.ergut@turktelekom.com.tr

Avea

Alper Kirbiyik

Phone:

e-mail: Alper.Kirbiyik@avea.com.tr

VTT

Tapio Suihko

Phone:

e-mail: Tapio.Suihko@vtt.fi

Alcatel Lucent

Sabine Randriamasy

Phone:

e-mail: Sabine.Randriamasy@alcatel-lucent.com

Montimage

Edgardo Montes de Oca

Phone:

e-mail: edgardo.montesdeoca@montimage.com

Executive Summary

The MEVICO project aims at analyzing the actual 3GPP LTE-mobile broadband network and the identification of the technologies for its evolution. The target is to innovate and develop new network concepts for meeting the future requirements of the evolving mobile networks. The work related to this document encompasses smart traffic management for the next generation of mobile networks. As mobile and wireless communication networks move toward broadband converged networks and applications, the demands on the infrastructure will increase tremendously.

Within the context of the project this document focuses on traffic management mechanisms for mobile broadband networks based on state-of-the-art technology and new concepts, which have not yet been deployed within commercial mobile networks. First of all we have described a set of general objectives for traffic management in order to understand how various mechanisms are suited to reduce or avoid the traffic related problems. Subsequently we have introduced the concept of traffic management building blocks. This serves two purposes: On one side the various mechanisms can be categorized according to common functionality or solution space. Second these building blocks represent design options and therefore may be used later in the project to define the detailed solutions for traffic management in the evolved packet system. The document aims for a holistic approach, i.e. the building blocks and exemplified mechanisms provide an overview of the different approaches to meet traffic management objectives. It is obvious that the project consortium can't elaborate in detail all these different directions. Hence the conclusion section of this document provides information about the focus topics to be further investigated within the described building blocks. Since this project focuses on innovative approaches on packet core, mobile backhaul and operator service domain, the radio management aspects addressed in this document will not be taken into further considerations. The same applies to traffic management, which is restricted to the application layer or is mainly about business modelling aspects. The four building blocks of highest interest to the project consortium are 'microscopic traffic management', 'macroscopic traffic management', 'improved resource selection and caching' and 'deployment of new network resources'. Due to limited effort of the project partners some of the smaller aspects within the aforementioned building blocks are also not further elaborated. After the description of the building blocks, the document provides an overview about operational, functional and architectural requirements found relevant by the project consortium. The results from this document serve as input to the traffic engineering architecture.

List of acronyms and abbreviations

2G/3G/4G	2nd/3rd/4th Generation Cellular Mobile Phone System (GSM, UMTS, LTE,…)
3GPP	3rd Generation Partnership Project, based on GSM Technology
ALTO	Application Layer Traffic Optimization
AMBR	Aggregated Maximum Bit Rate
API	Application Programming Interface
APN	Access Point Name
ARP	Allocation and Retention Priority
ARQ	Automatic Repeat Request
AS	Autonomous System
AVP	Active Virtual Peer
BNG	Broadband Network Gateway
BS	Base station
CAPEX	Capital Expenditure
CDN	Content Distribution Network
CIF	Common Intermediate Format
CSCF	Call Session Control Function
CSP	Communication Service Provider
DL	Downlink
DNS	Domain Name System
DoA	Direction of Arrival
DPI	Deep Packet Inspection
DSL	Digital Subscriber Line
DSMIPv6	Dual Stack Mobile IPv6
E2E	End-to-End
eNB	eNodeB
EPC	Evolved Packet Core
EPS	Evolved Packet System
E-UTRAN	Evolved UTRAN
FTP	File Transfer Protocol
GAN	Generic Access Network
GBR	Guaranteed Bit Rate
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
GTP	Generic Tunneling Protocol
GW	Gateway
HA	Home Agent
HeNB	Home eNodeB
HLR	Home Location Register
HO	Handover
HSS	Home Subscriber Server
HTML	Hypertext Markup Language
HTTP	Hyper Text Transfer Protocol
IETF	Internet Engineering Task Force
IFOM	IP Flow Mobility
IMPEX	Implementation Expenditure
IMS	IP Multimedia Subsystem
IRSC	Improved Resource Selection and Caching
ISP	Internet Service Provider

LAN	Local Area Network
L-GW	Local Gateway
LI	Lawful Interception
LIPA	Local IP Access
LSP	Label Switched Path
LTE	Long Term Evolution
LTE-A	LTE-Advanced
MAC	Medium Access Control
MacTM	Macroscopic Traffic Management
MBR	Maximum Bit Rate
MicTM	Microscopic Traffic Management
MIH	Media Independent Handover
MME	Mobility Management Entity
MNO	Mobile Network Operator
MP2MP	Multipoint-to-Multipoint
MPLS	Multiprotocol Label Switching
MPTCP	Multipath TCP
MT	Mobile Terminal
MWR	Microwave Radio
NAPTR	Naming Authority Pointer [Resource Record]
NAT	Network Address Translation
NE	Network Element
NIC	Network Interface Card
OPEX	Operational Expenditure
P2MP	Point-to-Multipoint
P2P	Peer-to-Peer
P4P	Provider Portal for (P2P) Applications
PCC	Policy and Charging Control
PCRF	Policy and Charging Rules Function
PDN	Packet Data Network
PDP	Packet Data Protocol
P-GW	PDN Gateway
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RFC	Request For Comments
RNC	Radio Network Controller
RRM	Radio Resource Management
RTSP	Real Time Streaming Protocol
SCADA	Supervisory Control And Data Acquisition
SCTP	Stream Control Transmission Protocol
SDH	Synchronous Optical Hierarchy
SeGW	Security Gateway
SGSN	Serving GPRS Support Node
S-GW	Serving Gateway
SIP	Session Initiation Protocol
SIPTO	Selected IP Traffic Offload
SLA	Service Level Agreement
SMTP	Simple Mail Transfer Protocol
SRV	Service [Location Resource Record]
TCP	Transmission Control Protocol
TE	Traffic Engineering

TM	Traffic Management
UDP	User Datagram Protocol
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
UTRAN	UMTS Terrestrial Radio Access Network
VLAN	Virtual LAN
VoIP	Voice over IP
VPN	Virtual Private Network
WLAN	Wireless Local Area Network
WMN	Wireless Mesh Network

1. Introduction

The research project MEVICO investigates aspects of the 3GPP LTE-mobile broadband network for its evolution in the mid-term in 2011-2014 and beyond. The goal is to contribute to the technical drive and leadership of the Evolved Packet Core (EPC) network of the 3GPP, and thus support the European industry to maintain and extend its strong technical and market position in the mobile networks market. The project follows an end-to-end (E2E) system approach on evolution of the EPC. The focus will be on the connectivity layers of the system, for example on the part of the future LTE network which provides the efficient packet transport and mobility support for the applications and end-user services accessed over the LTE and LTE-Advanced radio systems. The technical research of the project covers relevant topics in the areas of network architecture, mobility and routing, packet transport, traffic management, network management and engineering, and techno-economic aspects. The project will include both conceptual research and demo/trial system implementations.

The work related to this document encompasses smart traffic management for the next generation of mobile networks. As mobile and wireless communication networks move toward broadband converged networks and applications, the demands on the infrastructure tremendously increase. Data, voice, video, and wide-scale of applications are running on the same network, simultaneously. Advances at the terminal side increase the need for smart traffic management solutions as smart phones, tablet PCs and other mobile devices are able to run multimedia applications. Consequently high amount of wireless bandwidth will be available (possibly equal to typical PCs on broadband connections). LTE network will follow the same path as wireline networks in the past and will thus become quickly dominated by CDN and P2P traffic. It will become 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 most of the time. Ensuring that mobile networks are application-aware, we can achieve flexible adaptation to any new application and traffic pattern as they emerge in the future. Operators need to install effective management tools to manage traffic using QoS policies, bandwidth allocation schemes, prioritised access and admission control, traffic shaping and rate control, and flow based processing. Only such advanced and active traffic management will ensure that operators can provide cost-effective data transfer with real-time speech and video on heterogeneous accesses.

The current document describes initial investigations on traffic management in the context of the project in order to set a base for subsequent concept and algorithm development tasks. The first aim is to derive general objectives for traffic management, which is explained in section 2. It has been identified that various approaches can be taken to improve resource usage in the network and quality of experience for the user. These mechanisms have been categorized into traffic management building blocks. The motivation behind and some more details of the mechanisms are explained in section 3. Another purpose of the document is to derive initial operational, functional and architectural requirements that advanced traffic management subsystems of future mobile networks should implement. This is described in section 4. The results of this document should be used later on in the project to design and evaluate a traffic management framework comprising new innovations such as concepts, protocols and other interfaces to fulfil the requirements of future broadband wireless telecommunication networks.

2. Objectives for traffic management

As mobile and wireless communication networks move toward broadband converged networks and applications, the demands on the infrastructure increase exponentially. The increase of such wireless networks ensures that data, voice, video, and wide-scale of applications are running on the same network, simultaneously. High capacity LTE network will follow the same path as wireline networks in the past, and will thus become quickly dominated by CDN and P2P traffic. The high amount of wireless bandwidth available makes it possibly to provide capacities equal to typical PCs on broadband connections. On the other hand, it can be recognised that wired networks of mobile operators remain static regarding available capacity and throughput. The latter increases the need of smart traffic management (TM) solutions for wired backbone networks and for the terminal side to be able to run multimedia applications efficiently.

It will become essential that the network must be aware of each applications traffic type and enforce TM and control (i.e. priority, routing, bandwidth, etc.) required for ensuring improved QoE for every user most of the time. Ensuring that mobile networks are application-aware, can help to achieve flexible adaptation to any new application and traffic pattern emerging in the future. Operators need to install effective management tools to manage traffic using QoS policies, bandwidth allocation schemes, prioritised access and admission control, traffic shaping and rate control, and flow based processing. Only such advanced and active TM will ensure that operators can provide cost-effective data transfer with real-time speech and video on heterogeneous accesses.

TM forms part of the management process of operational networks with close relations to network planning and resource deployment. TM and traffic engineering cover all measures to dynamically control and optimize traffic flows in a network domain or in a global view of the interconnected Internet, aiming at ensuring a maximum throughput and sufficient QoS for the users. Therefore TM includes concepts for dimensioning, admission control, differentiation of services and failure resilience that should guarantee a well balanced load level for good performance in normal operation and maintain availability of important services for a set of main failure scenarios.

The objectives of TM for this project will be aligned with the definitions from the IETF [RFC 2702, RFC 3272] and 4WARD [4WARD] project, and will scope on:

1. **Traffic modeling and QoS experience**

Traffic modeling will constitute the basis of the proposed investigation. The main objective is to bring the model of traffic characteristics and the adaptive behavior of selected Internet applications as close to the reality as possible. To this end, the minimum requirements for constant QoE and corresponding minimum QoS requirements will be defined. This work is not in the scope of this document.

2. **Mechanisms for traffic control**

The procedures and mechanisms proposed for traffic management will form the main focus of the foreseen work plan. They will be divided into two levels: microscopic and macroscopic. For both levels an architecture containing the components for measuring and monitoring functions will be developed. From there, the relevant performance parameters will be derived, as well as the proper mechanisms for traffic management will be developed. These mechanisms will contain decision and coordination functions to select the best appropriate TM mechanisms. Investigation within the macroscopic and microscopic traffic management domain will mainly focus on proactive and reactive influencing of the traffic flows at protocol or packet level, aiming at a fair allocation of resources and taking into account the QoE requirements. Meanwhile, the macroscopic traffic steering will look from the higher perspective, at the mechanisms of access network and gateway selection, as well as the paths lying in between. For access network selection the scope of consideration will cover heterogeneity of different generation of networks (e.g., 2G, 3G, 4G), different standards (e.g., 3GPP, IEEE) and cell dimensions (macro-, micro-, pico- and femtocells). For the gateway selection process in order to achieve efficient and energy-optimal data transmissions, the influence of mobility management, the amount of data to be transferred, as well as the geographical location information and the user's profiles will be taken into account. Furthermore, the new mechanisms for intelligent resource selection should be developed. Consequently, improvements in the resource usage as well as the shortening of transport paths contributes to a reduction of active network elements (links, routers, switches) that does not only express in the CAPEX savings but also through the subsequent energy savings reduces the OPEX.

A comprehensive end-to-end consideration of the procedures and mechanisms for traffic steering and traffic management network optimization as well as the identification and exploitation of the dependencies will form the integral part of the planned work.

3. Techno-economical analysis

The techno-economical analysis focuses on the examination of the economic efficiency of the technical solutions foreseen in this project. Models provided within the scope of this analysis for evaluation of the cost structure will take into account not only the investment costs (CAPEX) for the network infrastructure but also the operating expenses (OPEX) as for example management costs, servicing costs and location costs as well as the energy consumption. Following, the studies will include network migration and business scenarios analysis.

Additionally, objectives for TM could include as well:

Good user experience for the applications running on top of the mobile access network. This could be defined according to a threshold values for the mean opinion score (ITU P.800).

Minimization of network cost:

- Low CAPEX: There is a considerable challenge to satisfy user experience with minimum of infrastructure resources and still be flexible to handle the possible large variation of traffic patterns over time. Low CAPEX is only possible, if the resources consumed by application flows are actually used efficiently. Inefficient usage of resources may occur, if for example an application may have allocated resources within the EPS but the user experience is not sufficient for some reason.

- Low OPEX: There are different possibilities to keep OPEX low. On one hand interconnection cost could be quite high, if identical content is repeatedly transferred from external or otherwise un-preferred external locations. Another example is the application of strategies, which can select appropriate traffic management mechanisms without human interaction.

Maximization of revenue: The operator may decide to steer user traffic in a way that helps maximizing revenue. For example some value added content might be available within the local network instead of an external CDN.

Further details of the techno economical analysis are out of scope for this document.

3. Traffic management building blocks

The building blocks described in this section provide an overview about potential mechanisms to improve quality of experience for the user and to enable efficient usage of infrastructure and IT resources. For the latter there is a benefit for other stake holders in the (mobile) communication business, such as communication service providers and mobile network providers (MNO), content providers and content distribution network CDN providers. We have identified six different categories, which can be used to assign the various building blocks and aspects. ***Figure 1*** displays the principle building blocks for traffic management with some relevant examples, which might be further investigated within the project. There are three blocks, which may be correlated with each other similar like lower layer functions providing services to higher layer functions in a communication stack.

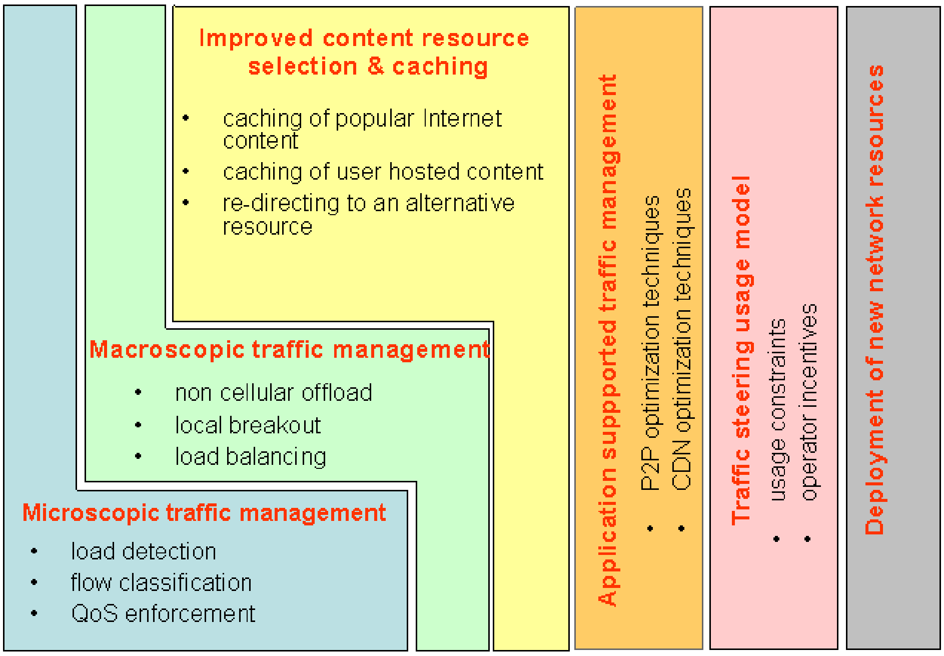


Figure 1: Traffic management building blocks

We introduce the new terms microscopic traffic management (MicTM) and macroscopic traffic management (MacTM) with the definition of traffic management building blocks. MicTM 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. MacTM includes all mechanisms with the primary objective to improve efficient usage of network resources. Parameters for optimization in the latter 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 (IRSC). 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. A resource in this context is associated with specific (multi-media) content, which is requested by users. All the above mentioned categories are associated with mechanisms, which may require support from lower layers (below application). In addition, we have identified three more building blocks which may require only little or no support from MicTM, MacTM and IRSC... These could be in place without dependence to other traffic management building blocks. On one hand there is application supported traffic management. There are many applications based on CDN and P2P, which try to optimize performance from end user perspective without getting support from network elements. Another building block has been identified, which is more relevant from business perspective without too many technical aspects. Mainly network operators but possibly other stake holders as well may influence user behaviour by defining certain constraints for usage of networks / services and certain incentive to comply with the usage constraints.

Finally we have defined one further building block, which 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. In contrast to the other building blocks mentioned previously, capacity extension is a process which will become effective in the network after a long time period, possibly up to several months. In addition to the building blocks there are some common function like policy control and traffic monitoring. How the building blocks are related to each other and how they interact with the aforementioned common control functions shall be investigated within follow up activities.

In conjunction with defining traffic management mechanisms associated with the building blocks, we want to introduce a simplified role concept in order to understand the challenges, problems and requirements associated with the different stake holders. These are the following:

- User – the person, which attaches to a network via user device(s) and consumes content from a remote resource.
- Communication service provider (CSP) – operator of an access network. Due to the focus of the project there is a special viewpoint on mobile operators
- CDN Provider – usually a company with large amount of infrastructure to distribute content as close as possible to the end user. Customer of the content provider is usually the content provider.
- Application / content provider – commercial or non-commercial entity, which inserts content for global or limited use in the Internet.

The roles of CSP, CDN Provider and content provider could be intertwined. For example, Google provides content but also has deployed a huge infrastructure to deliver it to the end user. On the other hand there are some operators, which want to make additional business by deploying CDN infrastructure or provide services via “walled-garden” business models. Hence end users should be motivated to connect to resources within the CSP or associated domains.

Each of the following subsections contains a defined set of structural elements. The assumptions are used as a basis for the general description and followed by the problems statements, which can be derived from the previous parts. The following points of view have been considered within the different descriptions:

- End user perspective, which includes user behaviour and user expectations. Some examples are the change of access by a user or group of users, excessive download by a user or hosting of content on a user device.
- Aspects related to the application: It needs to be clarified whether specific solutions are needed for the different applications. In some cases a differentiation is necessary between application client and application server. This includes also aspects related to the content provision and delivery.
- Communication service provider (CSP) perspective with a special emphasis on mobile network operators. Examples include planning / deployment of network capacities, switch on / off equipment but also principles of interworking with other providers (for establishing service level agreements, network resource sharing, etc.)

3.1 Macroscopic traffic management

Macroscopic traffic management include examples such as change of routing and dynamics for network infrastructure. Change of routing for instance may involve alternative routing paths in the mobile core network and in the backhaul.

3.1.1 Access technology reselection

Originally, the design of WLANs emphasized data service support in contrast to cellular networks focusing on voice services. This segregation is no longer applicable, nowadays WLANs and cellular networks have to support a complex mix of services. Furthermore, modern user devices are equipped with multiple network interface cards (NICs) and are thus able to use any of several available access technologies. Additionally, access providers extend the range of their products: it becomes quite common that a single provider is offering access to cellular and WLAN networks as well as to other technologies.

The issue of access technology selection becomes more and more interesting, as it is no longer uniquely determined by either contract with the service provider or by the type of used service. For example, it might be attractive to relocate a voice user holding a call in a cafe from the cellular system to WLAN—in order to release the cellular capacity for other users. On the other side, if the cellular network has enough capacity, it might be interesting to shift a voice user from WLAN to free more capacity for data transfer.

We will consider systems that consist of a WLAN access cell connected to a second heterogeneous access technology (denoted as ”AT2”). Thereby, the WLAN cell is completely within the coverage area of AT2. We assume that all mobile terminals are equipped with a network interface card for each access technology such that they can perform vertical handovers. Secondly, we assume that each access network has an own cost function reflecting the effort and the revenue serving a specific user who belongs to a certain traffic class.

For the handover decision process, the involved access networks are divided into two conceptual parts, namely the originator and the recipient network. In principal, there exist three general concepts regarding the placement of the handover decision. This can be realized within the originator network, the recipient network, or by a separate arbitration entity. In the following, we discuss the outstanding tasks for a handover if the decision is made in the originator network.

There, the originator network

- firstly identifies potential handover candidate(s),
- secondly estimates the gain due to the potential handovers
- thirdly requests cost function value estimates from the recipient network via appropriate means of signaling
- fourthly compares candidates’ cost function values within originator and recipient network
- finally decides for or against a vertical handover for each candidate.

Contrary, the recipient network firstly estimates the cost function value for each potential handover candidate currently served by the originator network, and secondly assesses the impact of a handover on other users.

Access technology re-selection likely will be further investigated within the project.

3.1.2 Selection of core network elements

The Evolved Packet System (EPS) supports resilience (through Network Element (NE) redundancy), optimized routing, balancing of data plane and control plane traffic load in the NEs, and sharing of eNodeBs among mobile operators. These features are enabled by

- the separation of control and data plane functions in different NEs: Mobility Management Entities (MMEs) and Serving Gateways (S-GWs);
- grouping of MMEs and S-GWs, which allows an eNodeB to be connected to one or more MMEs and S-GWs in a pool; and
- flexibility in Packet Data Network Gateway (PDN GW) selection.

Core network element selection likely will be further investigated within the project.

3.1.2.1 Initial selection of core network elements

In LTE access, when a UE attaches to an eNB, the eNB selects an MME to serve the UE, based on the UE’s previous attachment history and possibly based on MMEs’ capacity/load status and other vendor’s or operator’s configuration information ([3GPP_23.401] defines a procedure for the MME to indicate the load status to the eNB). The MME then selects an S-GW to serve the UE based on similar criteria and on the current tracking area. The MME also selects the PDN GW based on the supplied APN (Access Point Name), topology information (collocation with or closeness to S-GW), mobility protocol, and other subscriber information.

In gateway selection, the advanced features of DNS are extensively used (e.g. NAPTR and SRV resource records) [3GPP_23.261]. This involves adaptation of 3GPP concepts to a form that is suitable for DNS, which makes the configurations burdensome and error-prone. This is exacerbated with the increase in the number of gateway functions in the future distributed architectures and in various local breakout scenarios (see section 3.1.4).

The above-mentioned mobile NE selection procedures implicitly affect packet routing in the whole mobile network and they involve route optimization (e.g. in the form of topology awareness and in S-GW re-selection, in case the UE crosses a tracking area). However, the selection rules do not directly take into account the current load and failures in the transport network (which are out of the scope of the 3GPP specifications anyway). This leaves room for mobile network vendors and operators to devise more sophisticated NE selection criteria that allow load balancing among transport resources. To make this possible, the mobile NEs need to become aware of the available capacity in the transport network.

3.1.2.2 Reselection of core network elements

In the future EPS architecture the reselection of core network elements might become more important. This especially holds for distributed architecture building blocks where the gateway nodes are located closer to the access network. For highly mobile users (i.e. users which stay within the same radio access network or radio cell not more than a few minutes) the original user traffic routing paths become inefficient, and a gateway node reselection would be favourable. Also new mechanisms like IP traffic offloading (see section 2.3.4.1) take advantage of efficient methods to reselect user plane core network elements (S-GW, PDN-GW). MME reselection is less needed, because the control plane traffic has a much lower share of the overall traffic compared to the user plane traffic.

S-GW reselection is standardized since Release 8 (see [3GPP_23.401]). It can be performed either during a handover (if the UE is in active state), or during a Tracking Area Update (if the UE is in idle state). Currently the main reason of reselecting the S-GW is the mobility of the user. During a handover or Tracking Area Update the MME decides, if the change of the S-GW is necessary. This might be the case because the UE has moved into a tracking area which can’t be served by the old S-GW. Other reasons for S-GW reselection, like load balancing, are not specified in [3GPP_23.401] so far.

Since Release 10 the relocation of the PDN-GW is standardized as well (see [3GPP_23.401]). The current procedure is that the MME triggers a PDN disconnection procedure with a reactivation request after a Tracking Area Update. This means that the connection to an APN through a PDN-GW is shut down and a new connection to the same APN through a new PDN-GW is established afterwards. This method doesn’t allow a seamless change of the connection, because the IP address changes as well. In general, this procedure can be performed at any time. However, it is convenient to perform the PDN-GW reselection when the UE changes its state from idle to active. The trigger for this relocation procedure is when the MME determines that another PDN-GW is more favourable to serve the UE during a Tracking Area Update..

3.1.2.3 Coordination of access network and core network element selection

The optimum selection and reselection of core network elements might be dependent on the currently selected access network - e.g. when changing the access network, it could be useful also to change the anchoring P-GW to ensure an optimum routing path through the network Therefore a function which coordinates the different selection procedures might improve the overall system performance.

3.1.3 Change routing within backhaul

Mobile backhaul refers to the transport network and the equipment used for connecting together mobile system elements in the RAN to those in the mobile core (in LTE, connecting eNBs to MMEs and S-GWs). Routing-related considerations include the following:

- the transport service type and the underlying physical layer topology have to match with the connectivity requirements (non-optimal routing paths should be avoided to make efficient use of network resources and to minimize packet delay);
- redundant routing paths need to be provisioned in order to survive from node and link failures;
- traffic flows requiring special forwarding treatment may be routed via separate routing paths;
- dynamic re-routing of traffic may also involve re-allocation of transport capacity, if such flexibility is available;
- routing changes within the backhaul transport should be coordinated with any load-balancing decisions among mobile NEs, e.g. re-selection of core NEs (section 3.1.2); or traffic offloading or (forced) handover within or across radio access technologies (section 3.1.4).

In the traditional tiered mobile access networks, the routing topology has been a tree rooted at the controller site. This has changed since the introduction of distributed and pooled mobile NEs and local breakouts. Also user plane traffic between eNBs is expected to increase in the future, beyond the temporary tunnelling over X2 in handovers. Therefore, multipoint-to-multipoint (MP2MP) transport service is desired between mobile NEs instead of the traditional point-to-multipoint (P2MP) connectivity. Still, during the migration phase only P2MP service may be available because typical physical topologies in the lower access are still trees.

The mobile NEs, as customer equipment to the transport network, have only limited means for affecting routing within the backhaul (such that the re-routing would not involve re-selection of mobile NEs). If the NEs are multi-homed they can use alternative IP source addresses (i.e. alternative egress interfaces) or alternative IP destination addresses in Stream Control Transmission Protocol (SCTP), in S1 control plane. In Ethernet transport, the NEs can tag frames with one or two Virtual LAN (VLAN) Ids. The tags may designate different operators, traffic types, and/or service classes, which may be mapped to separately routed Label Switched Paths (LSPs) in an MPLS-based Virtual Private Network (VPN). Here the scalability with respect to the number of VLAN Ids is an issue. For splitting of TCP flows across multiple paths, see section 3.2.2.2.

A traffic routing/steering solution that selectively moves traffic to alternate routes should be aware of the traffic load and available capacity along whole end-to-end paths between the UE and PDN GW. Therefore, there is a need for co-operation and coordination between the backhaul transport network and the radio network layer. The concurrent steering mechanisms might be categorized as belonging to the network's self-organization in one extreme and to traffic policing ("routing" packets to the garbage bin) in the other extreme. One of the main success criteria for the solutions is the end users' QoE, which is hard to measure and which implies that traffic steering needs to be application-aware.

Routing change within backhaul likely will be further investigated within the project.

3.1.3.1 Routing in a first mile wireless mesh access network

In the first mile access, microwave radio (MWR) technology is a viable option in cases where the deployment of fibre is not profitable for the operator. Moreover, a MWR mesh network may fulfil the identified MP2MP connectivity requirements. In this scenario, the wireless mesh network (WMN) forms a self-organizing routing area with traffic engineering capabilities. This requires a routing and wireless link scheduling solution that adapts to changes in traffic load and to varying environmental conditions (like degradation of wireless link capacity due to rain). Traffic flows that require expedient forwarding are routed along low-delay primary paths while best effort traffic can be directed to secondary paths. For resilience, redundant links in the mesh are overlaid by node-disjoint routing paths, and the WMN must be connected to the wireline aggregation network through several gateways.

In the WMN, the available link capacity changes depending on radio conditions (modulation and coding changes), adaptive MAC scheduling, and load balancing. Therefore, there is a need for mechanisms that improve mobile NEs' awareness of the existing transport capacity in order to facilitate the NEs' own access control, congestion control, and load sharing decisions.

3.1.4 Offloading techniques

Traffic offloading refers to techniques of distributing users' data traffic through localized wireless access points (femtocells or WLAN) and by locating service gateways (breakout points) near to those access points in order to avoid non-optimal routing and needless overloading of the network elements.

In 3GPP, there are parallel efforts to address traffic offloading in LTE and UMTS. The LTE variants of traffic offloading are examined in the following (see also [WE2E]). For more detailed 3GPP-defined LIPA and SIPTO scenarios and architecture options the reader is referred to [3GPP_22.220], which is part of Release 10. Although the examined solutions are specific to LTE they deal with traffic routing requirements that are relevant in any future mobile network architectures.

Some of the 3GPP based offloading techniques (LIPA, SIPTO, IFOM) likely will be further investigated within the project.

3.1.4.1 LIPA, SIPTO, and IFOM

Local IP Access (LIPA) [3GPP_23.829] provides access to a residential/corporate local network interconnected to a femtocell, called Home eNodeB (HeNB). LIPA allows access to a subnet within home or an office for shared resources like printers and media servers while a mobile device is attached to the 3GPP operator's network.

The LIPA scenario addresses the need for the user to have access to resources connected to a local network in a home or office environment. In most cases this would provide Internet access as well that would offload in addition to the operator Radio Network the Core Network too.

This is made attractive with the modern UE devices that provide considerable computing power and enhanced displays. LIPA is a proposed method for making local IP network access seamless for the user when moving between LTE macro and home network.

The LIPA scenario is somewhat similar to the access technology reselection scenario depicted in section 3.1.1, but the incentives for using alternative access networks are different in these scenarios.

LIPA is based on the presence of a HeNB and a co-located Local Gateway (L-GW) function in a residential or enterprise network. The LIPA scenario is shown in *Figure 2.*

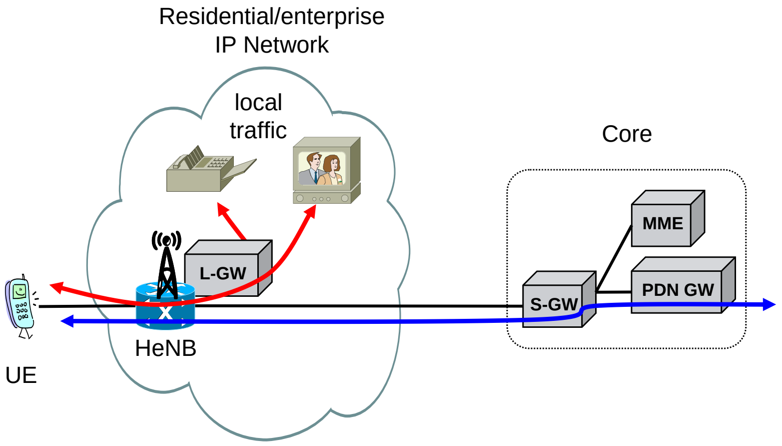


Figure 2: LIPA scenario

L-GW includes a subset of PDN GW functions, which brings some additional implementation complexity in the base station equipment. The home network has to connect to the core network through a Security Gateway (not shown in *Figure 2*).

A discovery mechanism for LIPA capability must be built into the architecture, and dedicated APN need to be assigned for the L-GWs. LIPA introduces also a number of changes in existing mobile NEs (HSS, MME, and eNB) and in the existing signalling protocols. For example, the MME must be involved in L-GW selection. Scaling for large number of customer premises devices can be a significant burden. Mobility among HeNBs is currently defined in 3GPP Rel.11. Also Lawful Interception (LI) for local access is difficult to arrange. The success of LIPA depends on large scale deployment femtocells for capacity and coverage enhancement, which has not happened so far.

Selected IP Traffic Offload (SIPTO) [3GPP_23.829] provides the ability to selectively forward different types of traffic via alternative routes to/from the UE with the objective of reducing load in the primary access and core network paths. There are two flavours of SIPTO, each having the breakout point in different locations in the network:

- SIPTO for home network assumes the use of a HeNB. In this mode, specific traffic determined by operator policy and/or subscription is transferred to/from HeNB directly to Internet/Intranet bypassing the mobile operator’s access and core network.
- SIPTO for macro network covers the ability to offload traffic at or near the S-GW as opposed to traversing the operator's core network (PDN GW).

SIPTO for HeNB is similar to LIPA in terms of necessary changes to network capabilities and additional functionality in the form of a L-GW collocated with HeNB. It does not provide any substantial benefits as compared to LIPA. Mainly for HeNB without LIPA support, 3GPP considers a SIPTO solution in Release 11 that might be similar to SIPTO for macro cells/network.

SIPTO for the macro-cellular network is a feature that many equipment vendors already provide for UMTS in some form or another. The idea (in LTE) is to select a PDN GW that is typically located closer to RAN (*Figure 3*) and allows to offload the traffic to another (transport) network operator instead of routing to central GWs in the EPC.

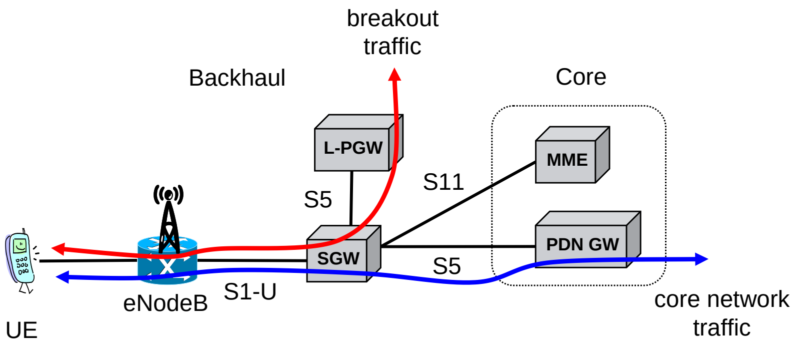


Figure 3: SIPTO scenario

Solutions are needed to allow the user access to the new APN (in the subscription) and to map the APN to a gateway (e.g. by using DNS) to support gateway selection. Since SIPTO for the macro network allows traffic breakout only above the RAN, it does not reduce traffic on the radio interface or the backhaul.

SIPTO requires minor changes in the existing HSS and MME functions. For example, SIPTO does not require any special support from the UE. Moreover, it has only minimal impact on mobility management and no impact to Lawful Interception. The MME mobility management function may check in case of UE movements (tracking area update) whether the locally selected GW becomes non optimal for routing and select a new PGW. This forces the UE to re-establish the IP connectivity and is performed in IDLE mode.

IP Flow Mobility (IFOM) [3GPP_23.261] IFOM selectively assigns different traffic flows originating from a UE to separate radio access network connections that represent different technologies (e.g. LTE and WLAN). Network-based solutions are now being addressed for IP flow mobility for GTP based S2a and S2b interfaces. This may have significant impact on traffic engineering possibilities inside the network based on information available in the network and not available at the device side.

The current smartphones offer only limited support for simultaneous access to networks with different technologies (e.g. restricted to sending MMS over the macro network while being connected to WLAN). Moreover, the integration of packet core and WLAN access does not yet support session continuity across technologies. IFOM provides simultaneous attachment to alternate access networks and allows fine granularity of IP flow mobility between access networks, *Figure 4*.

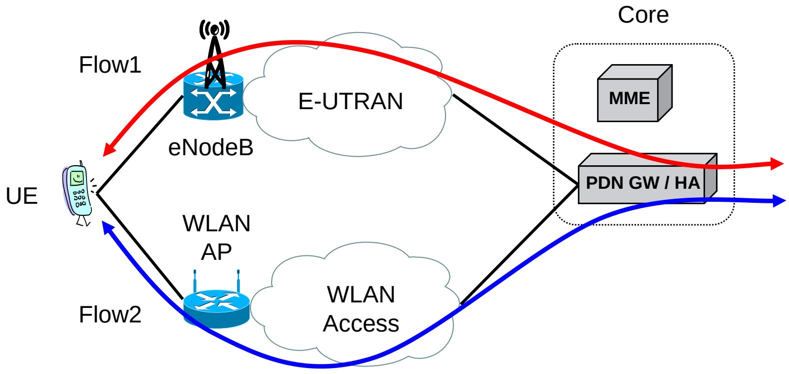


Figure 4: IFOM scenario

In LIPA or SIPTO, route optimization is based on the use of special gateway nodes in the mobile network. In contrast, IFOM relies on the UE’s capability of simultaneously attaching to both LTE and WLAN access networks (or in general, any different access networks). Using IFOM, it will be possible to select particular flows in a UE and bind them to one of two different tunnels between the UE and the Dual Stack Mobile IPv6 Home Agent (DSMIPv6 HA) that can be implemented within a PDN GW. To achieve this the UE must implement DSMIPv6 functionality defined by the IETF [RFC5555]. DSMIPv6 supports the creation of multiple tunnels between the UE and the Home Agent (HA). The dual-stack capability allows access to both IPv4 and IPv6 content without Network Address Translation (NAT).

The main drawback in IFOM is the need for upgrading UEs to support DSMIPv6 (besides the need for implementing HA functionality in PDN GW). It is still open how to map between flows and networks based on different criteria like cost, QoS, security, and user preference.

3.1.4.2 Traffic Engineered HandOvers

The traffic contains an increasing part of highly demanding applications such as video streaming. On the other hand, mobile terminals still have limited capabilities in energy and signal strength where as the radio resource will always remain naturally limited. Relieving the RAN of traffic burden is a fundamental objective and offloading techniques enabled by LTE a major contribution.

As written in the previous sections: “routing changes (within backhaul) should be coordinated with load balancing decisions among mobile NE e.g. re-selection of core NEs or forced handovers between base station or radio access technologies” (see section 3.1.3); “It is still open how to map between flows and networks based on different criteria like cost, QoS, security, and user preference”. This stresses the need for Traffic Engineering (TE) mechanisms for NE selection that integrate various criteria such as QoE and network state awareness, associated models, policy and user preferences.

This integration requires the involvement of TE components associated to NE in the RAN but also in the backhaul and core network.

This section focuses on RAN-based decision mechanisms to be involved in cell or RAT changes. They are referred to as Traffic Engineered Handovers (TEHO) some of them can be involved in route selection.

Traffic Engineered Handovers (TE HO)

In an LTE network, the handover process is usually based on the measurement of the mobile equipments and the HO is usually initiated by a significant change in radio conditions (heavy interference, weak signal etc.) sensed by the mobile. The effective decision on the HO initiation and execution is made by the network. .In spite of this the HO can be initiated for different other reasons such as excessive cell load, QoE degradation or offload to another RAT, e.g. HO of a flow to a WLAN covered by the cell. .

The concept of TEHO covers the HO initiated by a set of one or more factors that include more than only the usual triggering factor which is the degradation of signal quality received by the UE. TEHO is typically triggered and managed by TE entities located in and beyond the RAN. TEHO can also be referred to as Forced HO.

A TEHO results in a movement to another cell, involving all the flows of a connected UE, or to another radio access network involving one or several application flows of the UE. A HO can result in the change of core NE such as the SGW and the PGW. The selection of the target cell therefore involves information at the RAN level but also at higher levels.

Examples of flow-based HO

Different types of traffic are transmitted in the network with different priorities. In a modern wireless and mobile telecommunication network many types of traffic have a different priority, for instance voice call, video call, file transfer, interactive data etc. Another important case is the emergency call for which proper transmission has to be ensured even in an overloaded network. In such cases, the HO is initiated by the network and based on the QoS and priority requirements associated to the flow application.

Information involved in TEHO:

The parameters of the TEHO solutions are related to both the signal quality and the traffic. The traffic demand has to be considered as well as the available free capacity in the involved access networks. Depending on the relation of these two values and their distribution in the network the usage of priority parameters may be necessary. Also, the QoE sensed in a UE and its attached preferences may be considered by the network to initiate a TEHO. The final decision may also involve policy rules. So the TEHO decision stake holders are not restricted to the RAN. Likewise, RAN-based information used in TEHO may also be involved in decisions such as selection of core network elements.

TE solutions that make the HO procedure more QoE and network state aware involves the circulation of additional information w.r.t. signal quality. Therefore they must on the other hand minimize the additional Control Plane procedures

Requirements for the design of Traffic Engineered HO solutions

TEHOs can significantly increase the transmitted control information in the core network. Thus for an efficient traffic engineering solution appropriate handover algorithms are needed to keep the additional control traffic at a low level and the following traffic engineering problems need to be addressed:

- Specification of transmission costs to and within a given access network, involving multiple decision metrics.
- Associated decision algorithms to trigger a TEHO and to select a target cell or other RAN.
- Appropriate management of TEHO, given the triggering conditions and policy rules.

We consider the following observations and their associated challenges with the subsequent solutions that need to be designed to face them.

1. Handovers (HO) can be initiated by signal degradation, cost reasons, but also a threat of congestion or a threat on the QoS conditions.
    - Therefore, the decision metrics initiating the HO should include signal and interference status on one hand and also network state information such as cell load and QoS metrics.
2. The scanning of candidate target cells is particularly costly in time, energy and resources. The number of candidate cells ever increases in hilly or dense urban areas..
    - Therefore, the choice of target cells should be accelerated by reducing scanning operations. One way is to avoid useless scanning of neighbor cell stations, for example, the cells to which the terminal is unlikely to move to or to HO successfully.
    - The path of a MT is not deterministic. So statistics should relate the direction the MT comes from with the chosen cells and the associated HO success rate, in order to give advantage to the cells that have a higher probability to be selected.
    - Another important information is the geographical position of the requesting terminal. The systems should enable cell ranking according to position attributes such as location and navigation information, or 3D position of the requesting terminal.

3. Handovers can be initiated for several reasons and involve various decision parameters. Therefore:
    - a robust HO Decision Management function needs to be designed, to support various triggering factors (signal, load, QoS needs).
    - This HO Decision Manager (HODM) located in the network should maintain: (i) service continuity through signal quality, (ii) QoE continuity through QoS-aware handover, (ii) load balancing in the controlled access network to maximize the success of (i) and (ii).
    - It should maintain seamless HO whenever possible, in particular in inter-RAT HOs.

The 3 considerations above advocate for an HODM that supports Multi-Criteria Cell or access network selection, according to the different possible triggering factors.

3.1.5 Switch on / off equipment

The amount of traffic in wireless communication systems is getting higher and higher; however, it is not uniformly distributed in time and spatial dimension. In addition a lot of energy is used for backup resources. This fact may leads to a requirement to switch off some elements of the network and to use different traffic routes in given periods. This method, called equipment on/off switching, can be used not only for energy saving purposes but also for traffic management solutions. We also can mention that the on/off switching is related to the power control, however it is not really special case of the power control (i.e., when P=0). The on-off switching includes not only the switching of the radiated power but it also affects the consumption of other elements of the equipment. Therefore it is more suitable for saving purposes.

The on/off switching is also related to the forced handover mechanism. When an on/off process is conducted in the network an optimization process is performed in order to ensure advantages in, for instance, traffic engineering point of view. After the optimization the results must be enforced. This solution is ensured by the mechanism of the forced handover. The whole process consists of four steps. In Step 1 the optimization is performed and the nodes for switching on/off are determined. According to the result of the optimization the nodes being necessary for the new traffic situation must be switched on in Step 2. Then the adequate forced handover commands are ordered in Step 3, which also means traffic redirections. Finally in Step 4 the nodes without any load must be switched off.

When introducing a new solution in a network, one of the most important viewpoints of the operator is the installation and the operational cost. In this case the “cost” can include not only the financial expenditures (capital expenditures – CAPEX and operational expenditures – OPEX), but also has a general meaning. Based on special demands, the operator can create a cost function that contains several parameters with different weights according to the requirements. For instance, parameters can include energy consumption, system capacity or service outage. After the determination of the cost function, the next step is the optimization process in which the operator’s purpose is to find the optimal or, in most of the cases, the sub-optimal solution.

In order to perform the optimization process, the statistics of the traffic in a given network segment or at least the estimation of it is required. For instance, the daily statistics of data traffic generated by the mobile subscribers can be used for making an approximation of the network load in given periods. This also defines active users’ percentages as a function of the time. The subscribers connected to a segment of the network can be modeled with traffic aggregation points. Therefore it is also possible to model the distribution of the network load over the segments. Based on this theory and by the usage of the cost function, the optimization, e.g. the on/off function of the network nodes can be performed.

The mentioned general optimization solution can be applied for different goals. The cost function (or in other terminology: the objective function) can be defined for traffic engineering purposes. Of course, when designing the proper objective function several other theories should be applied such as graph theory, etc. Assuming that the adequate cost function and optimization method are used, their application will also result in better performance of the overall network via better load balance, more efficient exploitation of network capabilities and advantages in traffic engineering issues.

The next important parameter is what main segment of the network is involved in the optimization process: the access network or the core network. As the nodes of the access network are much closer to the elementary traffic generator entities, the relative fluctuation in their aggregated traffic will be higher than in the core network. This fact can significantly affect the on/off switching decisions. Related to this idea the fluctuation of the aggregated traffic is also influenced by the level of network hierarchy of a given node. For instance, the traffic fluctuation in an eNodeB significantly differs from the fluctuation in S-GW. The evolution of the network architecture is also involved in this problem. To illustrate this issue one can mention that a network with flat architecture requires different solution compared to the conventional hierarchical architectures.

Continuing the previous chain of ideas one can say that the deviation in the data traffic has also a time related parameter, the frequency of variations. The operator has to decide which periodicity is required in on/off switching. According to this decision the proper threshold values have to be defined which is not a trivial problem to solve. Moreover, the operator has to consider the equipment abilities as the switching process is usually rather slow in network elements; however the latest products perform quite well in this aspect. The fast switching is very advantageous in case of energy saving purposes as many short periods of low consumption idle states can be summed up in significant savings. But one has to consider that the possible rate of switching also depends on the responsiveness of the applied traffic routing algorithms if the redirecting is required in the given situation. Finally a physical problem also has to be considered, i.e., the type of the applied power amplifiers of the used devices. Efficiency of these amplifiers is highly related to the switching periodicity and the amplifier load. For instance, many types of power amplifiers perform better in case of heavy load, while frequent switching deteriorates their efficiency. We should denote that both facts are in contrast with the original energy saving purpose. Therefore the system designers also have to find the trade-off in this aspect.

In some thoughts we have to return to the expenses. As mentioned before, basically these can be divided into two classes: capital expenditures (CAPEX) and operation expenditures (OPEX). The CAPEX includes costs of new equipments, their installation or special software (containing the on/off algorithm), respectively. The OPEX consists of cost of energy consumption, maintenance, necessary software upgrades, etc. We should notice that, for some reasons, the operator can decide to neglect the CAPEX, for instance, and perform the optimization only based on the OPEX. In this case the CAPEX has zero weight in the mentioned cost function in the general model.

A special question for the operator is how to handle the CAPEX, when there is an investment related to the network development being necessary to keep the system up-to-date and which is independent from the on/off developments. In accordance with this problem one can state that the mobile network operators serve the subscribers with different technologies, moreover nowadays the question of co-existence of the different networks becomes more and more important. In the latest standards the spectrum aggregation (using different frequency bands with common control) is one of the most popular radio technologies. Therefore in case of on/off switching solutions the technology realization in a heterogeneous network is also a challenge. For example, a general question can be how to perform the switching in a co-existing GSM and LTE system in order to transmit data and voice traffic in optimal (or sub-optimal) manner.

The standards that describe the different communication systems eventually define frames for operation and determine special network elements and functions that have to be realized in order to ensure proper operation of the system. For example, 3GPP standards define NodeB, RNC, SGSN, GGSN, HLR etc., their functions and interfaces for the 3G networks. However, if someone builds a network only by the use of the defined elements, we can say it is a rather theoretic solution. The real systems of the operators contain many other auxiliary devices with different roles and functions, for instance, aggregator or repeater nodes. When calculating the necessary throughput, energy or other parameters in the on/off switching algorithm these elements also have to be taken into consideration for realistic assumptions on network operation.

Finally the following simple theoretic example illustrates parameters and circumstances that have to be considered when designing an on/off algorithm and its cost function, respectively. Let us assume a small segment of a mobile cellular network with 7 cells. According to the measurements we know that the subscribers use only 10 percent of the capacity in a given period of night, in contrast to the peak hours when the load is almost 100 percent. The idea is to switch off 6 of 7 base stations at night and raise the power level of the remaining one. This idea also implies tasks to solve in traffic engineering. The mentioned operation is shown in the *Figure 5* (the raised power cell is in the center, blue colored):

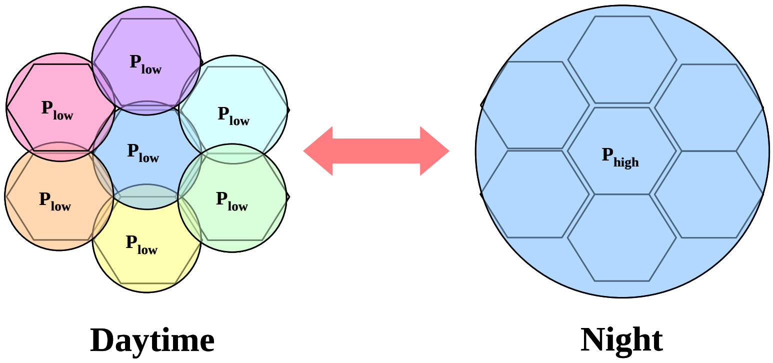


Figure 5: Example scenario for switch on/off equipments in cell level

The question raises whether it is worth to realize this kind of on/off mechanism. The operator has to create an adequate cost function. The trade-off is between the raised power required for the traffic and the control power of the switched-off base stations. When the load in the network is higher, more base stations have to be operated in order to achieve higher traffic density (higher overall throughput). This situation also involves higher power consumption and higher operational costs. In spite of this, however, the overall power consumption will be lower when some of the base stations are switched off but only lower traffic density can be ensured.

The relation is not linear between the signal attenuation in the radio channel and the received power. It is approximately $P \sim \frac{1}{r^n}$, where r is the distance between the communicating parties, P is the received power (being the function of the transmitted power, radio channel attenuation and some other parameters, e.g., antenna gain, cable loss etc.) and n is the propagation exponent (of which value is usually between 2 and 4 in realistic situations).

Because of the mentioned reasons above, the operators have to optimize non-linear functions, define thresholds and determine which base stations can be switched off even in this simple case. However the realistic situations are much more sophisticated. Moreover, the restrictions of frequency planning, minimum required bit rate for expected quality of service and other issues also have to be considered. Therefore one can state that the on/off switching algorithm has to be planned carefully in order to operate properly in a real network. Moreover, as mentioned before, from traffic engineering point of view the usage of on/off switching methods are also advantageous. Therefore on/off switching of equipment likely will be further investigated within the project.

3.1.6 Cell site re-configuration / modification of cell footprint

Even though these aspects are not investigated in detail within the project, the mechanism is described here to give a complete picture of traffic management mechanisms. The following methods are mainly related to the access network; however these also have effects on the backhaul and core network traffic, also depending on the given architecture.

3.1.6.1 Cell diameter (via power control)

<u>Assumptions:</u>

Cell diameter in a given service area is basically designed with fixed parameters assuming a given number of subscribers. However, the increasing number of subscribers or increasing demand on the overall cell capacity is a valid assumption. The increasing number of subscribers can result in varying required cell range. In a deliberated design solution, the importance of uplink transmission also has to be considered such as the low signal-to-interference problems at the cell edges. The varying cell sizes cause traffic redirections between neighboring cells. Regarding the core network the mentioned redirections have significant effect in border regions (areas) of the network, i.e. where the affected eNBs belong to different SGWs. Therefore traffic redirection is also requested in the core network.

<u>Mechanism description:</u>

As the number of subscribers is getting higher, the rate of fluctuation is also increasing. As a proper load balance among the cells can significantly increase the overall throughput in a given area, this is an important aspect for traffic engineering. The diameter of the adjacent cells with lower load can be increased and diameter of higher loaded cells can be decreased meanwhile the appropriate handover process has to performed by redirecting subscribers in lowered range cells to the raised range cells. The phenomenon is called cell breathing.

Moreover, the cell edge area is rather problematic as the signal-to-interference ratio is much lower as in case of the most cellular systems (the distance of the user from the connected base station is higher and the interferer is closer). Besides the downlink radio parameters, the uplink conditions can also become relevant in case of many types of communications, for instance if symmetric data transmission is assumed as in a videophone application, the importance of uplink transmission is equal to the downlink. It is worth to consider what direction (up or down) will bottleneck the transmission in a given situation.

Problem statements:

During the solution of the traffic engineering problem, the engineer has to determine where the most critical problem arises: In downlink or uplink. This problem highly depends on the application over the cellular network. The problem is even more difficult in cell edge regions. This implies that it is worth to differentiate the cell edge regions and cell centre; however it requires further analysis.

In case of cell diameter determination the following problems are recommended to examine and solve regarding the traffic engineering:

- examination of cell load balancing solutions in case of significant traffic demands in adjacent cells
- dynamic detection of minimal required DL power for each user to be subjected lower co-channel interference from adjacent cells – resulted in smaller non-overlapping cells to provide higher bit rates at cell edges
- examination of uplink interference (besides downlink considerations), strict control of uplink power avoiding serious uplink interference in a less-analyzed environment up to now (assuming serious increment in importance of uplink connections)
- statistical examinations of cell size deviations
- management of uncoordinated cell insertions (see IR1.1 Section 3.4) (e.g., addition of femto cells in a residential area where radio interference occurs by existing radio cells). Uncoordinated means that no previous radio planning is performed. (also may be controlled by coordinated beamforming)

Criteria used for description:

The solution requires realistic data and its statistics about the traffic in the neighbouring cells and cell range parameters in a given region. The mentioned statistics describe the traffic change in time, so various periods with different lengths can be examined in order to give an adequate description of the traffic. Without these realistic parameters only theoretical assumptions can be established and in aspect of practice these calculations will not give real applicable results. Unfortunately the operators handle this data confidentially and usually give it to a third party for research purposes only in a limited form. This means that only a part of the needed data is available, the data is filtered (e.g. because of law issues), moreover the usage and results are restricted by non-disclosure agreement.

General characteristics of the solution based on the aspects of RFC-3272:

- usability: the solution can be used in the existing cellular network as the power control mechanism is used in the base stations and the cell breathing mechanism can be interpreted as a result of the properly applied power control solution (the higher radiated power, the higher cell diameter and vice versa). However uplink power control (e.g. for cell range change) can be critical
- automation: the power control mechanism is automated in cellular networks but the proper operation of coordinated cell-size modifications requires sophisticated algorithms
- scalability: the power control algorithm (for cell size modification) works independently in small areas including some cells. Then the neighbouring cells are affected. Therefore increasing size of the network means only growth in complexity that can easily be handled
- stability: the method ensures stable solution, however oscillation is possible but one can overcome the problem by selecting proper thresholds in cell size modifying decisions
- flexibility: assuming the cell size modification can be easily solved by the existing radio network elements.
- visibility: the visibility of the cell size modifications are important in radio network controller units as they have to control the base stations. The other elements of the network are not affected, these only give some input for the mechanism (e.g. the required traffic)

- simplicity: as mentioned before power control algorithms are well-known tools in a cellular network and only some simple extensions are needed for coordinated cell-size modifications
- interoperability: the cell size modification eventually a spatial load balancer in the access network of the cellular radio systems therefore it has to interoperate with the communicating elements of the radio network (base stations, mobiles) and their traffic generating algorithms, e.g. in case of base stations their controller unit
- security: basically this question is out of the scope of the cell diameter modifying techniques, however it can be an interesting part of the future work

3.1.6.2 Reclustering of (relay-extended) cells

Assumptions:

In order to ensure mobile wireless service for large number of subscribers the service provider has to divide the service area into smaller parts, cells. Within a given cell only a part of the available carrier frequencies can be used to avoid interference coming from the neighboring cells. The operator can reuse the given frequency in cells being far away from the mentioned cell. The larger the distance between the two cells the smaller the caused co-channel interference. This also means higher data rate is achieved when higher reuse factor is used and the distance between the cells is increased. A group of cells, in which all the available frequencies are used, is called cluster. Finally, one can state that larger cluster size enables higher bit rates. The reclustering is also resulted in changing load balance among cells in border regions of the network. Depending on the structure of the aggregation nodes the varying traffic can cause significant load deviations in several core network links.

Mechanism description:

The size of a cluster in a cellular network can not be arbitrary. The possible cluster size is given by $i^2+i*j+j^2$, where $i,j \ni N$. *Figure 6* shows an example for different cluster sizes.

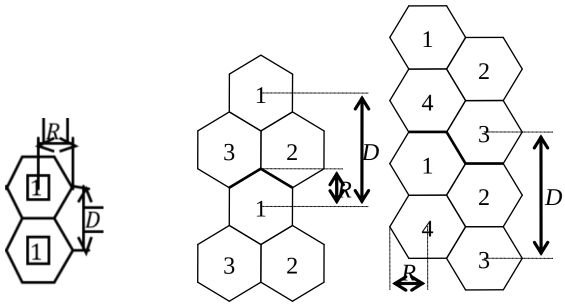


Figure 6: Scenario with different cluster size

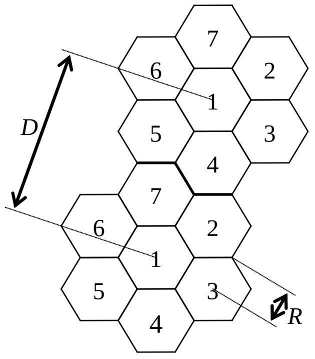


Figure 7: Aspect of sectorization

The cell radius (*R*) and the distance (*D*) of the cell using the same frequency are also represented. Together with the cellular division so called sectorization is also used by the operators. This technique means that not the whole cell is served by a transceiver but only a section of it, for example a cell consists of 3 sectors with angle of 120° like seen in *Figure 7*. This solution ensures lower co-channel interference and therefore higher data rates as well.

The mentioned cellular division and sectorization method has another useful extension, so called partial frequency reuse, which is represented in the *Figure 8*.

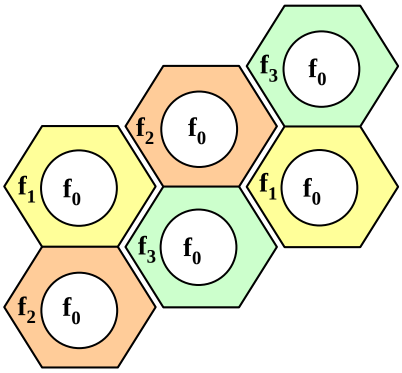


Figure 8: Partial frequency reuse

According to the basic idea, different power levels are used in the same cell for different frequencies. The carrier frequency f_0 is only used at low power and thus near the base stations where high data rates can be ensured. Because of the low radiated power at larger distances, e.g. at the cell edge any radio sign can not be detected and the interference is avoided. Therefore the operator can reuse f_0 in the neighbouring cell. The carriers f_1, f_2, f_3 with higher power can be used at the cell edges, therefore reuse factor N>1 is required. The coverage is larger for these frequencies, however the ensured bit rate is lower.

Problem statements:

In a cellular network the cell division, cluster forming, i.e. the frequency planning requires very thorough and deliberate work. The engineer has to consider the radio environment, the expectable subscriber demands, etc. Therefore reconfiguration of the network is rather difficult and time consuming. This means that fast or periodic (e.g. daily) reconfiguration of the network is not possible. The more realistic reconfiguration is the trend-like reclustering e.g. according to the network development.

When performing a reclustering solution in a given part of the network, the border zone between the two segments with different frequency reuse factor will always induce more problems to solve.

In case of cell reclustering, the following problems are recommended to examine and solve regarding the traffic engineering:

- examination of LTE clustering solutions, frequency band division, partial frequency reuse (LTE Advanced)
- possible frequency division solutions in adjacent cells, adjacent channel bandwidth division compared with total channel usage
- measurements based on signal power distribution in frequency domain for definition of the overall channel quality and reclustering solutions

Criteria used for description:

The usage of the mentioned solutions depends on several parameters related to the radio propagation environment and the subscriber's data rate demand. However, there are some auxiliary and derived parameters (like co-channel interference rejection, bit error rate) that have to be used in calculations. But during the design of the proper algorithm several other conditions and restrictions need to be considered, such as the mentioned cluster border problem.

General characteristics of the solution based on the aspects of RFC-3272:

- usability: this solution is only usable in large-scale variations the time, fast reclustering is not possible, and operators do not apply the fast reconfiguration

- automation: this is not the most relevant feature regarding the mentioned “slow” configuration possibility, and we expect that cell reclustering requires serious manual installation and configuration procedures
- scalability: scalability of reconfiguration in time is rather poor, i.e. only large-scale variation is possible without periodicity. Regarding the spatial scalability one can state the reclustering can be performed in any network independently from the network size, however, one has to consider that the required time is getting larger proportional to the size of the network
- stability: assuming proper frequency design stability problems usually will not arise in the network when reclustering is performed. The only critical zones are the borders of clusters with different frequency reuse factors.
- flexibility: we can state that the cell reclustering is rather an inflexible solution, only slow variations are possible.
- visibility: the visibility of the reclustering is important in radio network controller units as they can coordinate the operation of the radio traffic aggregating nodes, i.e. the base stations
- simplicity: reclustering algorithms are usually not complex, but the pre-design is rather difficult and requires serious human resource
- interoperability: the reclustering solution has to interoperate with the communicating elements of the radio network operator, e.g. base stations and their controller unit
- security: basically this question is out of the scope of the cell reclustering techniques, however it can be an interesting part of the future work

3.1.6.3 Coordinated beamforming

Assumptions:

Mobile subscribers have non-uniform spatial distribution in a given cell of a cellular network. Moreover, this statement is valid for the whole network. In some regions of the cell it is unnecessary to radiate power and cause interference for subscribers situated in the neighbouring cells, meanwhile it is worth to concentrate the antenna beams (lobes) on the wanted directions. Less interference means better radio channel quality that allows communication at higher data rates. The network load is expected to be significantly increased in the beamforming regions of the network. Meanwhile the traffic load remains lower in non-beamforming areas. Regarding the backhaul network this mechanism can significantly change the load balance of the core network elements. This effect is even significant when installing the beamforming elements in given parts of the network, meanwhile other areas are unchanged.

Mechanism description:

The coordinated beams should be provided with adaptive (or also called “smart”) antenna arrays. These are kinds of antenna types in cellular radio systems that confine energy in a narrow beam instead of broadcasting it over the whole cell. This radiation mode has many advantages. For instance the signal gain is increased depending on the antenna radiation pattern, i.e., on the antenna directivity. Range of the signal path is greater, spectral efficiency is improved, multipath reflection is reduced. Similarly to the last one, the level of interfering signals from adjacent cells or beams is also reduced. Due to these properties the network capacity increases, and this is the main effort in smart antenna systems.

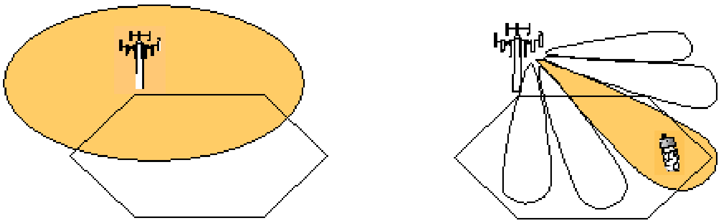
In the context of smart antennas, the term “antenna” has an extended meaning. It consists of a number of radiating elements, phase shifters, a combining/dividing network and a control unit. The antenna beam problem of coordination can be solved by different methods. Smart antennas can be grouped from this aspect.

The switched lobe or also called switched beam is the simplest technique. It comprises only a basic switching function between separate directive antennas or predefined beams of an antenna array. The setting that gives the best performance, usually in terms of received power, is chosen. Because of the higher directivity compared to a conventional antenna, some gain is achieved. Such an antenna will be easier to implement in existing cell structures than the more sophisticated adaptive arrays, but it provides only limited improvement.

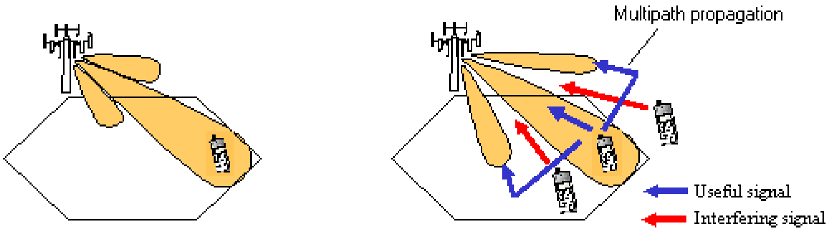
The next, improved and more difficult variation of smart antennas is the dynamically phased array. By including a direction of arrival algorithm for the signal received from the user, continuous tracking can be achieved and it can be viewed as a generalization of the switched lobe concept. The purpose is similar to the former one, the received power is maximized.

The third type is referred as adaptive arrays. In this case, a DoA (Direction of Arrival) algorithm to determine the direction toward interference sources (e.g., other users) is added. Then radiation pattern can be adjusted to null out the interferers. In addition, by using special algorithms and space diversity techniques, the radiation pattern can be adapted to receive multipath signals, which can be combined. These techniques will maximize the signal to interference ratio (or Signal to Interference and Noise ratio (SINR)).

The different types of antenna beams can be seen in *Figure 9*. The sector beam is provided by a simple sectorial transceiver, and in this situation no smart antennas are used. The first smart antenna applying system, which operates with switched beams, is also called multibeam system. In the figure it is shown in the second picture. The next more improved antenna concept is the steerable beam and it concerns to dynamically phased array, which is mentioned as the second one in former paragraph. Finally when zeros or minimums of antenna pattern are fitted to the directions of interfering signals, moreover, to the useful signals from the multipath propagation, the system is referred as adaptive array.



Sector beam **Multibeam (switched)**



Steerable beam **Beams of adaptive array**

Figure 9: Coordinated beamforming

Problem statements:

Apart from advantages they also have some drawbacks, of course. The most important of these is that beams have to be redirected continuously in order to track the angular position of mobile terminals situated in the cell. The angular position is usually called Direction of Arrival (DoA) in smart antenna systems. The measurement and mainlobe positioning require some additional devices in the transceiver system, which increases the complexity of the system.

Apart from the hardware and physical layer issues, the following problems are recommended to examine and solve regarding the traffic engineering:

- examination of the OFDM-MIMO transmissions on multi-cell and co-channel interference burdened environment (beamforming solution) and their effect on local or overall network capacity
- examination of various transmission data streams with different independence per antenna elements, space-time coding solutions (maybe without real beamforming but different feeding) to discover the possible benefit of different coding and evaluate for different assumed traffic mix
- elaboration of dynamic beamwidth control algorithm based on traffic demand with adaptive antennas (narrow beams in case of higher traffic, broader beams in case of lower traffic) based of previous traffic statistics
- examine the advantages of time-dependent periodic (e.g. daily) beamforming that can ensure for example higher rate communications in narrower beams at low traffic hours for more demanding users (in lack of other demands) due to interference avoidance

Criteria used for description:

The usage of the mentioned solutions depends on several parameters related to the radio propagation environment and the subscriber traffic demands in given segments of the radio access network. From the aspect of radio propagation basically the rural and city environment have to be distinguished. The propagation can be modelled differently in these areas, therefore the proper signal transmission requires different beamforming techniques. On the other hand, the subscriber density also differs in the mentioned regions, which is resulted in necessity of different beamforming solutions. Moreover, the spatial density distribution of the users differently varies in rural and urban areas. When planning a proper beamforming algorithm that can ensure suitable enhancement traffic and network capacity parameters all the mentioned criterions have to be considered.

General characteristics of the solution based on the aspects of RFC-3272:

- automation: this condition can be fulfilled by installing the adequate software in network nodes that are responsible for control of the RAN and ensuring the proper co-operation of these elements
- scalability: regarding that the propagation of radio waves is limited, assuming given maximal transmission power the beamforming solution will only have effect on a limited part of the network. This statement is also valid for small range radio networks because in their case low power transmissions are allowed to avoid interference. Therefore a well-defined upper limit can be determined for the maximum number of involved network elements and the complexity.
- stability: the stability of the beamforming techniques are based on the elaborated algorithms and can be ensured by proper mathematical considerations, for instance, to avoid harmful oscillations (unnecessary periodic deviations in the network beamforming state) etc.
- flexibility: assuming the now available radio network elements the network reconfiguration can be easily solved
- visibility: the visibility of the beamforming is important in radio network controller units as they can coordinate the operation of the radio traffic aggregating nodes, i.e. the base stations
- simplicity: this is a very important factor as complexity of the beamforming algorithms may increase easily that will result in slow operation
- interoperability: the beamforming eventually a spatial load balancer in the access network of the cellular radio systems therefore it has to interoperate with the communicating elements of the radio network (base stations, mobiles) and their traffic generating algorithms, e.g. in case of base stations their controller unit
- security: basically this question is out of the scope of the beamforming techniques, however it can be an interesting part of the future work

3.2 Microscopic traffic management

3.2.1 QoS differentiation based on applications and user profiles

The mechanisms described in this subsection likely will be investigated in more detail by the project consortium.

3.2.1.1 Application based differentiation

There are two possible scenarios for application based QoS differentiation [EKS+09]. One is based on network-initiated QoS control where the network initiates the signal to set up a dedicated bearer with a specific QoS toward the UE and the RAN. In this case, the user application does not need to be aware of the QoS model of the access network (QoS unaware client application). The second is based on terminal or UE-initiated QoS control where it is the terminal that initiates the signal to setup a dedicated bearer with specific QoS toward the network. The trigger for this signal is carried over a terminal vendor-specific QoS API. In this case, the client application must be aware of the access QoS model in order to specify the QoS information for the bearer (QoS aware client application). The following elements intervene in both scenarios: UE, LTE RAN, transport network, PDN gateway, policy controller (PCC, Policy and Charging Control, or PCRF, Policy and Charging Rules Function), CSCF application function (IMS Call State Control Function). In the UE-initiated control, the Serving gateway is also involved. The difference between the two is that in the network-initiated control the PCRF pushes the rules to the PDN GW, whereas in the UE-initiated control the PCRF waits for the UE to trigger a pull of rules from it.

Assumptions:

The subscriber is engaging in two services, e.g., Internet browsing and peer-to-peer file sharing. These services are both mapped onto the default bearer. The IMS application in the client is preconfigured with the IP address of the CSCF so that signalling messages are directed toward this node.

Problem statements:

Initiating a dedicated bearer can be done by the UE or by the network. In the case of network initiated QoS control, one needs an Application Function (e.g. IMS CSFS) or a DPI function and the signal is carried over Rx or Gx interfaces. In this way the application does not need to be aware of any QoS issues. However the application can negotiate QoS requirements (network agnostic) using SIP or RTSP protocols. On the other hand, UE initiated QoS control makes it necessary that the application knows about the QoS model used in the access network (using a vendor specific QoS API) and there is no policy controller involved. For the sake of flexibility, both of these techniques are necessary, particularly because some services are to be controlled by the operator but others not and some are QoS access agnostic while others need to control their QoS.

The dedicated bearer is associated to a QoS class (QCI), allocation/retention priority (ARP), bit rates (GBR and MBR), total bit rate allowed for non-GBR bearers (AMBR). This association is done by the PCC providing operators with service-aware QoS and charging. It is not generally possible to overprovision for radio networks, making it necessary to ensure efficient utilization of resources. Likewise, services and users have very different needs that must co-exist on the same network. The PCC not only allows controlling traffic aggregation over a bearer with a given QoS characteristic but also allows finer-grained control by acting at the session level.

Criteria used for definition

The choice of the QoS control technique depends on the user service requirements and on the type of service. A gap could exist between the network QoS provision and the QoS required by some applications. The dilemma is how to maximize network usage while providing the adequate QoE. Network optimisation may imply ignoring some service requirements but this forcibly has impact on the perceived user service experience. Over-dimensioning to achieve QoE implies inefficient use of the network resources. A possible solution lies in enabling several different techniques that are complementary (e.g. over-provisioning, fixed provisioning, terminal initiated QoS, network initiated QoS, etc.) to be able to ensure that different possibilities can be offered to the users that can decide according to their needs and the price they are willing to pay.

Typically, applications that might need to maintain control of the QoS are real-time and safety-critical applications, business/industrial support systems (SCADA), dedicated enterprise networks, multicast and broadcast services where it is difficult to individually select optimum interfaces for the large number of users, etc. Other factors also have impact, such as the use of strong encryption that makes DPI techniques and Lawful Interception more complicated or even impossible.

On the other hand, services where the access network operator wishes to maintain control are Internet access, mobile TV, IMS services (e.g. voice), network server supplied FTP and Web Services, etc.

3.2.1.2 User based differentiation

The network operator provides different subscriber classes (2-3; e.g. premium, business, standard) for differentiated service experience. According to the contract, the premium class subscriber gets highest priority for the downstream traffic and possibly for the upstream traffic too. In case there is no congestion in the network (i.e. there is sufficient bandwidth available) there is no need for differentiation between traffic associated with different subscriber classes from a solely technical point of view. A further specialization of this scenario is the combination with application differentiation. In this context a traffic for premium user may get highest priority only for certain types of applications and services, e.g. certain delay sensitive and interactive applications. All other applications from the premium user would be handled with the same priority like traffic associated with users from other subscription classes.

In order to enhance flexibility of the QoS mechanism, the concept is to apply traffic control based on application differentiation together with user differentiation. For example video conferencing may get higher priority compared to multimedia file upload or a user associated with a certain user class may get higher priority over users associated with a different class.

This scenario should also consider the roaming case, where service experience in the visited network might be different for the user compared with experience in the home network. This may apply to the user connection or just to a single or group of applications. Ideally the user is informed that there is a mismatch what can be expected in the home network and in the visited network.

An example policy for the combined usage of application based and user based differentiation mechanisms is given in the following, if the UMTS QoS model is used with the four standard QoS classes ‘conversational’, ‘streaming’, ‘interactive’ and ‘background’ [3GPP_23.107]:

- conversational and streaming traffic of user A gets higher priority over any traffic of user B
- conversational and streaming traffic of user B gets higher priority over interactive and background traffic of user A
- interactive traffic of user A gets higher priority over interactive and background traffic of user B
- there is no differentiation for background traffic and interactive traffic of user B

Details of algorithms for traffic flow distinction are subjects of further study. It is also a challenge to ensure that the number of combinations between user and application prioritization can be mapped to the maximum possible number of QoS classes. In this sense prioritization should not be limited to forwarding behaviour but may consider access control as well. Details are as well for further study.

3.2.2 Improvements on TCP layer

The mechanisms described in this subsection likely will be investigated in more detail by the project consortium.

3.2.2.1 Cross Layer Interference Detection

Interference between adjacent cells is one of the challenging problems in wireless communication. Interference will corrupt packet flows resulting in traffic overhead in the wireless and as well in the wired network. Especially a combination of ARQ based link layer protocols and TCP will suffer from packet losses caused by interference. The link layer tries to hide losses to TCP, not taking into account TCPs retransmission behavior. Every unnecessary TCP retransmission will add unwanted traffic to the network and will reduce application throughput and response times. Similar examples can be found for real time traffic.

Traditional interference situations are detected by mobile end-systems by means of signal strength indicators. End-systems will signal interference situations to the access point the end-system is associated to. The access points can normally not recognize interference, due to the fact that access points will in most cases not interfere with other access points. As commonly known, signal strength indicators depend on the sensitivity of the hardware equipment of the end systems.

In this scenario access points or network devices in the core network, which are located near to the wireless border, will be made capable to identify interference by means of cross layer traffic monitoring.

Such devices will stochastically probe specific TCP flows and correlate them by itself and with parameters of lower layer protocols with the aim to identify characteristics of interference. This method opens the possibility to dynamically reconfigure radio cells to mitigate disturbing interference.

3.2.2.2 Support of multipath flows

In future mobile networks, it can be assumed that the capacity of wireless access network is higher compared to the capacity of the wired mobile network. The increasing number of high bandwidth radio cells will push for the modernisation of the wired mobile network to meet the increasing bandwidth demand. Especially, numerous WLAN and femto cells might make the bandwidth demand unpredictable for the mobile network. Therefore the network is expected to become the bottleneck of the end-to-end path. A common method to support high bandwidth applications is to split a single flow into multiple flows and carry each flow over uncongested regions of the core network. This splitting mechanism can be also used to distribute multiple flows over overlapping radio cells of different radio access networks. This means, splitting must be done at the end-system. If the flow is carried over a single radio technology, splitting functions can also be placed within the core network. Also a combination of both, end-system and core network splitting support is possible.

A lot of proposals to split TCP connections over multiple flows have been devised. In general, these solutions are difficult to implement or the resulting performance is lower than expected, especially for TCP bulk applications. The main problem is caused by the fact that the design of TCP is not able to cope with the characteristics of multiple flows, meaning extremely high round trip time variations and increase in per flow losses resulting in packet reordering and TCP retransmissions

3.2.3 Ad hoc change of QoE

Subscription profile of the user may allow changing QoE for certain types of applications. Hence there is a default QoE for some applications associated with the user profile and a certain type of UE. There are three purposes to change QoE for an application or set of applications during runtime.

- user requests more bandwidth for an application: This can be applied e.g. for a file transfer application to speed up the download or to change resolution for video streaming from standard to HD format.
- user changes access and / or device: This situation is detected again by the network and the bandwidth is changed to fit with the changed environment.

The mechanisms described in this subsection likely will not be investigated in more detail by the project consortium.

3.2.4 Dynamic routing of EPS bearers

In order to transport user data from the mobile devices through the radio and core network to the Internet, the EPS uses the concept of bearers. Once a PDN-connection to an APN, i.e. to a PDN-GW chosen from a subset of available gateways, is opened, a default bearer to that gateway is established. The chosen PDN gateway cannot be altered during the lifetime of the connection.

In contrast to the single radio path from a UE to the eNodeB, usually there are multiple entities of core network elements and possibly also several network paths between them. While some form of internal load balancing is involved when initially selecting the gateways, this by no means ensures an optimal user and core network performance or load distribution over the course of a connection lifetime. In order to deal with this limitation several schemes were proposed solving only parts of the overall problem space by incorporating different methods for dynamic re-allocation of bearers (e.g. [Xue01], [Shiao01]). However, 3GPP still lacks in a general well established and standardized solution.

In a typical mobile Internet connection, one default bearer (or PDP context in a GPRS/UMTS environment) is established, and the traffic of the whole session, consisting of any number of flows, is routed through it. After that, up to ten additional dedicated bearers can be established up to at total of eleven bearers (cf [3GPP_24.007, Section 11.2.3.1.5],. Traffic flows are mapped to these bearers by Traffic Flow Templates set up in the UE and P-GW with several packet filter attributes (e.g. remote address or local/remote port) to choose from. The always-enabled default bearer serves as a catch-all tunnel for all flows which are not yet bound to specific filters.

Today's Internet is not the classical, strongly tiered network of networks any more that it once was. It now has many shortcuts from local ISPs to content providers and their CDNs to improve the quality of experience for the users. So, one mobile gateway may have access to a better route to the target while another gateway in the same network may not. If the gateway that was chosen for them is not “shorted” for the remote flow endpoints they are interacting with this could result in a degraded performance for the mobile users. The mobile user could experience degraded performance, if the selected gateway can’t establish connectivity with sufficient QoE to the remote endpoint.

One can envision several ways to improve on this situation and dynamically start to choose better routes for flows through the core network:

- Allow and encourage the establishment of multiple default bearers per UE and map the flows into them. This would further differentiate the needs of a mobile user's application profile.
- Enable bearers, targeting one APN (this could be the Internet access point), to terminate at different P-GWs to allow for an improved routing from and towards specific Internet hops, especially in light of the mentioned shortcuts.
- Improve the mapping of traffic flows onto bearers and make it more dynamic to increase the responsiveness and granularity of the core network flows.

Making the UE visible at more than one gateway router to the Internet might have several issues. If a device is reachable through more than one gateway this could pose significant problems to the routing from the remote node to the UE. And if another access provider internal (transport) network follows directly behind the mobile gateways with only a single connection point to the Internet, the core network improvements would probably have less effect.

Also the core network itself may come in a multitude of variants, with some less suited to the mentioned alterations.

All in all, extensive studies are needed to show if there can be measurable effects from this class of core network routing improvements.

The mechanisms described in this subsection likely will not be investigated in more detail by the project consortium.

3.2.5 Upstream flow control

Up to the 90s, upstream traffic was not an issue at all, because the volume compared to downstream communication could be neglected. For TCP there was only low background traffic related to ACK message returned to the sender from time to time. Other applications are electronic mail (based on SMTP), text based chat and messaging and network diagnostic tools. But recently there is a trend of increasing data volume for upstream traffic injected by the UE. This can be especially a problem for the radio access, because it is not dimensioned for the same high data rates like in downstream direction. It should be assumed that a bottleneck could also be within the mobile backhaul, e.g. for microwave links. As a consequence prioritization for upstream flows might be necessary for QoE sensitive applications with significant volume injected by the user terminal. Applications which may fall into this category are

- multimedia upload e.g. via social networks (video and sound clips, photos, images, etc.)
- video conferencing (smart phones and notebooks with built-in cameras)
- user hosted content, e.g. within P2P networks or open telco platforms

QoE differentiation might be primarily important for video upload.

The mechanisms described in this subsection likely will not be investigated in more detail by the project consortium.

3.2.6 Selective admission control

The fulfilment of the paradigm of an admitted user being a satisfied user, and more precisely quality of service (QoS) support in terms of low packet loss and delay at a guaranteed throughput, is important in context of an increased traffic volume in carrier networks nowadays, and becomes even more critical for future Internet with coexistence of data, voice and high-quality video (e.g., IP-TV services) traffic. For coexistence of such traffic mix simple capacity over provisioning is not sufficient, and thus motivates the need for appropriate admission control schemes. Additionally, the admission control algorithms are required to be efficient, robust and scalable.

To this end, a new admission control mechanism will be proposed. In order to prevent situations in which network cannot grant connectivity for all users at any time, selective admission control algorithm will provide the network with the means to deliberately reject incoming users,. e.g., by not responding to the SYN packet of an incoming TCP connection. Evaluation of the proposed admission control scheme from the perspective of the individual traffic flows is subject of further studies.

3.3 Improved resource selection & caching

All mechanisms described for in the following subsection likely will be further investigated in more detail by the project consortium.

3.3.1 Caching of popular Internet content

In this case content is transferred from a node in the Internet to a cache node within the operator domain. How to manage re-direction of connection requests to a new location is described in section 3.3.3. At first place it is necessary to detect that caching of popular Internet content is beneficial. The following criteria could be used:

- the number of accesses to specific content within a certain time period
- a trigger sent by a CDN / content provider (based on SLA with CSP)
- volume of the content
- copyright status of content
- optimization potential (e.g. in terms of QoS parameters such as throughput / delay / number of hops etc.)
- value of content for the user
    - e.g. price charged to the user

- user profile: caching of user content may be restricted to a certain subscription profile. Also content caching may be suppressed for users attached to a visited network.

Caching might be used in principle for all Internet based non-streaming content, i.e. data which is accessed as a file, i.e., P2P files, multimedia files, etc.

In addition there is the challenge to keep content up-to-date. This needs synchronization between the cache engine and the application.

The content may be replicated possibly among different cache nodes within the operator domain.

What should be assumed for a roaming user, e.g. if the home network does not support caching, but the roaming network does.

3.3.2 Caching of user hosted content

There are applications which send data in a repetitive way into the network. Optimization can be reached, if the hosted content is stored in the network.

A cache may be located at the base station or at some other location further upstream. Caching at base stations could be beneficial, if there is congestion in the mobile backhaul. Caching makes sense for data which is accessed repetitively, e.g. P2P applications, or for content hosting.

Applications, which may fall in this category and which may benefit from user hosted caching, are:

:

- P2P applications
- content hosting on UE
    - UE is part of a P2P overlay
    - UE hosts web server within the public Internet
    - UE provides content within open telco platform or as part of a CDN

Devices, which generate traffic, might be mobiles, notebooks or stationary computers connected to cellular access (replacement for other access technology).

Caching might be explicitly triggered from the UE, e.g. by an application or by the user.

In a different situation the network may recognize that there is benefit for storage of UE hosted content. An entity within the network may trigger the usage of network located cache service via an application or send an announcement to the user.

Further challenges are:

- Identification of a suitable location for caching
- Level of redundancy – how many replicas of the content are stored concurrently within the local domain
- Selection of suitable locations for content
- Hosted content might have to be updated from time to time

3.3.3 Redirect requesting node to alternative resource

As network resources in that context are both limited but growingly solicited, it becomes a necessity for both users and network providers to carefully select the content locations with which to deal content, be it in a P2P, CDN, or any application involving exchange of payload. P2P and CDN based data transfers involve a majority of terminals located in wireless networks, where maintaining the QoE is even more challenging. The IETF Application Layer Traffic Optimization (ALTO) working group defines protocols to support TE solutions that provide guidance on where to get or download the content from given the transport network topology viewed from the transport network operator. In *Figure 10* the resource selection mechanism is displayed for a P4P service, which provides information to a P2P application like investigated within IETF WG ALTO. The figure shows critical situations at some locations in the network. One example is congestion in upstream direction caused by a P2P server, which is frequently requested for content. This becomes a problem especially for mobile peers, since capacity in upstream direction is less compared to downstream. The second example displayed addresses frequent downloading of the same content from external sources. This causes significant extra cost to operators which have volume based service level agreements with the IP backbone provider. In both situations a P4P server may identify an alternative resource, which is more suitable from operator perspective.

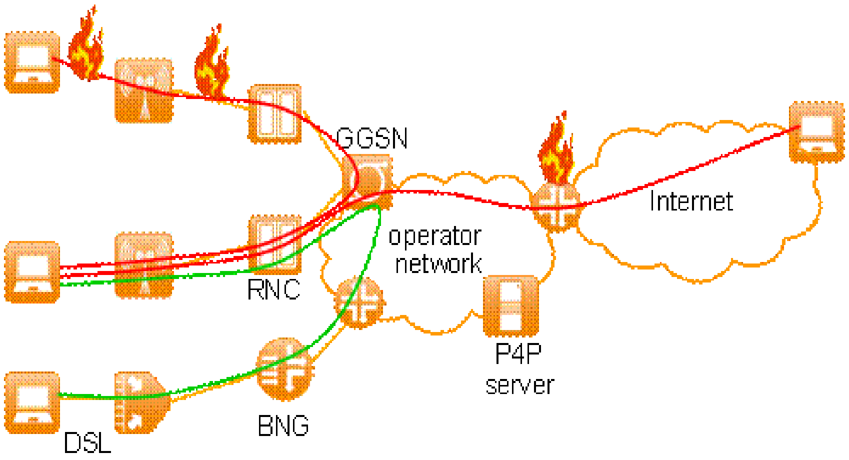


Figure 10: Resource information service

However, the ALTO work is tailored for fixed provider core networks. So it is necessary to define particular ALTO services and related protocol extensions supporting traffic optimization on mobile core networks such as the LTE EPC.

- One way is to provide solutions based on cross layer cooperation between the application and the network layer.
- It is crucial to gear the evolution of ALTO related functions to mobile networks and define the implementation architecture for ALTO functions in mobile networks.
- A requirement for that is the presence of features in the access and EPC to support ALTO based functions.

It should be distinguished between non-transparent re-direction to an alternative resource and transparent re-direction w.r.t. the user or application client (AC). The redirection is called “transparent” when the user or client application is not involved in the re-direction to another location. Examples are HTTP server re-direct or manipulation of IP address mapping functions, such as DNS, NAT and others. On the other hand, when the user or application client exchanges information with a serving entity in order to obtain knowledge about preferred candidate locations, the operation is qualified as non-transparent. Transparent re-direction is already described e.g. by a mechanism like ALTO, where the AC, linked to an ALTO Client, explicitly sollicits guidance from the ALTO server to select an appropriate location to get content from

3.3.4 Detect unfavourable resource usage

Unfavourable resource usage means that a user is connected to a resource, which is unfavourable from operator perspective or which can be accessed with limited QoE for the user. There are several situations why connection to a current resource is not preferred:

- a user changed access and the path to the connected resource is not preferable any more
- a better resource has become available after some time
- characteristics of the path have changed (e.g. due to congestion)
- a user / P2P application initially selects a sub-optimal peer

In some cases where re-direction could not be handled transparently the user needs to be informed via some notification mechanism.

3.4 Application supported traffic management

This subsection describes the mechanisms on application layer to improve performance for the end user. Functions related to this building block will not be investigated in detail by the project consortium.

3.4.1 P2P optimization techniques

3.4.1.1 Introduction to P2P networks

A peer-to-peer, or “P2P,” file transfer service allows the user to share computer files through the Internet. These services are set up to allow users to search for and download files to their computers, and to enable users to make files available for others to download from their computers. Since their inception in 1999 with the Napster file-sharing service, peer-to-peer networks have grown to become a predominant source of Internet traffic [Basher08]. Even though the relative volume share of P2P traffic has declined in the recent years, it is still significant within operator networks. P2P is based on unicast, the natural way of transmitting information on the Internet. Peers transfer data to one another in an end-to-end fashion. What makes P2P so powerful is its “exploding” capability. It uses peers as exploders, making each peer simultaneously receive and send data to several peers. This scheme provides scalability in that it decreases the amount of resources needed for a producer to distribute its contents. In other words, peer-to-peer networks facilitate information dissemination by spreading the distribution load in contrast to a centralized server that would have to bear over a swarm of hosts, thereby reducing infrastructure costs. Thus, a much larger audience can be addressed simultaneously at lower costs.

There are different levels of peer-to-peer networking such as Hybrid P2P, Pure P2P and Mixed P2P. For Hybrid P2P, there is a central server which keeps information about the P2P network and the peers are responsible for storing the information. If they want to contact another peer, they query the server for the address. There is absolutely no central server or router for Pure P2P and each peer acts as client and server at the same time. Pure P2P is also sometimes referred to as “serverless” P2P. Last, Mixed P2P combines “hybrid” and “pure” P2P networking. An example is Gnutella™ which has no central server but clusters its nodes around so-called “supernodes”.

A phenomenon like P2P that involved some years ago over 66% of all the Internet traffic is vital to an information society and needs to be investigated to unlock further potential [P2PNext]. New technologies in multimedia distribution such as on-demand content distribution, Internet TV and VoIP services require the switch to P2P technology at lower costs. As mentioned before, in a P2P system, peers communicate directly with each other for the sharing and exchange of data as well as other resources such as storage and CPU capacity, each peer acts both as a client who consumes resources from other peers, and also as a server who provides service for others. Compared with traditional streaming techniques such as IP multicast and CDN (Content Delivery Networks), a P2P based streaming system has the advantages of requiring no dedicated infrastructure and being able to self-scale as the resources of the network increase with the number of users. However, switching to these systems is not easy since there are some challenges to P2P systems.

Peer-to-peer (P2P) file-sharing systems generate an estimated 50 to 80% of the total traffic depending on geographic location [Ipoque]. This file sharing method makes possible to distribute data to a large population of users without the need for big investments of the provider in server parks. The costs of the information sharing are carried by the end users and the Internet Service Providers (ISPs).

In P2P systems file contents are divided into a large number of pieces. Peers exchange these pieces among each other. Peers that do not have the entire content are called “leechers” while peers owning the entire content are called “seeds”. The set of peers involved in the distribution of the same content is called a swarm. Peers can find other peers interested in the same content via a centralized tracker, a distributed hash table or unstructured overlay. Neighbors of a peer are a subset of the peers in the swarm. A peer exchanges data with a subset of its neighbors. This later subset is dynamically determined by the "choking “ algorithm in BitTorrent™, or might be static as in Gnutella ™.

3.4.1.2 Problems and motivations

One problem with many P2P solutions is the proximity-unawareness. P2P file sharing protocols were not originally designed with the network topology in mind, causing non-optimized volume of inter-AS (Autonomous System) traffic. One research area of P2P systems is to reduce transit traffic in the network caused by P2P applications. Solutions introduce locality awareness into P2P protocols, or add supplemental localization services such as P4P [P4P].

On the other hand, any locality aware solution favours the apparition of partitions, i.e., if during a period of time the leecher cannot find a piece it needs among its neighbours, it connects to the tracker with a

"Partition Merging" flag. It means that the leecher believes there is a partition and that it needs a connection to a new peer outside its home ISP. In case of large torrents, there might be an implosion of requests at the tracker, known as the feedback implosion problem. This problem has been extensively studied and can be solved using several techniques [Nonnenmacher99].

Another way to reduce inter-ISP traffic is the introduction of caching in ISPs so that popular contents do not have to be downloaded from remote peers, i.e., reduce incoming transit traffic. There exist several types of caches. Transparent caches intercept requests to external peers from the ISP, and transparently upload the pieces to the peer. Another category of caches are the ISP Managed "Ultrapeers" such as OverSI's OverCache P2P [Oversi], which appear as regular high upload capacity peers to the local peers. The neighbour selection algorithm decides how often a peer downloads from the cache peer. A third category of caches are the ISP Managed Caches. Via some cache discovery protocol such as BitTorrent, or approaches defined in the IETF SIPTO, DECADE, working groups, the peers get aware of the local caches, and can select downloading from the cache.

Frank Lehrieder et al. [Lehrieder10] have recently analyzed the impact of caches on inter-ISP traffic. The results show that inter-ISP traffic is not proportionally decreasing by increasing the cache size and local upload rate from the cache to the local peers. Even if caches serve the local peers in the ISP, those local peers will seed the entire P2P system. The identified benefit of Local caches is to add additional upload capacity to the P2P system. The reduction of inter-ISP traffic depends on several other factors and in some cases, caches even increase the overall inter-ISP traffic.

3.4.1.3 State-of-the-art solutions to reduce inter-AS traffic

Introducing proximity-awareness in P2P protocols is one way to reduce considerably inter-ISP traffic.

Stevens Le Blond et al. [LeBlond10] introduce Tracker-based peer selection mechanisms that enable high traffic locality. They show that they save up to several orders of magnitude of inter-ISP traffic compared to random peer selection, without adversely impacting peers download completion time. In an experiment they show that whereas the torrents they crawled generated 11.6 petabytes of inter-ISP traffic, their locality policy implemented for all torrents could have reduced the global inter-ISP traffic by up to 40%.

D. R. Choffnes, F. E. Bustamante [Choffnes08] propose an alternative, scalable technique to provide biased peer selection that requires no cooperation or trust between ISPs and their subscribers, no additional infrastructure and no network topology information. Their technique is based on the observation that the information necessary for peer selection is already being collected by Content Distribution Networks (CDNs). CDNs use dynamic DNS redirection to send clients to low-latency replica servers located in thousands of ISPs worldwide. CDN-based "hints" can inform a biased peer selection algorithm that significantly reduces cross-ISP traffic.

R. Bindal et al. [Bindal06] propose also a biased peer selection algorithm which maintains the nearly optimal performance of BitTorrent in a variety of environments, and fundamentally reduces the cross-ISP traffic by stopping it from growing linearly with the number of peers. A key reason for its performance is the "rarest first piece replication algorithm". Compared with existing locality enhancing approaches such as bandwidth limiting, gateway peers, and caching, biased neighbour selection does not require dedicated servers, and scales to a large number of BitTorrent networks. Furthermore, it can be combined with bandwidth limiting and caching to improve their performance further.

T. Koulouris, R. Henjes, K. Tutschku [Kouloris03] propose the use of "Active_Virtual Peers" (AVP). An AVP is a virtual entity which interacts with other peers inside a P2P network. An AVP is a representative of a community of peers. Its purpose is to enhance, control and make the P2P relation more efficient inside that community. AVPs enable flexibility and adaptivity by the use of self organization.

The AVP performs certain functions, not expected from an ordinary peer. The proposed approach based on AVPs includes for example a *dynamic* forming and maintaining of peer-to-peer overlays, and an adaptive routing of signaling and download traffic. They claim that the approach is also extensible to enforce security or ownership rights and to include mechanisms of charging.

The authors believe there exist four areas where the enforcement of control will be beneficiary for P2P applications:

- The first is *access control*. Participants of P2P overlays are typically granted access to all resources offered by the peers. These resources are valuable.
- The second area is *resource management*. The resources of individual peers have to be treated with care, e.g. low-bandwidth connected peers should not be overloaded with download requests and exploited equally.
- A third area of interest is *overlay load control*. Overlay load control copes with traffic flows inside the overlay. The majority of P2P systems operate by creating overlays on top of the

application layer, where peers form their application-level virtual topologies. These overlay topologies rarely match the underlying network infrastructure leading to inefficient operation and large operational costs for the network and Internet Service Providers (ISPs).

- The forth area of command is adaptive topology control.

By inserting an AVP inside the P2P network an ISP can manage aspects of the service's behaviour in a manner that is transparent to the rest of the peers while having an "insider's" view of it at the application level.

In spite of the fact that security is not the main focus of interest in the MEVICO project, it would be beneficial to follow the results of Framework 6 European research project PEPERS (Mobile Peer-to-Peer Security Infrastructure). PEPERS focuses on mobile networks and most of previously mentioned work does not consider mobile P2P. The other reason is that building secure networks would increase the efficiency, hence helping the optimization goal.

One such work has been performed by G. Spanoudakis, C. Kloukinas and K. Androutsopoulos [Spanoudakis08] who propose a method for mobile P2P systems where dependability and security can be enhanced through the runtime monitoring (a.k.a. dynamic verification) of the compliance of the system behaviour to specific dependability and security properties, and the execution of control in cases where properties are violated. They use peer-specific monitoring policies to specify application-level properties.

P2P communication has a great deal of attention in today's technology and centralized network structures are also being transformed to distributed peer-to-peer networks with the idea of distributing storage and computational capacity. Considering mobile communication, this transformation of course needs special requirements such as new P2P schemes that have been defined in many ways by various researchers. In these studies, P2P schemes for mobile communication seem to include new applications based on physical location and context, different from traditional sharing applications and considering limitations in mobile communications. Although current P2P applications and architectures are mainly designed for wired networks, recently developed mobile communication technologies and devices such as LTE can allow to overcome the listed challenges.

3.4.1.4 Device, Network and Wireless Environment Challenges

In *Figure 11*, device and wireless environment challenges are illustrated. User mobility can cause a "challenge", enabling roaming between different topologies. In this sense, especially, a network that provides the lowest data rate, can dominate end-to-end performance. Some standards are developed to solve this problem, namely GAN (General Access Network) and MIH (Media Independent Handover). Beside this, in the next generation mobile communication (4G), there will be a large increase of mobile devices usage. Unfortunately, today's devices have a poor battery capacity and some applications such as P2P can require significant battery and processing power. This trend must be mainly considered for the 4G. But, in the future, the contents may be stored in mobile devices with high memory capacity, instead of storing in centralized network. This enables us to propose point-to-point and meshed network communication as a solution for challenges in P2P (see *Figure 12*).

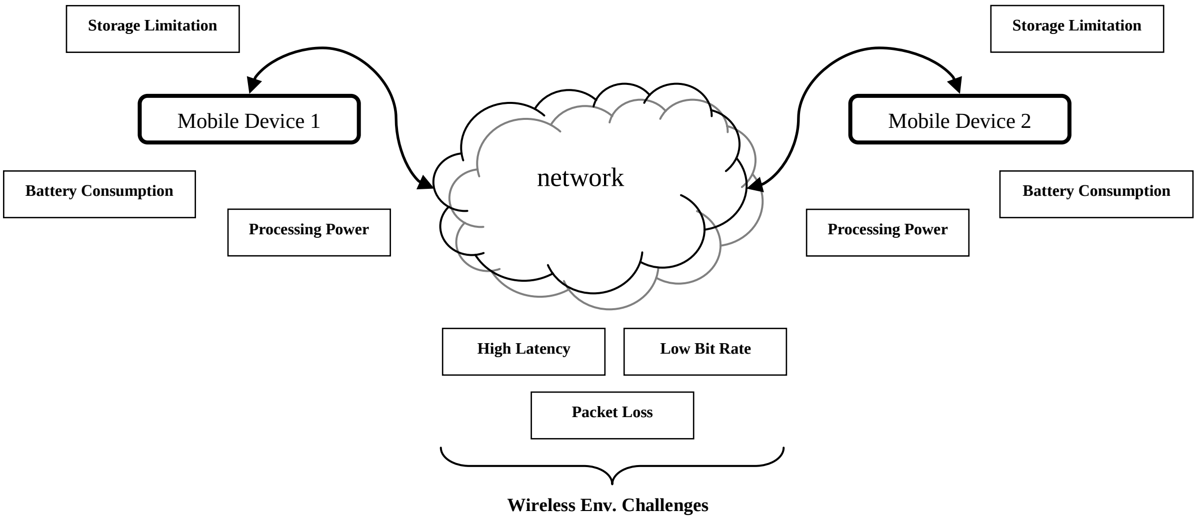


Figure 11: Device, Network and Wireless Environment Challenges

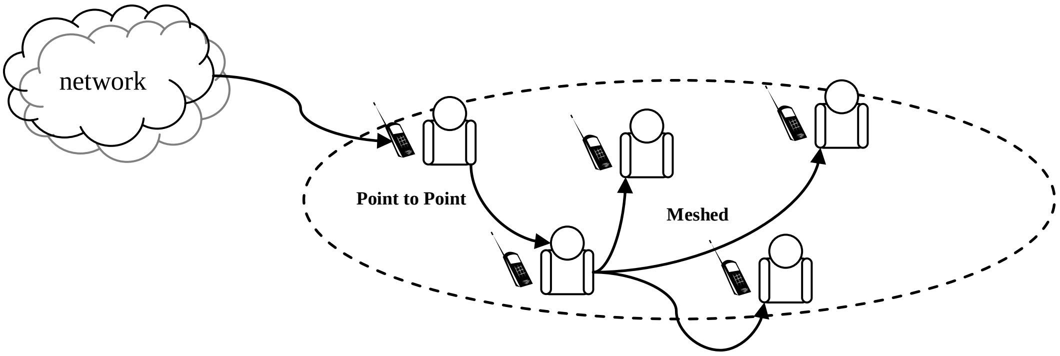


Figure 12: Point-to-point and meshed network communication

In our opinion, mobile user demands in the future mostly include multimedia (music/video) files generated by P2P applications over mobile networks and this user behaviour can be a huge threat to mobile communication networks. The new proposed P2P may combat these challenges such as limitations in transmitting huge multimedia files over wireless networks, supporting cooperative or relay based communication that creates virtual links between the users and source of content. Cooperative/Relay based communication can be probably preferred by a service provider due to providing higher throughput without overcharging in existing traffic load

One of the main challenges in P2P distribution is the lack of topological-awareness [P2PNext]. Flows of data are not topology aware in P2P networks. Currently, peers connect to randomly selected peers. Therefore, the flow (the P2P overlay) is completely independent from the network topology. This means that a single piece of data is likely to be transmitted through the same network link several times which brings higher loads and degrades the performance of the network.

Several approaches are being developed to provide P2P topology awareness. P4P, for instance, uses an oracle service provided by the Internet Service Providers (ISPs) which hints peers with information about network topology [P4P]. Unlike P2P, which selects random peers to share with, when using P4P the peers are intelligently selected as the protocol utilizes network topology data to maximize the efficiency of routing between the peer-to-peer connections as seen in *Figure 13*.

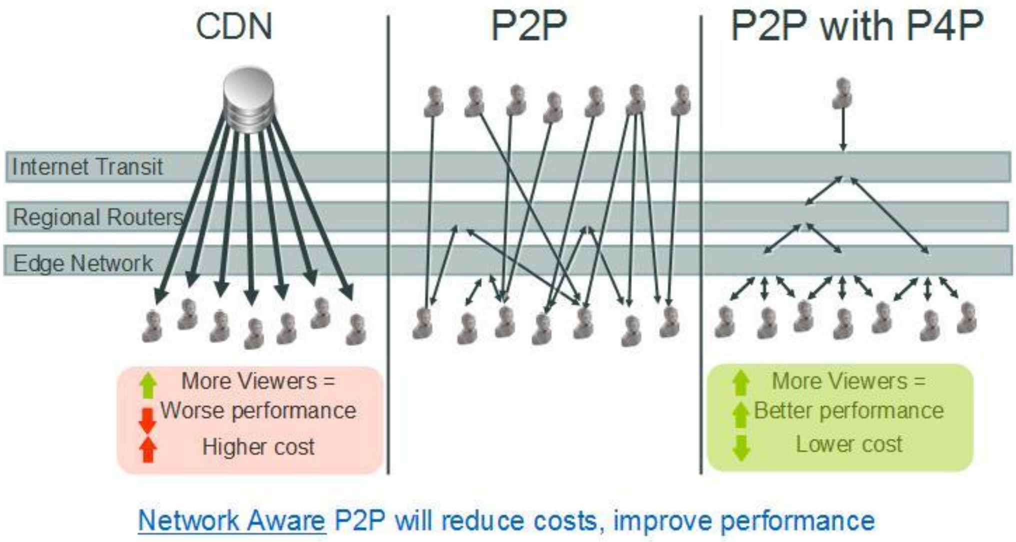


Figure 13: Data distribution in CDN, P2P and P4P architecture [P4Pslide]

In order to improve the various performance aspects of the P2P networks underlay awareness is also an important topic. Underlay awareness is defined as the collection and usage of underlay parameters as ISP Location, Latency, Geolocation, and Peer Resources which affects the overlay [Abboud09]. The underlay abstracts the physical, Medium Access Control (MAC), network and transport layers.

ISP-location is the identification of the ISP through which a peer connects to the Internet. ISP-location awareness in P2P systems will result in more locality of traffic and this information is usually used to improve the neighborhood selection of the peers in such a way and so as to reduce inter-AS (Autonomous System) traffic. Latency-aware systems aim at reducing the communication delays between peers by choosing shortest paths. Awareness of P2P applications on latency is a necessity for next generation IP applications ranging from multimedia content to live communication services. Peer resources, including available bandwidth, processing power, harddisk space, etc., are defined as the set of parameters that estimate different capabilities of a peer. Peer resources information is used to construct P2P systems with shorter search and download times, higher responsivity and stability. This is achieved by arranging the overlay in such a way that different roles in the network are taken by appropriate nodes [Graffi08]. Geolocation is the identification of the geographic location of a peer and can also be used to calculate the geographical distances between two different nodes. Geolocation information is used to build an overlay where neighboring peers are geographically close. Geolocation awareness can be applied to construct the P2P network in such a way that peers, which are geographically close, are also closely located in the routing overlay network. One parallel advantage is that this kind of routing may be more scalable since with increasing load, congestions are less probable due to locality of traffic.

At the same time, also the networking capabilities of mobile devices are constantly evolving. Their usage and the number of users are increasing. This has raised the question about adapting the P2P model into the mobile environment too. Despite the constant evolution of the mobile devices, the mobile environment is still very limited when compared to the fixed Internet. Mobile devices have limited networking capabilities, storage, electrical power and processing power. In addition to this, the mobile environment architectures set limitations on mobile applications. In a mobile P2P network, also a fast-changing topology of the physical network may be a significant problem from P2P point of view. As the underlying physical network keeps changing, it is hard for an overlay P2P network to maintain an optimal or reasonable topology. Thus, adaptive P2P systems should be designed to adapt rapidly to network resources and topology changes for these mobile networks. The usage areas (real-time collaboration, instant MP3 sharing, etc.) and other challenges (lack of effective addressing mechanisms, data sharing and synchronization, security and etc.) of mobile P2P systems are explained in detail in [MobileP2P]. For the mobile P2P systems, underlay mechanisms that support low footprint mobile devices should be incorporated. By knowing these challenges, new methodologies can be developed in order to reduce the traffic load and increase the performance of the mobile P2P networks.

3.4.2 CDN optimization techniques

A CDN replicates the content from the place of origin to the replica servers scattered over the Internet and serves a request from a replica server close to where the request originates. A client accesses a copy of the data close to the client, as opposed to all clients accessing the same central server, so as to avoid bottlenecks near that server.

Content types include web objects, downloadable objects (media files, software, and documents), applications, real time media streams, and other components of Internet delivery (DNS, routes, and database queries).

CDNs generally deliver content over TCP and UDP connections. TCP throughput over a network is impacted by both latency and packet loss. In order to reduce both of these parameters, CDNs traditionally place servers as close to the edge networks that users are on as possible. Even though traffic is carried over TCP connections, IPTV CDNs such as Shanghai telecom IPTV CDN carries 90% of the traffic over UDP

There is a group of free CDNs, mostly based on P2P and a group of commercial CDNs among which the widest known is Akamai™.

Content delivery for mobile devices has become an increasingly important factor. However, commercially available and free CDNs are mostly not tuned for mobile usage. There is value in working on this issue so that there are CDNs for mobile network users. A sample benefit may be dynamically adjusting video bit-rate according to device type to overcome the fluctuations in mobile networks.

3.4.2.1 CDN Architecture

It may be good to know the CDN architecture and understand the difficulties in routing the requests when considering forming a CDN.

The components of CDN architecture are given in [Peng03cdn] as follows:

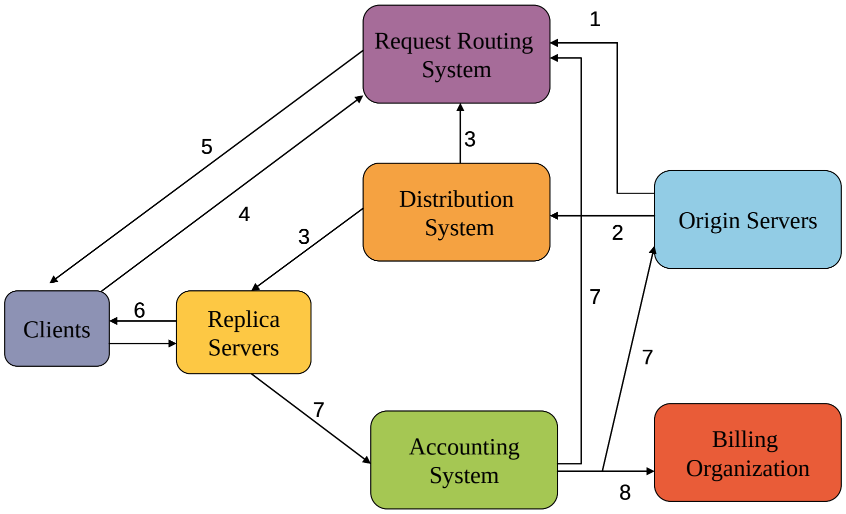


Figure 14: CDN architecture components

1. The origin server delegates its URI (Uniform Resource Identifier) name space for document objects to be distributed and delivered by the CDN to the request routing system.
2. The origin server *publishes* content that is to be distributed and delivered by the CDN into the distribution system.
3. The *distribution* system moves content to replica servers. In addition, this system interacts with the request routing system through feedback to assist in the replica server selection process for client requests.
4. The client *requests* documents from what it perceives to be the origin. However, due to URI name space delegation, the request is actually directed to the request routing system.
5. The *request routing* system routes the request to a suitable replica server in CDN.
6. The selected replica server delivers the requested content to the client. Additionally, the replica server sends accounting information for delivered content to the accounting system.

7. The accounting system aggregates and distills the accounting information into statistics and content detail records for use by the origin server and billing organization. Statistics are also used as feedback to the request routing system.

8. The billing organization uses the content detail records to settle with each of the parties involved in the content distribution and delivery process.

One of the main issues in CDNs is replica placement. There are several request routing techniques, to use a particular server among a set of replica servers.

Another important issue is routing the requests. A variety of algorithms are used to route the request. These include Global Server Load Balancing, DNS-based request routing, Dynamic metafile generation, HTML rewriting and anycasting. Some routing techniques are described in [Peng03cdn]. Proximity—choosing the closest service node—is estimated using a variety of techniques including reactive probing, proactive probing, and connection monitoring.

3.4.2.2 Operator owned CDN:

Mobile networks will carry a huge amount of traffic of their own subscriber base.

Networks may consider deploying their own CDNs since this may help to reduce inter-domain traffic to a certain extent.

Investigation of network packet statistics in batch manner and classification of users and application types periodically would help networks to host their own CDNs.

In case network operators consider deploying their own CDNs, examining the example of a Content Delivery Network such as Akamai™ would be helpful. Akamai hosts very popular websites, like Yahoo and Monster.com, and websites of many big companies, like IBM and FedEx; it has built its own DNS network, that ensures fast delivery of the requested content by resolving the host name of the URL for the requested content to the IP address of the Akamai replica server that will deliver the desired content to the user most quickly.

Building operator owned CDNs in mobile networks may also bring additional revenues.

It may be of interest to investigate the U.S. situation since LTE is launched in the U.S. The demand in broadband grew so fast that the operators are leaving the unlimited data tariff plans. Operators are also seeking ways so that they do not become bit pipes. Launching CDNs may be one of the solutions to the problem. Sample business case from [gigaom09]:

> AT&T Case: AT&T launched a private content delivery service for video inside company firewalls in April 2009. The explosion of video inside corporate networks is straining resources, according to the carrier. But unlike the tiered service that AT&T is experimenting with for its last-mile consumer networks, it’s offering enterprise customers a service that helps them track, compress and prioritize video traffic within the network.
>
> AT&T explains pretty clearly why this is necessary:
>
> “On the average business day, about one-third of the more than 17 petabytes of traffic traversing AT&T’s global backbone network is video content.”
>
> AT&T, rather than become a dumb pipe that acts solely as a conduit for the video deluge, is hoping to monetize that video traffic. On the enterprise side, it’s offering products such as its AT&T Private Content Distribution Service for inside the firewall; it launched a content delivery network last summer that will likely tie into this offering by delivering content from outside, too. Other carriers such as Verizon, and even Internet backbone providers such as Level 3, have CDN efforts as well.

3.4.3 Video optimization techniques

This subsection addresses video coding methods over mobile networks.

3.4.3.1 Motivation:

For several years the mobile communication systems have been improving with the fastest growth. In addition, mobile communications can be defined as one of the most active areas of technology development of our time. Beside this, the video multimedia has experienced massive growth as a distinct technology from mobile communication. But the globalization in the mobile communications sector and increased demand for high data rate/low latency video transport of course cause development in communication systems and create a merging between mobile communication and video technologies.

To this end, a significant amount of research has recently been done in this area to make use of existing radio spectrum in an efficient way.

Motivated by this vision, the broadcast video streaming applications offered as one of the most important services of a mobile service provider, will play a greater role in using radio spectrum in future mobile communication systems. In the following subsections, the challenges and research directions are therefore introduced for video streaming over mobile networks.

3.4.3.2 Challenges in video Streaming for mobile communication:

Video streaming of news and entertainment is widely used in mobile communication and needs a steady flow of information with delivery of packets by a deadline. However, a mobile user suffers from fading, shadowing and co-channel interface, the effects of which can cause low data rate/low quality for video streaming. A common way of combating low data rate/low quality in mobile communication can be summarized as shown in *Figure 15*.

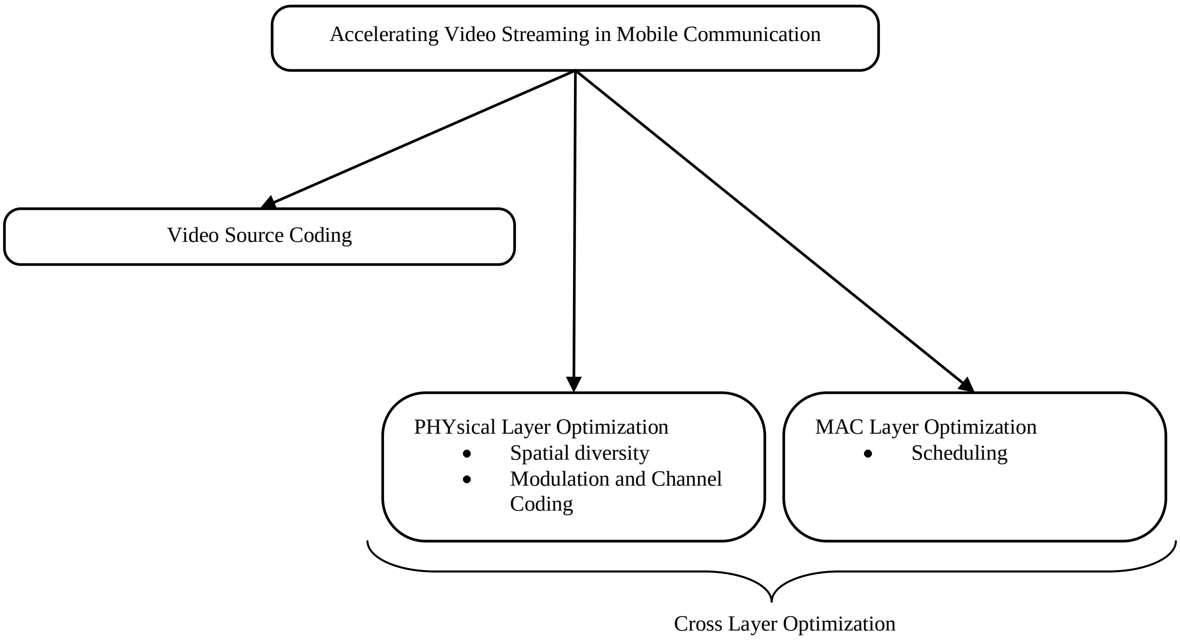


Figure 15: Video acceleration techniques

Cross Networks Optimization

Video streaming is supported by various wireless networks with different characteristics. For example, although a cellular network can provide a wide coverage, it supports a low transmission rate. On the contrary, a WLAN (Wireless Local Area Network) covers a relatively small area but supports a high transmission rate. The coming communication system 4G will provide better service quality. But no single wireless technology can eliminate all other wireless networks. For this reason, cross network approaches are needed to support QoS over such heterogeneous wireless networks.

Cross Layer Optimization

Cross layer optimization is designed to provide the best end-to-end performance based on routing, scheduling and transmission algorithms in any case of network stress.

Although video coding, reliable transport and wireless resource allocation must be considered jointly for such systems to deliver the best end-to-end performance, the video coding methods can be more feasible to implement than studying on PHY and MAC layer due to network complexity.

3.4.3.3 Video source coding optimization

Focusing on the main challenges of video coding for mobile communication, three items can be listed below.

Coding Efficiency:
A real time transmission of a CIF (Common Intermediate Format) video at 15 frame/sec over a 9.6 kbit/s GSM channel requires a compression ratio of about 1900:1. Although new coding technologies are capable of providing such compression ratio, there is a need for higher coding efficiency to improve quality of video at low bit rates.

Computational Complexity
Unfortunately, the needed processing capacity for coding is very limited due to power constraints in mobile devices. For this reason, usage of reduced complexity methods becomes important in mobile communication. The motion related processing is computationally more complex than all other encoding steps. This means that by reducing complexity in motion estimation, the overall system complexity can be reduced.

Error Resilience:
As mentioned in the motivation part, a video stream can be affected by a hostile channel environment that causes high bit rate error and degraded service quality of video communication. Especially, in video transmission, the coded bits are more sensitive to bit stream errors. Therefore, error-resilience techniques are vital to transmit video streaming over mobile channels.

In the future, the data rates in mobile communication will be improved and extended by WiMAX or LTE technologies supporting rate of 72 Mbit/s per cell and 100 Mbit/s per cell, respectively. Improvements are also expected for the short range technologies such as Bluetooth in mobile communication. All mentioned extensions for these techonologies create new user demands for higher throughput and correspondingly higher traffic load. In this sense, video acceleration methods may gain prominence, but implementation stages of the new proposed techniques can cause new challenges for several reasons, especially complexity in PHY and MAC layer. Therefore, a service provider can focus more on new coding/compression techniques than on cross layer optimization between PHY and MAC layer.

3.5 Steering user behaviour

In this sub-section we deal with business related mechanisms, which can be used by the network operator as a business-driven toolset in order to influence behaviour of masses of mobile users. Functions related to this building block will not be investigated in detail by the project consortium. More precisely the traffic usage of users should meet certain pre-defined conditions. On one hand there are conditions for targeted network usage. On the other hand there are incentives in order to influence user behaviour to comply with certain rules. The following usage conditions are envisioned:

- (not) to connect to network or specific access points at a certain time
- subscriber stays within a certain traffic volume
- user restricted to certain applications
- user restricts to certain devices
- user restricts to certain usage modes and scenarios

Some of the restrictions are already used today in a rather inflexible way. These conditions may apply in a general way for the user subscription or are bound to a fixed time interval, e.g. on monthly base. This leads to inefficient resource usage.

The following incentives could be envisioned

- reduced charging
- increased traffic priority
- additional services to be used
- different kind of allowances and discounts

A major concern of the used approach is that the notion of 'network neutrality' is violated, which postulates fair treatment of communication services regardless of users, used applications, devices, type of access, etc.

A basic challenge for this approach is to make user behaviour transparent in order to enable users to react in a proposed way. Usually traffic patterns are influenced by applications, which is non transparent for the user.

3.6 Deployment of new network resources

Traffic-driven upgrades are a dominant cost factor on fixed and mobile Internet access platforms. Functions related to this building block will be likely further investigated by the project consortium. Studies on Internet traffic growth observe annual growth factors in the range 1.5 – 2 since 1990 [Odlyzko03], which are slightly slowing down around 1.45 in the forecast for the next years [HaHa10]. For mobile broadband access, growth rates are much higher in recent years such that traffic is more than doubling within a year [Cisco, AustralianBoS]. As a consequence, bandwidth provisioning has to be steadily adapted to increasing demands where scalability of technological solutions is desirable to avoid too many shifts to next generation platforms in shortening lifecycle periods. The distribution of operational (OPEX) and capital (CAPEX) expenditures depends on the lifecycle duration of technology generations and on the demand for capacity upgrades, as depicted in an example of Figure 16, which also refers to implementation expenditures (IMPEX) that are often included in CAPEX.

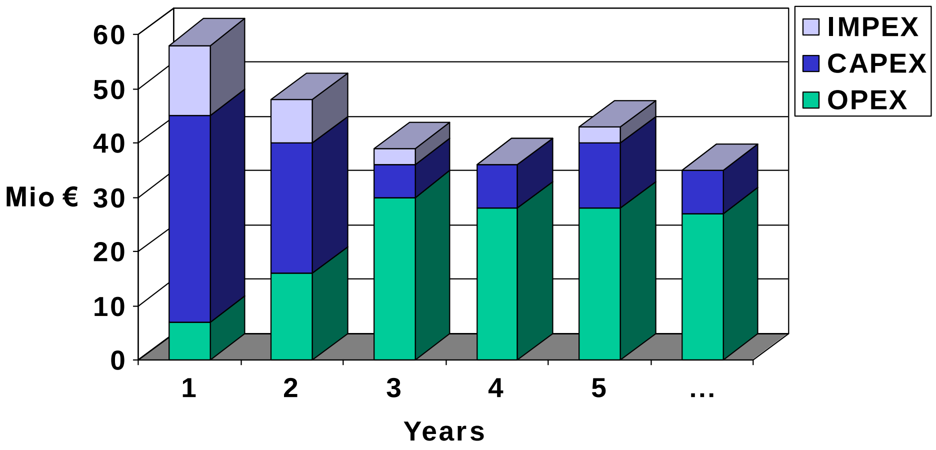


Figure 16: Example of expenditures in the lifecycle of an IP networking platform

Nonetheless, IP network resource utilization is usually low. Among the main reasons contributing to over-provisioning for IP traffic and applications are

- variability in usage pattern due to churn in on-/offline mode and daily or seasonal busy hour profiles,
- high variability of source traffic for many IP applications, e.g. web browsing, video conferencing and streaming, gaming etc.,
- demand for resilience in the aggregation and the backbone,
- inflexible traffic management procedures leading to unbalanced load and congestion due to delayed or insufficient adaptation to changing demands and changing resource availability or changing network topology.

Traffic variability occurs on different time scales from short term bursts on the milliseconds scale, periodical daily/weekly profiles and seasonal effects up to long term growth trends [HaHa10].

Automated load balancing mechanisms are important not only to handle sudden shifts in traffic demand matrices, but also for including new resources becoming available during a network upgrading procedure. In pure IP networks, link upgrades are usually triggered when a load threshold is exceeded. After a non-congested link has been upgraded to *k*-fold capacity, the load on the link goes down by a reciprocal factor 1/*k*. If no redirection of traffic for load balancing is applied, then this leads to a substantial reduction of the mean utilization of link bandwidths. Depending on the upgrade factor, we can quantitatively determine the impact when we assume an exponential traffic growth over time and denote

- t_{Up} for the load threshold, which triggers the next upgrade ($0 < t_{\mathrm{Up}} < 1$),
- f_{Up} for the factor of capacity extension of an upgrade ($f_{\mathrm{Up}} > 1$),
- d_{Up} for the duration of an upgrade period.

Figure 17 shows the usual process of link loads increasing with traffic until they drop down again at the next upgrade, where upgrades by factors 2, 4 and 10 are depicted. IP link capacity usually scales in 4-fold steps (155Mb/s, 622Mb/s, 2.5Gb/s, 10Gb/s, 40Gb/s …), whereas Ethernet proceeds to 10-fold capacity (10Mb/s, 100Mb/s, 1Gb/s, 10Gb/s). Naturally, upgrades to twice the capacity of a single link are always possible by replicating a link. Ethernet standardization is currently developing towards 40Gb/s as well as 100Gb/s.

We compute the mean link load ρ_{Mean} during an upgrade period of arbitrary length d_{Up} in terms of the ratio of the areas below the growth curve and the step functions for capacity upgrades

$$\rho_{\mathrm{Mean}} = \frac{t_{\mathrm{Up}}}{f_{\mathrm{Up}} d_{\mathrm{Up}}} \int_0^{d_{\mathrm{Up}}} \exp(\omega\, x)\, dx = t_{\mathrm{Up}} \frac{1 - 1/\, f_{\mathrm{Up}}}{\ln(f_{\mathrm{Up}})} \qquad \text{where} \;\; f_{\mathrm{Up}} = \exp(\omega d_{\mathrm{Up}}).$$

As a result, the mean link load undercuts the threshold by a factor $(1 - 1 / f_{\mathrm{Up}}) / \ln(f_{\mathrm{Up}})$ independent of other parameters (ω, d_{Up}, t_{Up}). In particular, the mean load ρ_{Mean} on an IP link with usual upgrading process is bounded by

- $\rho_{\mathrm{Mean}} < 0.72\, t_{\mathrm{Up}}$ for $f_{\mathrm{Up}} = 2$, i.e. upgrades to twice the capacity,
- $\rho_{\mathrm{Mean}} < 0.54\, t_{\mathrm{Up}}$ for $f_{\mathrm{Up}} = 4$, i.e. upgrades to 4-fold capacity,
- $\rho_{\mathrm{Mean}} < 0.39\, t_{\mathrm{Up}}$ for $f_{\mathrm{Up}} = 10$, i.e. upgrades to 10-fold capacity.

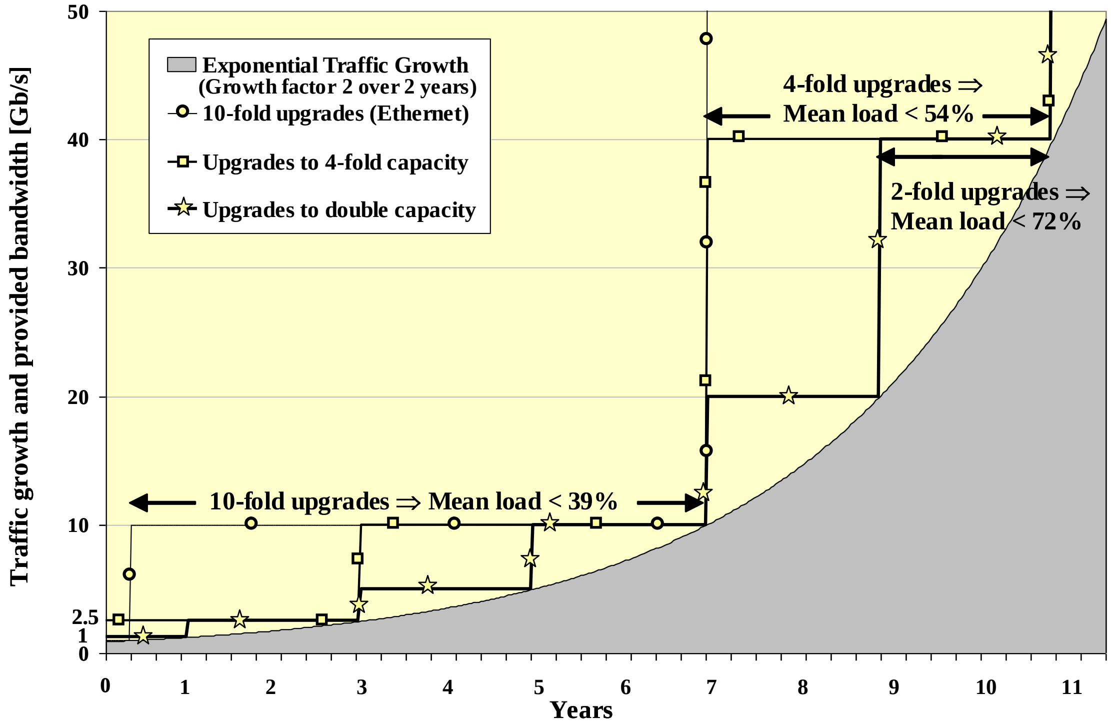


Figure 17: Link upgrades and resulting utilization gaps for exponential traffic growth

Even if the traffic does not strictly follow an exponential growth in practice, the estimates still hold for random variability around an underlying growth tendency, provided that the variability is independent of upgrade times, i.e. if upgrades are not delayed until congestion and then would lead to an immediate rise of traffic after widening an existing bottleneck.

In order to avoid those utilization gaps after upgrades, traffic engineering has to react by redirecting transport paths from links in the surrounding to an upgraded link, such that newly installed bandwidth is instantly exploited to smooth down higher load on other links.

Multiprotocol label switching (MPLS) has been developed by the Internet engineering task force (IETF) as a networking sub-layer providing advanced traffic engineering support in meshed IP networks covering a basic set of required functions. Explicit source-to-destination paths can be set up in MPLS in normal operation and backup paths for failure cases, both with measurement options per path. MPLS is deployed in the backbone and aggregation of many large ISPs and is currently still in development by one of the most active working groups within the IETF [IETF-MPLS].

As a precondition to perform and control load balancing on meshed telecommunication networks, a set of traffic engineering functions has to be provided as a basic part of network management including

- monitoring and situation awareness of the current network topology and traffic on the links and nodes,
- the ability to direct traffic on explicit paths from source to destination for load control,
- threshold based decisions to adapt to changing load conditions and to trigger upgrades,
- online tools to compute and configure optimized traffic path designs in a network-wide view and
- fast failure resilience by automated and instantaneous switching to backup paths that have been prepared for a set of most relevant failure cases, at least for single link breakdowns.

When looking at current networking scenarios we seldom find full support of all those traffic engineering prerequisites. In pure IP networks, measurement of the link loads in 5- or 15-minute intervals has been the only available information for network planning for a long time. Although IP routing automatically adapts to failures, it often lasts for several seconds until routing information and rerouting paths stabilize to adapt to changes in topology. In addition, it is difficult to enhance the shortest path first principle with load balancing and control by manipulating the routing weights. Some types of optical ring networks (e.g. Synchronous Optical Hierarchy) can react on an interruption of the ring within less than 50ms by switching the traffic in the opposite direction of the ring but without network wide load balancing.

In next generation mobile networks, different opportunities for load balancing are relevant. The deployment of new resources, e.g. new or upgraded base stations can improve coverage up to its capacity limits, while other congested cells in the neighborhood may be able to concentrate on a smaller area to reduce their load. Load balanced paths direct a portion of traffic not on the shortest path and thus may increase the total load when traffic traverses more than the minimum number of links. But the optimization goal of minimizing the maximum load over all links allows only a small amount of deviations to unload congested links and on the other hand allows for a maximum throughput in terms of a linear scaling factor of the traffic matrix.

Traffic engineering tools have been developed, e.g. TE-Scout [HaSnFr05], to compute optimized path designs for load balancing for arbitrary network topologies and traffic demand matrices based on operations research methods including linear programming and simulated annealing. The application of optimization tools has to be embedded in a complete traffic engineering cycle for monitoring of the current topology and traffic demand matrix, via re-optimization after relevant shift are observed until reconfiguration of the new adapted path design into the routers.

As main conclusions of the gain analysis of traffic engineering for full use of new installed bandwidth during upgrading processes in broadband access networks reveals a considerable optimization potential for higher throughput as compared to upgrading processes without load balancing. Under the assumption of exponential traffic growth and a cost model involving exponential decrease in costs per bandwidth unit over time, analytical results are derived for the benefit of load balancing in full mesh networks for normal operation as well as failure cases [Ha-et-al 1]. The evaluation with realistic traffic growth and cost decrease rates shows that flexible adaptation of traffic paths can save at least 20% of the CAPEX and the same percentage of energy consumption as part of the OPEX.

4. Traffic management requirements

4.1 General requirements for traffic management

N°	Title	Description	Priority	Reference
0.1	Application-awareness	The traffic management must be able to provide means for traffic classification based on application types. An additional remark here is that not all applications can be handled according to their type and content but need to be classified into a categrory such as “Internet” or “Background” traffic for scalability reasons.	High	
0.3	Scalability	The traffic management solutions should scale well with the increasing traffic demands stated in IR1.1. This requirement must be fulfilled by all functions including measurement, collection of data, analysis of statistics, modelling (network topology, link states, prognosis of traffic characteristics), evaluation of triggering events of Traffic Engineering (TE) controls, execution of TE controls.	High	
0.4	Privacy	The inspection of user traffic must remain within the boundaries of the ePrivacy Directive and governmental legislations.	Medium	

4.2 Requirements for macroscopic traffic management

N°	Title	Description	Priority	Reference
1.1	Transport resource availability in mobile NE selection	When assigning mobile NEs (e.g. PDN GW, S-GW, MME, eNodeB) to UEs the available transport network resources should be taken into account in the NE selection decisions.	Medium	3.1.3
1.2	E2E traffic steering	Traffic steering along alternative paths must take into account transport resources on the whole routing path between the UE and PDN GW (and not only radio resources).	High	3.1.3
1.3	Local traffic offloading and mobility	Local traffic offloading solutions should not degrade mobility support.	Medium	3.1.4.1
1.4	Multi-criteria cell selection	Signal quality is a mandatory metric in cell selection. But several popular applications have specific needs in e.g. bandwidth and delay. Metrics reflecting them should be involved in the cell selection. Information on bandwidth should be distributed by an efficient Radio Resource Management (RRM).	Low	3.1.6.1
1.5	Optimising scanning operations	Scanning is costly and should be reduced to the “best possible” candidate cells. The network should prepare an ordered success probability candidate list to minimize the scanning time. The cell ordering should also involve the HO success rate w.r.t., the 3D position of the UE, which could be stored in the BSs.	Low	3.1.4.2

1.6	HO manager	HO can be initiated by either the network or the terminal to maintain the quality of signal, QoE, or load balance. The HO phases then are different. A management tool is needed to guide the HO process w.r.t. these modes. Hosted in the network, it should also be connected with Subscriber Service maps or RRM.	Medium	3.1.4.2
1.7	Keep additional forced handover traffic low	The forced handover can often significantly increase the amount of transmitted control information, but the efficient handover algorithms are expected to keep the additional control traffic at low level.	Medium	3.1.4.2
1.8	Usage of proper on/off switching cost function	In on/off switching methods it is important to find an adequate cost function in order to attain a result that is a good approximation of the optimal solution.	High	3.1.5
1.9	Uplink bottleneck detection in cell breathing	In some high rate communications (e.g. video telephony) the traffic is symmetric; therefore it should be possible to detect uplink bottlenecks in the transmission.	High	3.1.6.1
1.10	Keep the rules of frequency planning in reclustering	The frequency planning requires thorough and deliberate work and in a reclustering solution Trestrictions also have to be considered for traffic management.	Medium	3.1.6.2
1.11	Pattern adjustment to interferers in beamforming	Apart from the antenna radiation pattern adjustment to the wanted signal, the pattern is also expected to be adjusted to null out the interferers.	Low	3.1.6.3
1.12	Access technology reselection	It may be beneficial to relocate user(s) between the different RANs in order to release available capacity for other users.	Medium	3.1.1

4.3 Requirements for microscopic traffic management

N°	Title	Description	Priority	Reference
2.1	QoS support in upstream direction	Several applications require frequent uploading of content into the network. For delay critical applications, QoS support is needed in this context.	Medium	3.2.5
2.2	Ad hoc change of QoE	The user should be able to change QoE on demand for running applications	Medium	3.2.3
2.3	Support for multipath flows	In order to increase flexibility of balancing load among multiple paths and to improve QoE for the user, it should be possible to make use of multipath flows	High	3.2.2
2.4	Combine user prioritization with application differentiation	A combined scheme may enable improved resource usage.	Medium	3.2.1
2.5	Selective admission control	An incoming flow may be deliberately rejected by the network	Medium	3.2.6

2.6	QoE support in roaming case	The user should be able to have the same QoE in case of roaming.	High	3.2
2.7	Cross-layer interference detection	Access points or network devices in the core network, which are located near to the wireless border, will be made capable to identify interference by means of cross layer traffic monitoring.	High	2.2.2.1

4.4 Requirements for improved content resource selection & caching

N°	Title	Description	Priority	Reference
3.1	Peer or storage selection optimised for mobile networks	Application Layer Traffic Optimization (ALTO), as designed in the IETF involves the operator centric transport layer topology but focuses on fixed provider cores. It should be adaptable to mobile networks. The mobile network architecture should be able to host and support ALTO functions (e.g. for supporting mobility-aware P2P applications).	Medium	3.3.3
3.2	Resource partitioning	It should be possible to distribute content into smaller portions in order to improve resilience and QoE.	Medium	3.3
3.3	Detect unfavourable resource usage	In case that a user is connected to an unfavourable resource, the network should be able to detect this incident.	Medium	3.3.4
3.4	Reduce P2P transit traffic	The P2P traffic management should minimize inter-AS traffic between the different domains of the operators.	High	3.3.3

4.5 Requirements for steering user behaviour

N°	Title	Description	Priority	Reference
4.1	Monitor user compliance	The operator should be able to verify user compliance of usage constraints.	Medium	3.5

4.6 Requirements for deployment of new network resources

N°	Title	Description	Priority	Reference
5.1	Seamless deployment of new resources	Whenever possible, new resources should be integrated in a self-organizing and seamless way including self-testing. Existing network entities should become aware of new entities	Medium	3.6

		based on automatic discovery protocols and the new entity should learn about the environment through data exchange with its new neighbors.		
5.2	Support for upgrading processes	The transmission capacities should be easily adaptable to steadily increasing traffic within a sufficiently wide scalability range.	High	3.6
5.3	Immediate usage	New resources should be made fully available from the start, e.g. by redistribution of the workload to reduce load on other networking entities especially those in high or over-load.	Medium	3.6

5. Conclusions

The motivation of this document is to provide an overview of the general objectives for traffic management, introduce the concept of building blocks and formulate requirements in the context of the project. It is assumed that these results can be a valuable input for the definition of the traffic management architecture.

The general objectives of traffic management include traffic modelling, traffic control mechanisms and techno economical aspects. The main interest in this document is on the initial analysis of traffic control mechanisms while the other aspects are described within other documents.

The concept of building blocks has been introduced in order to categorize traffic management mechanisms with common or similar functionality and solution space - six building blocks have been identified altogether. One aspect is the documentation of the different approaches of state-of-the art and novel mechanisms. Secondly those mechanisms are highlighted within this document, which will be further elaborated by the project consortium when developing the traffic management architecture. For the different mechanisms first analysis has been carried out for state-of-the-art technology and based on usage scenarios and assumptions, high level problem statements have been described.

The building block microscopic traffic management includes all mechanisms with the primary objective to improve performance of individual flows. The following mechanisms will be likely further investigated by the project consortium: QoS differentiation based on applications and user profiles, improvements on TCP layer, selective admission control and upstream flow control

The building block macroscopic traffic management includes all mechanisms with the primary objective to improve efficient usage of network resources. The project partners will likely put focus on the following topics: access technology reselection, selection of core network elements, change routing within backhaul, offloading techniques and equipment switch on / off.

The building block ‘Improved Resource Selection and Caching’ represents storage and selection functions in distributed data management systems (such as CDN and P2P). The following topics will be likely further investigated by the project partners: Caching of popular Internet content and user hosted content, redirection of requesting node to alternative resource and detection of unfavourable resource usage.

The building block “Steering User Behaviour” is more relevant from business perspective without too many technical aspects. Mainly network operators but possibly other stake holders as well may influence user behaviour by defining certain constraints for usage of networks / services and certain incentive to comply with the usage constraints. Details are not further analysed by the project consortium.

Application supported traffic management describes mechanisms which aim to optimize performance from end user perspective without getting support from lower layers, such as many CDN and P2P based implementations. Since network functions are not needed within this building block, the included mechanisms are not further investigated by project partners.

Finally we have defined one further building block, which is about capacity extension in case the available network is frequently in high load conditions. It is the challenge to apply an intelligent planning process for extending the available resource. In contrast to the other building blocks mentioned previously, capacity extension is a process which will become effective in the network after a long time period, possibly up to several months. These aspects are also further elaborated in the context of the project.

Even though the building blocks and the associated mechanisms 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 will be a follow up activity to map the diverse mechanisms into concrete 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. Common functionality related to the traffic management building blocks. such as policy control and network monitoring are also for further study.

Finally we have defined general requirements related to traffic management based on the architecture investigations in the project. In addition building block specific requirements have been defined based on the above mentioned problem statements.

References

[3GPP_22.220] 3GPP TS 22.220 V10.4.0, Technical Specification, “Service requirements for Home Node B (HNB) and Home eNode B (HeNB)”, Release 10, September 2010.

[3GPP_23.107] 3GPP TS 23.107 V9.1.0 Technical Specification “Quality of Service (QoS) concept and architecture”, (Release 9),. June 2010

[3GPP_23.261] 3GPP TS 23.261 V10.1.0, Technical Specification, “IP flow mobility and seamless Wireless Local Area Network (WLAN) offload, Stage 2”, Release 10, September 2010.

[3GPP_23.829] 3GPP 23.829 V1.3.0, Technical Report, “Local IP Access and Selected IP Traffic Offload”, Release 10, September 2010.

[3GPP_23.401] 3GPP TS 23.401 V10.3.0, Technical Specification, “General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access”, Release 10, March 2011.

[3GPP_24.007] 3GPP 24.007 V10.0.0, Technical Specification, “Mobile radio interface signalling layer 3; General Aspects”, Release 10, March 2011.

[3GPP_29.303] 3GPP TS 29.303 V10.0.0, Technical Specification, “Domain Name System Procedures, Stage 3”, Release 10, December 2010.

[4WARD] 4WARD project of the European Commission, Deliverable D-4.3: Inherent Network Management (INM) Design, <www.4ward-project.eu/index.php?s=Deliverables> (2010)

[Abboud09] O.Abboud, A. Kovacevic, K. Graffi, K. Pussep, R. Steinmetz, “Underlay

awareness in P2P systems: Techniques and challenges”, IEEE International Symposium on Parallel&Distributed Processing, 2009.

[AustralianBoS] Australian Bureau of Statistics, Internet activity <www.abs.gov.au/ausstats/abs@.nsf/mf/8153.0> (2011)

[Basher08] Basher, N., Mahanti, A., Mahanti, A., Williamson C., & Arlitt, M. (2008). A comparative analysis of web and peer-to-peer traffic. In *Proceedings of the 2008 WWW Conference*. Beijing, China.

[Bindal06] Ruchir Bindal, Pei Cao and William Chan, Improving Traffic Locality in BitTorrent via Biased Neighbor Selection.. ICDCS 2006, pp. 66-77.

[Choffnes08] D. R. Choffnes, F. E. Bustamante. Taming the torrent: A practical approach to reducing cross-ISP traffic in P2P systems. ACM SIGCOMM 2008.

[Cisco] Cisco Systems, Visual Networking Index / Global IP traffic forecast and methodology; White paper Series <www.cisco.com> (2008-11)

[Ekström 09] .Hannes Ekström, QoS control in the 3GPP evolved packet system, IEEE Communications Magazine, v.47 n.2, p.76-83, February 2009

[Gigaom09] http://gigaom.com/2009/04/21/att-creates-private-cdn-for-corporate-video/

[Graffi08] K. Graffi, A. Kovacevic, S. Xiao, and R. Steinmetz. Skyeye. kom: An information management over-overlay for getting the oracle view on structured p2p systems. In In Proc. of IEEE ICPADS08 Conference, Melbourne, Australia, 2008.

[HaHa10] G. Haßlinger und F. Hartleb: Internet-Verkehr: Plattformen zur Datenverteilung, Zugriffsstatistiken und Optimierung von Transportpfaden, PIK – Praxis der Informationsverarbeitung und Kommunikation 33, de Gruyter (2010) 85-94

[Haßlinger11] G. Haßlinger, G. Nunzi, C. Meirosu, C. Fan and F.-U. Andersen: Traffic Engineering Supported by Inherent Network Management: Analysis of Resource Efficiency and Cost Saving Potential, to appear in the International Journal on Network Management (2011)

[Haßlinger 10] G. Haßlinger und F. Hartleb: Internet-Verkehr: Plattformen zur Datenverteilung, Zugriffsstatistiken und Optimierung von Transportpfaden, PIK – Praxis der Informationsverarbeitung und Kommunikation 33, de Gruyter (2010) 85-94

[Haßlinger 05] G. Haßlinger, S. Schnitter and M. Franzke: The Efficiency of Traffic Engineering with Regard to Failure Resilience, Telecommunication Systems Vol. 29/2 (2005) 109-130

[IETF-MPLS] Internet Engineering Task Force (IETF), Multiprotocol Label Switching (MPLS) Working Group, <www.ietf.org/html.charters/mpls-charter.html>

[Ipoque] Ipoque. Internet Studies 2008. [Online] http://www.ipoque.com.

[Kouloris03] T. Koulouris, R. Henjes , K. Tutschku. Implementation of Adaptive Control for P2P Overlays. Proceedings Fifth Annual International Working Conference on Active Networks (IWAN 2003)

[Lehrieder10] Frank Lehrieder, György Dán, Tobias Hoßfeld, Simon Oechsner, Vlad Singeorzan. The Impact of Caching on BitTorrent-like Peer-to-peer Systems, IEEE P2P 2010, 2010.

[LeBlond10] Stevens Le Blond, Arnaud Legout, Walid Dabbous Pushing BitTorrent locality to the limit,. Computer Networks : Elsevier, 2010. DOI:10.1016/j.comnet.2010.09.014.

[MobileP2P] Haitao Wang, Lihua Song, "Applications and Challenges of Mobile P2P Systems in Ad hoc Network ", Advanced Materials Research Vols. 171-172, pp 575-579, 2011.

[Nonnenmacher99] J. Nonnenmacher, E.W. Biersack, Scalable feedback for large groups,. IEEE ACM Trans. Netw, 1999., 7. kötet, old.: 375-386.

[Odlyzko03] A. Odlyzko, Internet traffic growth: Sources and implications, Proc. SPIE Vol. 5247

[OverCache] P2P, OverCache. http://www.oversi.com. [Online]

[P2PNext] P2P-Next. http://www.p2p-next.org/

[P4P] H. Xie, Y. Yang, A. Krishnamurthy, Y. Liu, and A. Silberschatz, P4P:Provider portal for P2P applications ACM SIGCOMM, 2008.

[P4Pslide] www.dcia.info/activities/p2pmslv2008/1-6%20**P4P**WG.ppt

[Peng03cdn] Gang Peng, CDN: Content Distribution Network, 2003

[RFC3272] D. Adwuche, A. Chiu, A. Ewalid, I. Widjaja, X. Xiao, "Overview and Principles of Internet Traffic Engineering", IETF RFC 3272, May 2002.

[RFC5555] R. Soliman, Ed., "Mobile IPv6 Support for Dual Stack Hosts and Routers", IETF RFC 5555, June 2009.

[Shiao01] Shiao-Li Tsao: "Scalable gateway GPRS support node for GPRS/UMTS networks", 56th IEEE Vehicular Technology Conference (VTC Fall'02), Volume 4, pp. 2239, 2002.

[Spanoudakis08] G. Spanoudakis, C. Kloukinas and K. Androutsopoulos Dynamic Verification and Control of Mobile Peer-to-Peer Systems. The Third International Conference on Internet Monitoring and Protection, 2008

[Xue01] C. Xue, J. Luo, R. Halfmann, E. Schulz, C. Hartmann, “Inter GW Load Balancing for Next Generation Mobile Networks with Flat Architecture”, In proc. of IEEE 69th Vehicular Technology Conference, VTC Spring’09, pp.1-5, Barcelona, Spain, April 2009.

[WE2E] “Analysis of Traffic Offload: WiFi to Rescue”, White Paper, WirelessE2E LLC, September 2010. http://trusted-advisor-mobile.com/Analysis_of_Traffic_Offload_files/Wirelesse2e_Trafficoffloadwhitepaper_09262010.pdf