Traffic modeling in FMC network scenarios

Executive Summary of the Deliverable

The COMBO project will propose and investigate new integrated approaches for Fixed Mobile Convergence (FMC) for broadband access and aggregation networks. COMBO will target on an optimized network infrastructure ensuring increased performance, reduced cost and reduced energy consumption. Under this general objective, WP2 defines the preliminary work on fixed, mobile and converged networks in order to put the basis for the remaining work packages, specifying the overall reference framework for fixed and mobile networks, the most relevant FMC network use cases, the traffic studies and the requirements that will be the base of the COMBO project architecture COMBO WP2 will provide the preliminary work for fixed and mobile networks analysing their current status and evolution trends.

This deliverable (D2.3) of Task 2.3 is called *Traffic modelling in FMC network scenarios*. Its main targets are to analyze traffic scenarios for FMC networks from the current status with a 2020 time horizon; to analyze which are the key drivers of traffic growth and their impact; to study fixed and mobile traffic models; and to analyze how traffic and its related parameters will evolve in different FMC scenarios.

D2.3 contains valuable information for the evolution of the traffic, especially considering the effects of FMC scenarios. It presents current analysis results of high-speed aggregated traffic, as well as a comparison of traffic similar measurements carried out at two leading operators, pointing out similarities and differences in traffic patterns in order to provide valuable input for various aspects of traffic modelling and their effect on modelling. Furthermore, it provides traffic modelling as well as aggregation network planning methodology based on traffic analysis. These are important for COMBO in order to understand the drivers and the variables of traffic change, and the effect of these. It then helps to understand how the various FMC scenarios are able to solve the upcoming issues with the various traffic demands.

The change of traffic patterns at FMC network scenarios in comparison with fixed and mobile networks without considering FMC architectures can be seen from two different perspectives. On one hand, the content-specific characteristics of the traffic – and on the other hand, the way how traffic flows. The volume and the various traffic-descriptor statistics will change, including the application mixture changes in some links/connections. Due to the new possibilities provided by the convergent architecture, some traffic will take different routes instead of current ones.

This document summarizes the reasons of these changes, and makes an effort of predicting how the traffic-related changes will take effect in the future - especially in the relation of specific FMC scenarios. It starts with a methodology definition and a brief description of the current traffic situation in different network segments. After discussing current traffic demands, and various drivers for traffic growth – such as the increasing number of devices, the introduction of new services and applications, and the effect of these factors on bandwidth demands – a forecast on traffic trends is provided. The analysis of traffic can support whether our current models provide a good basis of modelling FMC traffic – and show what are the changes in methodology and parameters that should be applied. The modelling parts of the document cover packet-level arrival processes and long-range dependency; as well as modelling on macro-level (traffic mixtures, user behavioural patterns, etc.).

Furthermore, the document describes how traffic patterns and models are affected by some of the FMC scenarios. One of the models here is described for offloading scenarios with WLAN femtocell, and the other model describes traffic composition, and the expected evolution within aggregation networks.

This deliverable builds on the results on Task 2.1 (Reference framework) and Task 2.2 (Fixed and mobile network evolution) within COMBO.

The study on the analysis of current traffic demands and forecast provides insight and comparative data for WP5 (Techno-economic assessment), as well as for WP4 (Traffic and Performance Management).The document contains a description on traffic analysis and modelling, providing a general overview of the current traffic modelling methodology and publicly available reference data (traffic archives). The available modelling methodologies, and predictions on the traffic changes and growth, as well as the results and conclusions on traffic analysis and modelling have an impact on architectural issues, hence it affect the work in WP3 (Fixed Mobile Convergent architectures). This provides input on some of the traffic-related discussions at Task 4.2 (Performance Management), as well.

These previous studies and their results will be used in other COMBO tasks, for example: in Task 3.2 and 3.3 (Fixed/mobile convergence at protocol level; Fixed/mobile convergence at equipment and infrastructure level) Task 4.2 (Performance Management), Task 5.3 and 5.4 (Impact of convergence on business ecosystems; Analysis of energy consumption).

Beside this, the analysis results published here - as well as the procedures that invoked data capture and analysis activities at various partners' networks - can be used as reference points at some WP6 activities (Functional Development & Experimental Research Activities). The discussion on offloading scenarios and the FMC traffic model for aggregation networks support the work of WP3, WP4 and WP5 activities by providing traffic models for some of the scenarios referenced at these work packages.

This document supports the advances of FMC and traffic modelling through the following:

- current traffic demands, trends, drivers are identified – providing room and direction for plans and developments requiring such information;
- traffic analysis at given reference points were carried out at different operators' networks – similarities on the results reinforced that generalized models can be built, and differences reveal the important variables to be built into future models;
- measurements on Poisson and Self-similar behaviour showed that real traffic are inevitably complex;
- self-similarity should be considered when creating FMC traffic models to obtain valid results for capacity planning – we do have to use complex models for real cases;
- measurements at PoPs showed that traffic growth due to mobile traffic alone is not a potential driver to change the aggregation network structure or even re-direct traffic from mobile users closer to the edge of the network affecting the network architecture.

Furthermore, it describes models built for various aspects on FMC network scenarios:

- a traffic analysis and complex modelling methodology for macro-level;
- a model for traffic offloading with WLAN or femtocells;
- a model for FMC traffic for aggregation networks.

Table of Contents

1 Introduction

This document provides the basis of traffic scenarios and modelling for fixed and mobile convergent networks starting with the study of the state of the art and the current evolution trends of both networks from a non FMC point of view done in Task 2.2 and the reference framework, uses cases and initial market studies done in Task 2.1.

This deliverable (D2.3) of Task 2.3 is called *Traffic modelling in FMC network scenarios*. Its main targets are to analyze traffic scenarios for FMC networks from the current status with a 2020 time horizon; to analyze which are the key drivers of traffic growth and their impact; to study fixed and mobile traffic models; and to analyze how traffic and its related parameters will evolve in different FMC scenarios.

Figure 1 shows the scope of this document in the context of the whole COMBO project. In this respect, this document covers the box "Traffic" with the target to describe the traffic-related issues for the FMC studies in COMBO.

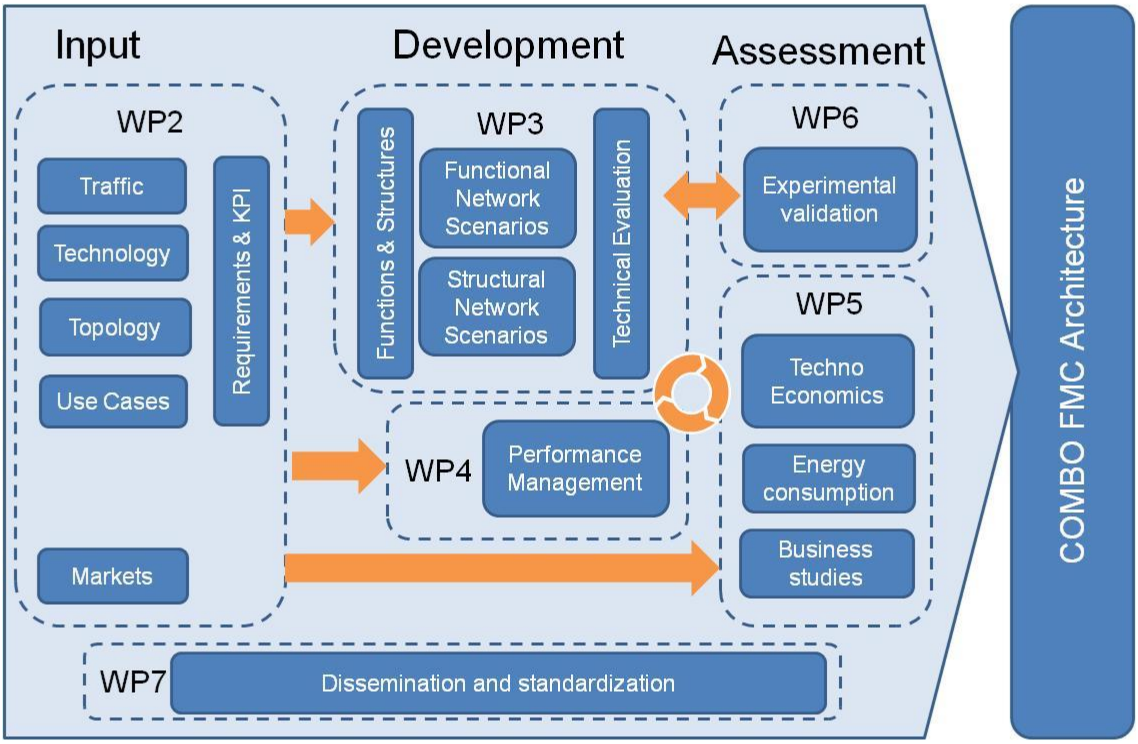


Figure 1: Work structure of COMBO project.

In the first technical chapter after the introduction, the document describes the overall methodology of gathering traffic-related data and attempting to set up models based on them. Furthermore, it aims to find some common understanding on what to analyse at, and how to model traffic scenarios. There is also a discussion here about traffic matrix (where and how the traffic flows in what volumes), about application mix and how is this expected to change with FMC deployments.

Traffic evolution with a 2020 time horizon should be analysed based on earlier experiences and the overview of current status. After discussing current traffic demands, and various drivers for traffic growth - such as the increasing number of devices, the introduction of new services and applications, and the effect of these factors on bandwidth demands - a forecast on traffic trends can be provided. The conclusions of these predictions include a 3.5x busy-hour Internet traffic increase from 2012 to 2017 (865 Tbps), and a 1000x traffic increase in a 20 years span. A brief overview of Content Delivery Networks - as an important optimization answer of some traffic routing issues, and signalling changes are also part of this chapter of the document.

In order to provide traffic models of current and future scenarios - including FMC -, traffic analysis of current patterns should be covered. Although the methods - and hence, the toolset - of modelling and analysis differ, they should be handled together, since they provide input requirements and feedback to each other. The traffic analysis and modelling section of this document covers methodology discussion and actual analysis results as well. These include analysis of publicly available traffic archives, as well as analysis of aggregated traffic patterns and bit-by-bit captures. In the modelling part the document covers packet-level arrival processes and long-range dependency, as well as modelling on macro-level (traffic mixtures, user behavioural patterns, etc.).

The last chapter of the document describes how traffic patterns and models are affected by some of the FMC scenarios. One of the models here is described for offloading scenarios with WLAN femtocell, and the other model describes traffic composition, and the expected evolution within aggregation networks.

2 Definition of traffic scenarios and models

2.1 Motivation and Methodology

This section briefly provides the main motivation behind the need for defining traffic scenarios, and modelling traffic. Furthermore, since partners have different understanding of what a traffic scenario or model may mean, we describe the agreed, simple method of consolidating these terms.

Since partners come from various organizations, there is a need for understanding what they – especially operators – mean by "traffic scenario", and to identify what is the common understanding of differentiating traffic types. This differentiation can be based on many factors, i.e. what kind of applications provide such a traffic; statistical properties; QoS demands; network type and network segment that carries such a traffic, etc.

Partners agreed on following a certain methodology for reaching an agreement on the above questions. This included steps such as:

- operators providing input, answering with their view of these definitions,
- consolidation of different naming conventions,
- definition of various traffic scenarios.

The first two steps were carried out within task 2.3, whereas the third step was a combined result of discussions within task 2.1 and two use cases of deliverable D2.1 "Framework reference for fixed and mobile convergence". The use cases UC04 "Universal access bundling for residential gateway" – and UC06 "Convergence of fixed, mobile and Wi-Fi gateway functionalities" were chosen to be part of the analysis within D2.3, due to their effect on changing traffic mix or changing traffic matrix.

2.2 Summary of views on traffic scenario and modelling

Based on the operators' feedback it seems that providing an overall, general description of FMC traffic mixtures is not an expectation. Rather than that, various scenarios are to be considered, analysed and modelled from traffic evolution point of view. The defining parameters of these traffic scenarios can widely vary depending on the network domain, the traversing architecture (node and connection types), the user behaviour, and many more.

This implies that the traffic modelling used for these scenarios can and should consider various levels of traffic descriptors. Depending on the underlying motivation, the traffic model can be defined at the level of packet-interarrival times, flow dynamics and other statistical distributions of packet- and flow-level parameters. On the other extreme it can be defined based on user behaviour characterization, architectural constraints and low-granularity descriptors of traffic volumes traversed across network segments.

2.3 Where that traffic flows: differentiation by traffic paths

This section describes a path in different segments of network for popular applications currently used in Fixed and Mobile Networks. This study will be used to make a qualitative assessment of the traffic volume in different segment of FMC networks in order to estimate how the traffic will change (or something similar).

Firstly we choose the applications most used and/or that generate most of the data traffic (e.g., Internet data traffic, traffic controlled by operator) based on real traffic data. Secondly we will show a qualitative description of traffic volume in different segments of Fixed Network. Thirdly we describe a qualitative traffic volume in different segments of Mobile Network; finally we indicate a qualitative description of traffic volume in the case of a FMC network based on use cases UC04 and UC06.

2.3.1 The choice of applications

Internet Traffic

For Internet traffic, we only consider applications that generate most of traffic in upstream and downstream. These applications are selected from section 3.1 coming from real traffic demand provided by FT and TID.

In downstream, we consider applications that generate most of total traffic and that represent minimum of 15% volume in different segments of the network. These applications are usually video as stated in section 3.1: Streaming Video and Download.

In upstream, Internet usage is strongly related to applications for file sharing in peer-to-peer mode (P2P). These files can be a video or other content (e.g., images, music). Peer-to-peer applications contribute majorly in terms of upstream volume in different segments of the network.

Traffic controlled and provided by operator

For this type of traffic we chose the applications that are frequently used although they can contribute little in load of different segments of the network compared to the Internet traffic. The most important aspect is to know the routing of these prioritized applications in different segments of the network. These applications are: TV Multicast, Video streaming (IPTV unicast) and VoIP.

2.3.2 Qualitative description of traffic volume in different segments of fixed Network

Downstream traffic

IPTV services are routed using multicast; the rate of multicast traffic decreases when this traffic approaches the final customers (e.g., core → aggregation → access → customers).

IPTV multicast is transported from the TV service provider (see Figure 2) via the core and the aggregation networks, arriving at the OLT or DSLAM which will distribute the necessary channels to all clients attached to the Passive Optical Network (PON) or to the DSLAM.

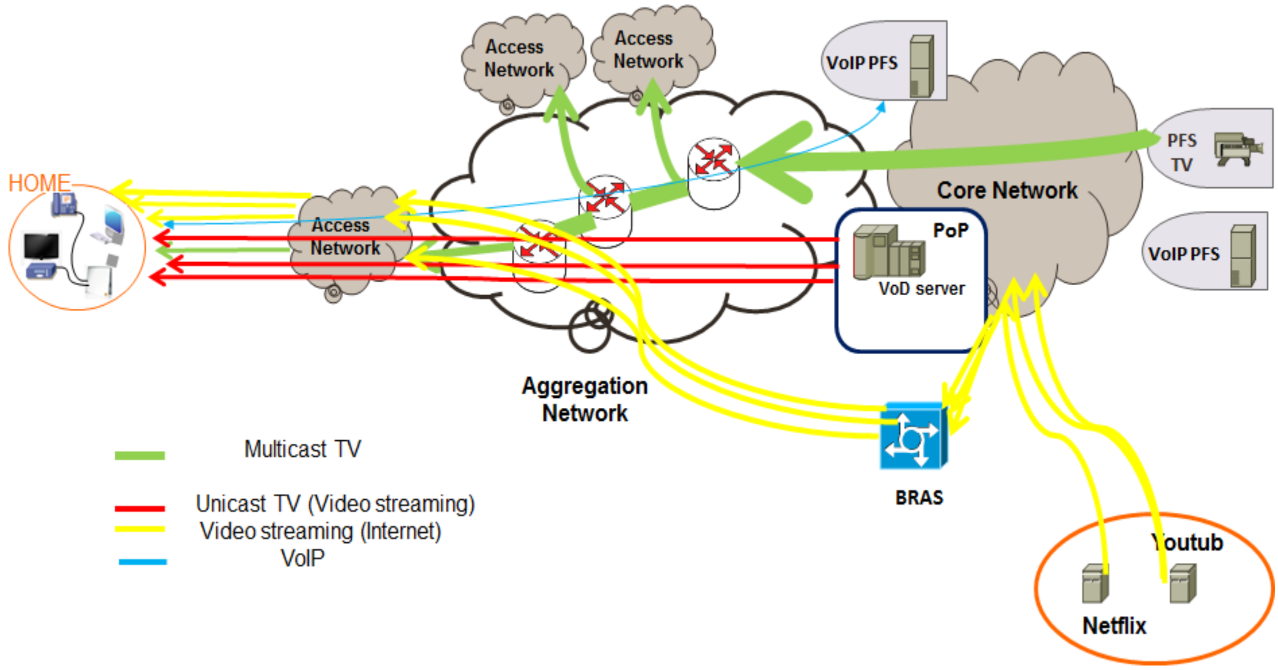


Figure 2: Architecture of Fixed Network.

The traffic generated by the multicast TV is modest and stable in the time. The contribution of multicast video in the network is not important (in terms of volume).

Other video services such as video streaming controlled by the operator are routed using unicast transport. This traffic is routed from the PoPs (Point of Presence) platform of video via the aggregation and access networks (DSLAM/OLT) arriving at the final customers.

Unicast traffic is very different from multicast traffic. Multicast traffic optimizes bandwidth, by broadcasting one flow for a set of customers, whereas in unicast traffic the required content is multiplied by the number of customers requesting it.

The unicast video managed by the operator represents a high contribution in terms of load in different segment of the network (i.e., core / aggregation / access).

A service that currently contributes poorly on the network in terms of load is the VoIP (Voice over IP). VoIP traffic is routed from the source to destination client via the aggregation and access network through a VoIP service platform.

According to our studies, Currently the Internet services (i.e., streaming and download) represent more than 50% of the traffic, and are increasingly loading the networks' resources at different segments.

The video contents (e.g., YouTube, daily motion, Akamai, Netflix) requested by users traverse the core network and BRAS (Broadband Remote Access Server), and then are sent to the end users via the aggregation and access networks.

Table 1 summarizes the qualitative downstream load in different segments of the Fixed Network for the different applications.

		Core Network	**Aggregation Network**	**Access Network**
Downstream	TV multicast	high	low	very low
	Unicast Video managed by operator	high	high	high
	VoIP	very low	very low	very low
	Video (Internet Data)	very high	very high	very high
Upstream	P2P	very high	very high	very high

Table 2: Qualitative load in different segment of Fixed Network.

Upstream traffic

In upstream direction, we observe that P2P generates the majority of traffic, therefore we will only consider the routing of P2P in different segments of the network (core/ aggregation/access).

A P2P service consists in is an exchange of contents between two end-users. Datagrams are routed from the access network across the aggregation, BRAS and finally through the Core Network.

Table 1 reports also table summarizes the qualitative upstream load in different segments of the network for the application P2P in the case of Fixed Network.

2.3.3 Qualitative description of traffic volume in different segments of mobile networks

Downstream

In this section, we chose the same applications in the Fixed Network based on data provided by COMBO partners TID and FT (see section 3.1). It is observed that in Fixed and Mobile Network, the customers use the same kind of applications.

For VoIP, we took the example of a call via Skype, but there are other solutions for VoIP whether via 3G (circuit mode) or via IMS.

The path of VoIP starts from the Internet service platform (e.g,.. Skype, Viber) through the core of the Mobile Network (PDN-GW and S-GW) then across the access network, arriving at the final customer. This path is valid in both directions of transmission (downlink and uplink).

VoIP is more and more used by the majority of users, but it currently represents a small contribution to the total load of the network (access and core network).

Video streaming in unicast mode (Internet, e.g.,. YouTube) represents the application that generates most traffic in downstream and contributes strongly in the load of different segments of the Mobile Network. Video streaming in unicast is sent from video head-end platform while crossing the Mobile Core Network (PDN-GW and S-GW) and the access network (E-node), and then the content is delivered to the end user device. Unicast video managed by the operator currently represents a small contribution in terms of load in Mobile Network and most users are likely to prefer to see the video at home with a classical TV, although this situation is changing lately.

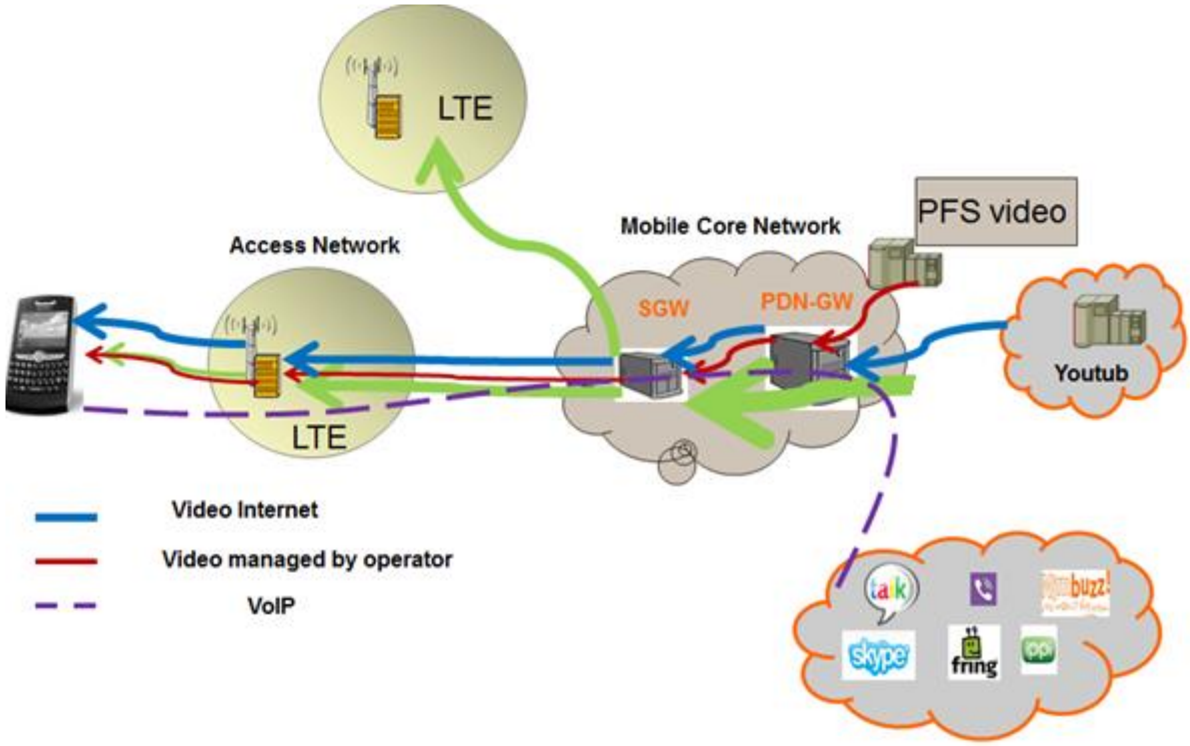


Figure 3: Architecture of Mobile Network (LTE).

Table 2 summarizes the qualitative load in different network segments for different services / applications in the case of Mobile Network.

Downstream	Core Network	Aggregation NW	Access Network
Unicast Video managed by operator	high	high	high
VoIP	very low	very low	very low
Video(Internet data)	very high	very high	Very high

Table 2: Qualitative load in different segment of Mobile Network.

Upstream

In upstream we find that web browsing is always the dominant application in terms of volume generated by all customers (Web represents 59% of total volume generated by Orange mobile customers and 47% of total volume generated by Telefónica mobile customers). But the total volume generated in upstream is very little compared to down-stream volume.

2.3.4 Qualitative description of traffic volume in different segment of FMC use cases

Taking the same applications considered used in upstream and downstream of Fixed and Mobile Networks (as discussed in previous sections), we now consider the path of each application in the upstream and downstream and to in what extent changes may occur in the different segments of an FMC network.

TV multicast is routed from the TV service provider to the end customer via the IP backbone, through the aggregation and finally the converged access (fixed/mobile) networks.

Generally, IPTV multicast traffic is viewed through a fixed access. The contribution of TV multicast traffic remains modest and stable over time.

VoIP is an application that requires little resource (flow) and is sent from the IP backbone through the aggregation and access network converged (in the use case UC06) and/or to fixed (e.g. Wi-Fi for Skype) or mobile access networks (in the use case UC04).

Streaming video managed by the operator is routed from the head-end, via the mobile IP edge and then the aggregation converged network of the mobile access network.

Video streaming in unicast (Internet) represents the application that generates most of the traffic in downstream and contributes strongly in the loading of different segments of FMC networks. Video streaming unicast (Internet) is sent from video servers (or CDNs) while crossing the core and access/aggregation converged network.

2.4 What is expected to change in Fixed Mobile Convergence situation?

2.4.1 General considerations

In the core and aggregation network the primary physical medium is optical fibre. Traffic on specific optical lines may depend on load balancing and resilience policies employed in the core network. On the other hand, traffic characteristics within the access segment is determined more by the individual user activities.

In the radio access network the traffic characteristics on the radio links are affected among others by subscriber mobility. Interface throughput also depends on radio channel characteristics, such as multipath interference and the effects of weather.

Different types of services are preferred by the users on the radio interface and on the fixed line (see the previous section). Service cost and available bandwidth also play a role in user preferences. Energy consumption of a channel will eventually be represented by the price of offered services.

The chosen user devices also affect the types of services used and thus the traffic mix. Some devices are better suited to specific applications. Factors here include processing power, display size, available peripherals, just to mention a few.

User behaviour may also depend on interface utilization. In an example situation, the user starts a large download - just to find that the connection is slow; the user cancels the transfer and postpones downloading activities to a later time. In another example, the user starts a VoIP audio and video session with a colleague; when the network becomes slow, they turn off video and use audio only.

We assume that users dynamically adapt their behaviour to the network capabilities: when there is much bandwidth, downloading becomes easy. When delay variation is high, audio and video conferencing tends to be frustrating and subscribers tend to refrain from using the service.

A further factor is the time available for a particular type of user activity. Specifically, a user will not start a lengthy software update if his/her available time is restricted and the procedure is unlikely to succeed within the available time frame. This also suggests that large data transfers would more likely be initiated from a fixed station rather than a mobile device.

Requirements on service availability can change dramatically when users' traffic patterns change by starting a new service. By introducing e.g. an attractive movie rental service affects the traffic characteristics within that network segment.

A considerable amount of network traffic is not directly attributed to subscriber activities; rather they are results of autonomous machine operations, such as software updates, relayed VoIP traffic, torrent upload, virus activities, sending/receiving spam, usage statistics reporting, advertisement dispatching, etc.

The users' requirements (Quality of Experience) are different for the different services. Channel characteristics should be controlled end-to-end, not only node-to-node. This can be maintained through watching the "bearer capability" parameter requested for the service/session.

Convergent networks will require functionalities from both mobile and fixed services. Mobile operators have an advantage in this transition, because they can extend their portfolio with fixed network services with less investment. On the other hand, fixed network providers would need a considerable effort if they wanted to enter the mobile operator and service provider market. This could result in today's fixed network service providers gradually falling out of business. Still, they can offer network services with high SLA-requirements, as well as infrastructure rental, or outsourcing.

2.4.2 Expectation on load changes

The load in the different segments of the network is completely different. We can imagine that the load must be important in the converged aggregation network (use case UC04) and reduced in fixed and mobile access networks which are separated. Whereas, in the use case UC06, we can have a very significant traffic load in the converged access and aggregation network which represents a single entity.

At busy hours we can imagine that even home users who prefer to connect their devices to Wi-Fi [1] or fixed access can also use their mobile access to download streaming movies or file sharing P2P, whenever the Wi-Fi connection and fixed access become overloaded.

The following tables summarize the total load in different segments of FMC networks.

Total traffic load	Fixed Access Network	Mobile Access Network	Converged aggregation Network	Core Network
Use case UC04	low	low	very high	Very high

Table 3: Qualitative load in different segments of FMC Network (use case UC04)

Traffic load	Converged Access & aggregation Network	Core Network
Use case UC06	very high	Very high

Table 4: Qualitative load in different segments of FMC Network (use case UC06)

3 CURRENT TRAFFIC DEMANDS AND FORECAST

3.1 Analysis of Current Traffic Demands

This section describes and analyses a real Internet traffic data set provided by two major European operators (Orange and Telefónica) [2][3]. This analysis enables allows to network or service operators to improve their knowledge on the current traffic and its possible evolution. The limited set of managed services, like IPTV, enable to enables the use well controlled rules for network dimensioning, this which is not the case for public Internet originated traffic. Indeed, due to the growing success of services delivered by Internet players. Which leads to pay significant attention to the customer's usage to make accurate forecasts, avoid future network congestion or propose new network architectures [4]. This present section provides insight on the applications generating the major part of the traffic, describes used cased for adopting fixed mobile convergence solutions communication services and compares these results with the study produced by CISCO VNI and SANDVINE [5].

3.1.1 Traffic probes

In an operational network, traffic measurements are mainly performed according to two methods. The first one is an active method based on the observation of some specific test flows of IP packets intentionally sent in the network. The second one is a passive method in which all the transmitted IP traffic is captured and analysed in real time. For this analysis, traffic probes are passive (e.g., using port mirroring as done in Telefónica network) and allow a detailed analysis of the traffic. Ina the fixed network of each network operator (i.e., Orange and Telefónica), traffic probes are located between the aggregation and core network, while in segment whereas within the mobile networks, traffic probes are located between Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN).

The probes capture all packets from customers: first they identify which customer is responsible for each packet and second identify which application is used by the customers using a Deep Packet Inspection (DPI) engine. In the case of Orange-FT traffic probe, the DPI engine is able to recognize traffic from about 500 applications which are further aggregated in about 20 application categories (e.g., P2P, streaming, file downloading, etc.). In the case of Telefónica, the traffic probes can identify more than 700 protocols and multiple applications categories.

For Orange fixed network, the average number of residential customers under measurement was approximately 2500 for FTTH and 5000 for ADSL. Concerning Telefónica fixed network (ADSL and VDSL), the average number of customers was approximately 100 000. On the other hand, in the Orange mobile network, Internet traffic observation is made on almost all Orange mobile customers. In Telefonica's mobile network, the traffic probes captured the traffic generated by approximately 1 500 000 customers.

3.1.2 Downstream and upstream profile analysis for fixed customers

Upstream profiles for xDSL and FTTH customers

The traffic measurement allows the operators to access a profile analysis over 24 hours. This section shows Internet data profiles in the case of xDSL and FTTH customers.

Thanks to this analysis, it is possible to clearly define the notion of "busy hour". Internet activity knowledge delivers the opportunity to schedule maintenance actions and firmware updates (e.g., customer devices) during low activity periods.

We have measured the upstream Internet traffic volume for FTTH/ADSL Orange customers between the 7th and 13rd of October 2013. We have also done the measurements for xDSL Telefónica customers between the 23rd and 29th of March 2013.

Figure 4 shows the normalized traffic volume, in relation with the maximum value, averaged over a week for each operator. For Orange customers, the upstream normalized volume is given hour per hour and every 3 hours for Telefónica ones. The volumes generated by Orange customers (FTTH and ADSL) achieve a maximum during the time period [19–22] which represents a "busy hour" period. These measurements show there is no real difference in the upstream Internet profile according to the day of the week. In the case of Telefónica network, the volumes reach their maximum between 6pm and 9pm. Figure 4 also shows a shift of one hour between Orange and Telefónica, whatever the day of the week.

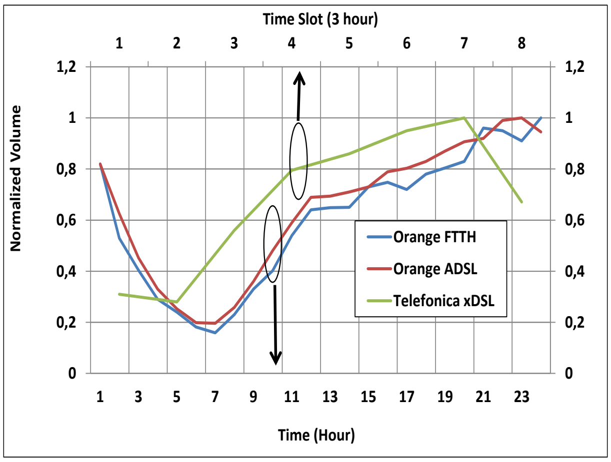


Figure 4: Daily average upstream Internet profile for fixed Orange / Telefónica customers, observation performed on October and March 2013 respectively.

Downstream profiles for xDSL and FTTH customers

Figure 5 summarizes the daily average downstream Internet profiles for FTTH / ADSL Orange and xDSL Telefónica customers.

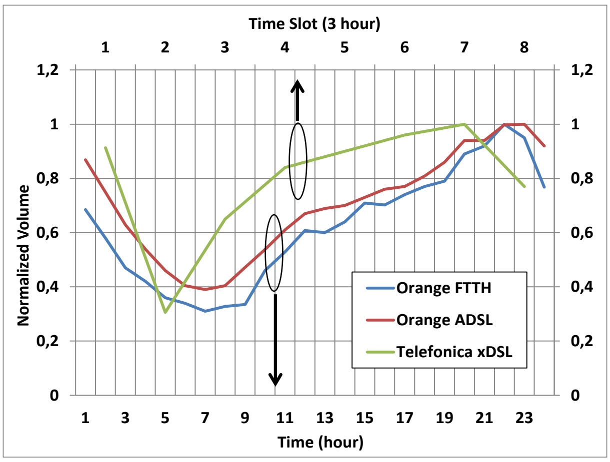


Figure 5: Daily average downstream Internet for fixed Orange / Telefónica customers, observation performed on October and March 2013 respectively.

According to our observation over a week, the average downstream Internet volume generated by Orange FTTH / ADSL customers reaches a maximum during the time period [19-23]. In the case of xDSL Telefónica customers, the maximum of traffic is observed during the time period [18–21]. These results do not depict a real difference between Orange and Telefónica customers: the common “busy hour” period [19–21] can be defined (see Figure 5). In addition, the profile shape remains the same regardless the day of the week and the involved technology. Figure 5 illustrates a classical fixed European residential customer traffic profile as the traffic load increases during the day and reaches a maximum after the working day when people are at home after 19:00 generally.

3.1.3 Downstream and upstream profile analysis for mobile customers

Upstream profiles for mobile customers

Figure 6 is representing the daily average upstream Internet traffic profiles for both Orange and Telefónica mobile customers between the 28th of November 2013 and the 4th of December 2013 for Orange and between the 23rd and 29th of March 2013 for Telefónica. We can see that, the traffic generated by Orange mobile customers during the week-day reached a maximum during two time slots. The first one is during the time period [12-14] and a second time slot is during the time period [17-19pm]. The first one represents the lunch break. In this time period, the customers use their mobile to watch a video streaming, web browsing, chatting with friends and checking e-mails. The second time slot of busy hours occurs at the end of the workday (e.g., in the public transport). Week-day and week-end profiles of Orange customers are quite different. During the week-end, the daily average upstream volume achieves a maximum during the time period 6pm-7pm. Week-end customers’ profile is flatter than the week-day one as customers connections are more distributed over this time period. Regardless the week day, Figure 6 shows that the average traffic volume generated by mobile Telefónica customers reaches the maximum during the period [19-21].

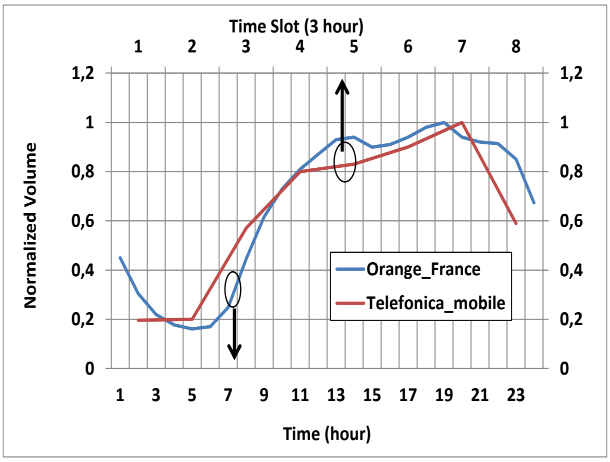


Figure 6: Daily average upstream Internet profile for Orange and Telefónica mobile customers, observation performed for a week.

Downstream profiles for mobile customers

Figure 7 shows the daily average downstream profile of Orange and Telefónica customers in mobile network. In Orange mobile network, the volume generated by all customers reaches a first maximum during the lunch period [12pm-2pm] and a second one during the period [7pm-11pm]. The traffic shape is similar whatever the day of the week.

Figure 7 indicates that the daily average volume generated by Telefónica customers achieves a maximum during the time slot [6pm-9pm]. Also, we can easily notice that the time slot [7pm – 9pm] is a common “busy hour” period for Orange and Telefónica. The results also depict that there is a difference in the downstream traffic shape between Orange and Telefónica as the second maximum found in Orange mobile network does not appear on Telefónica mobile one. Volume generated by the upstream mobile traffic represents the tenth of the volume generated in the downstream. This asymmetry is mainly due to the usage mode difference. Indeed, in upstream, mobile customers mainly use web browsing applications which generate a low traffic volume. In downstream, mobile customers mainly use video streaming which generates more traffic than web browsing.

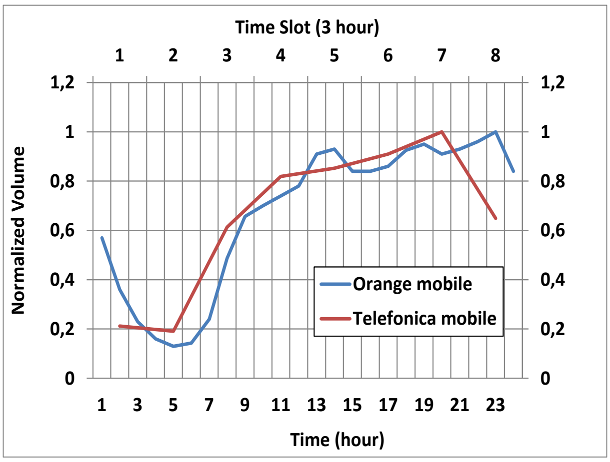


Figure 7: Daily average downstream Internet profile for Orange and Telefónica customers, observation performed for a week.

We can first conclude that, in upstream, a fixed profile is more flat than the mobile profile because the mode of use of wireless and wireline customers is different. As mobile customers are usually at home on evenings and largely use their mobile through the Wi-Fi access of their ADSL / FTTH Home Gateway, there is a rather limited traffic on evenings on the mobile network we do not observe on fixed ones.

3.1.4 Identification of the most bandwidth-demanding applications

In this section, a specific focus is performed on Internet applications generating most of the traffic in upstream and downstream. In order to identify communication services of residential customers, it is necessary to define some categories of applications like P2P. So, P2P category gathers applications like eDonkey, eMule, Bit Torrent, etc. file downloading applications include progressive media (video, audio…) downloading, basic file downloading and software updates. Streaming is used by many applications and allows visualizing for example a video or listening a piece of music without the need to store it on the end user device. With file downloading method, one content must be completely downloaded prior to be played.

Major services on fixed residential access networks

Table 5 depicts the composition of downstream/upstream Internet traffic for ADSL and FTTH Orange customers as well as xDSL Telefónica customers. These applications generate most of the total traffic (either in upstream or in downstream).

Downstream Internet traffic composition				
	Video Streaming	**P2P**	**Web**	**Downloading**
Orange FTTH	36%	16%	16%	26%

Orange ADSL	26%	12%	18%	21%
Telefónica xDSL	36%	20%	15%	3%
Upstream Internet traffic composition				
	P2P	**Web**		
Orange FTTH	78%	5%		
Orange ADSL	48%	18%		
Telefónica xDSL	50%	18%		

Table 5: Downstream and upstream Internet traffic for xDSL and FTTH customers.

Whatever the operator and the access network (i.e., FTTH, ADSL, xDSL), P2P is the application which mainly constitutes the upstream traffic. Thanks to massive FTTH deployment, new greedy applications (such as cloud storage, virtualization, etc.) could appear to the detriment of P2P applications which could revise network aggregation capacity and architecture [4].

In downstream, video streaming is now the main application generating most of the traffic (up to 36% for Orange FTTH and Telefónica customers). File downloading (resp. P2P) is the second family of applications generating most of the traffic in both Orange (resp. Telefónica) network. Massive use of streaming video may lead to congest peering links which can be mitigated by the capacity increase of these links to avoid bottleneck.

Major services of mobile customers

Table 6 describes the main applications composing the total Internet traffic in both Orange and Telefónica mobile networks. Video streaming is the application generating the main amount of the traffic in downstream within Orange networks competing with web browsing, which remains as the application generating the main part of the traffic (up to 47%) within the Telefónica mobile network. In total, video streaming and web browsing represent 73% (resp. 80%) of the total downstream traffic in the case of Orange (resp. Telefónica) mobile network. In the upstream, whatever the operator, web browsing is the application generating most of the traffic, up to 59% (resp. 47%) in the context of Orange (resp. Telefónica) mobile network. P2P applications are blocked in Orange mobile networks so the traffic remains lower than 0.5% and reaches 6% of the total upstream traffic in the case of Telefónica mobile network where P2P applications are allowed.

These results are consistent with a 2013 Sandvine report [5]. In Europe, P2P generates about 50% of upstream total traffic in fixed network. In the downstream direction, video streaming represents about 30% of total traffic followed by web browsing and P2P. In mobile network, network usage is related to web browsing and represents about 30% of total traffic generated in the upstream direction. In downstream direction, streaming video and web browsing represent together about 50% of total volume.

Downstream Internet traffic composition		
	Video streaming	**Web**
Orange network	39%	34%
Telefónica network	33%	47%
Upstream Internet traffic composition		
	Web	**P2P**
Orange network	59%	< 0.5%
Telefónica network	47%	6%

Table 6: Downstream and upstream internet traffic for mobile customers.

3.1.5 Towards a fixed and mobile convergence services

Some similarities between fixed and mobile networks

The previous sub-sections showed that video streaming is the downstream application used in either fixed or mobile networks for both operators. Web browsing is the other downstream application in mobile networks and is the third application category within fixed networks. Upstream Internet mobile traffic is mainly based on web browsing, while P2P application is the preponderant application in fixed networks. In mobile Orange network, P2P application is forbidden, but this could change with fixed/LTE plans. In the latter, the Internet access is possible but will remain limited since the data is capped within the mobile network. In the upstream direction, the usage mode of fixed and mobile customers is different for both operators with a little similarity which is mainly the use of Web application.

Evolution of traffic volume generated by mobile customers using their fixed access

Increasingly mobile customers tend to use their Wi-Fi interface embedded on their mobile devices. At home, this impacts on the traffic data profiles of the fixed networks. The Compound Annual Growth Rate (CAGR) of downstream average traffic volume generated by Orange mobile customers using Wi-Fi access between 2011 and 2013 represents an increase of 56%. In the same time, the number of customers using their mobile devices in fixed access networks increased by 38%. Our results confirm the achievements reported by CISCO VNI, where 33% of total mobile data traffic was offloaded onto the fixed network through either Wi-Fi or Femtocell in 2012.

The CAGR increase is due to both video streaming and file downloading applications. In the case of mobile networks, file downloading generally represents the software update of mobile operating systems. As such devices require more bandwidth for updating their operating system; customers prefer to use fixed Wi-Fi access. Video streaming is one of the main downstream applications in both fixed and mobile networks and represents 50-70 % of the traffic generated on Wi-Fi home access points.

3.1.6 Main Achievements and Conclusion

In fixed and mobile networks, the time slot [19-21] is a common “busy hour” period for Orange and Telefónica customers and traffic profiles are quite similar. Even though probe measurements in Orange mobile network report a second busy hour, [12-14], which represents a lunch break.

On ADSL/FTTH Orange and xDSL Telefónica network, P2P applications generate most of the upstream traffic. It generates 48% (or more) of the traffic in the case of xDSL and more than 70% in the case of FTTH.

In downstream, video streaming, file downloading, web browsing and P2P applications represent the applications generating more than 80% of the total downstream volume in FTTH and xDSL.

Some common services appear in the same time on fixed and mobile networks. Video streaming is commonly used at the same time on both fixed and mobile networks which becomes as a convergent application. As mentioned above, mobile customers leverage their Wi-Fi home access points to get Internet access which will impact fixed networks in the future.

3.2 Forecast of traffic trends

3.2.1 General trends

In the next years, all reports claim that mobile data traffic will grow faster than fixed data traffic. Indeed, during 2013 it was forecasted [15] that mobile data traffic will continue to double each year. Anyway, fixed data traffic volume will remain higher than mobile data traffic volume in absolute terms in all the foreseen periods (see Figures 8, 9) by at least one order of magnitude.

The rapid growth of the mobile traffic is mainly due to the rise of the smartphone traffic as a consequence of the increased number of smartphone sales and subscriptions. The total number of mobile subscriptions is expected to reach about 9 billion in 2017 [15]. In addition, the traffic generated by a single subscriber is forecasted to become 5 times larger in the next 6 years (Figure 10).

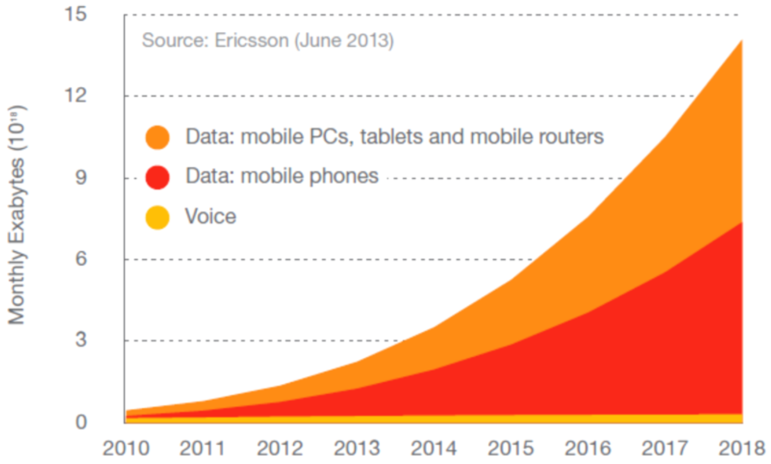


Figure 8: Global Mobile Traffic 2012 – 2018. [15]

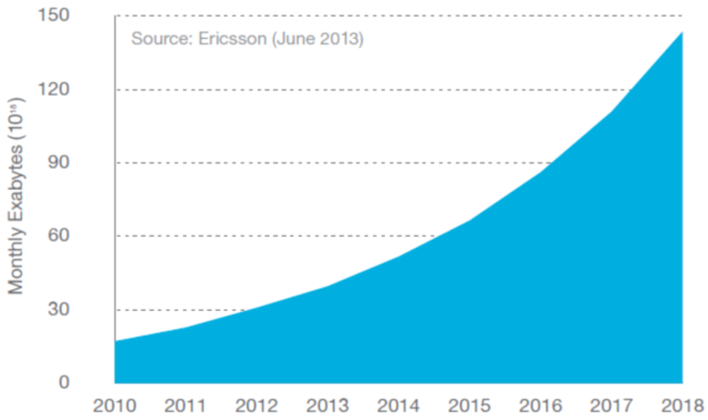


Figure 9: Global Fixed Traffic 2012 – 2018. [15]

According to the Cisco VNI Global Mobile Data Forecast 2012–2017 [7], the future outlook for mobile data traffic forecasts that by 2017 the global mobile data traffic will reach 11.2 exabytes per month (exa = 10^{18}),(134 exabytes annually.). Smartphones will represent the 68%of the total mobile data traffic in 2017, compared to 44% in 2012; 4G connections will be 10% of the total mobile connections in 2017, and 45% of the mobile data traffic.

It is expected that data traffic will be always more fairly split between mobile phones on the one hand, and tablets, mobile routers and mobile PCs on the other hand. Moreover, the increase of the utilization of the mobile data network is driven by the growth in the amount of content (i.e., applications and services) available and by the improved network speed which follows the development of HSPA and LTE.

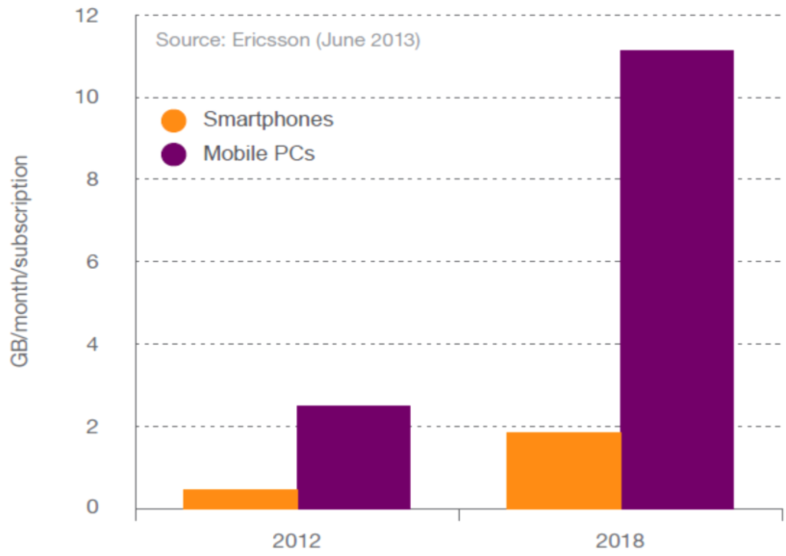


Figure 10: Mobile Device Traffic per month and per subscription forecasted in 2012–2018. [15]

A forecast study [15] related to the population covered by different mobile systems in the future years states that: it is expected that more than 90% of the global population will be covered at least by GSM/EDGE technology while 85% will be reached also by WCDMA/HSPA transmissions. Finally, more than 60% of the world population will be able to connect to the mobile data network through LTE systems. Such measurements (Figure 12) certify that in 2018 the overall mobile data traffic will grow due to a higher number of users use utilizing broadband mobile technologies.

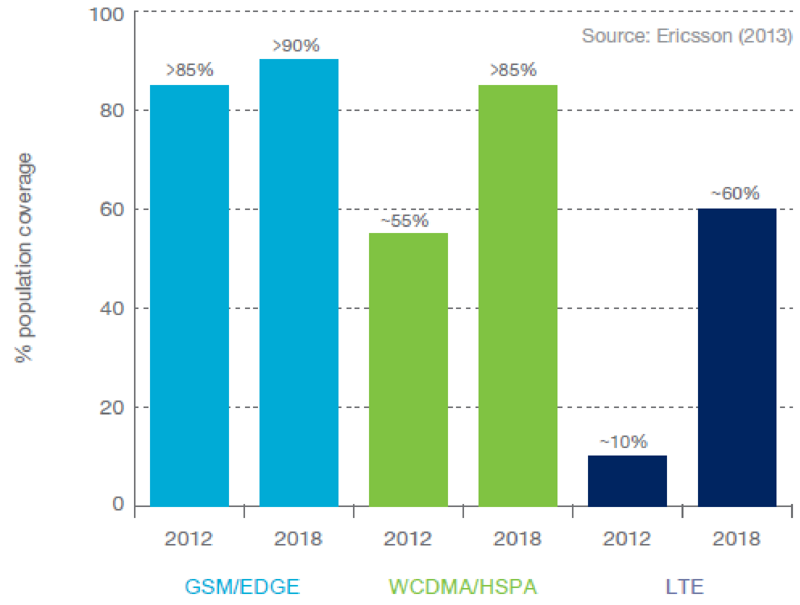


Figure 11: Forecast of population coverage divided by mobile access technology. [15]

Considering the traffic growth on per region basis, Figure 12 shows that in Asia-Pacific and Middle East the growth of the number of mobile subscriptions will be the strongest among all the regions [15]. On the other hand, the figure also shows that the market in Europe is saturated. Therefore, more data traffic will not come from more users, but from other end devices which allow more/different use, and thus, require more bandwidth.

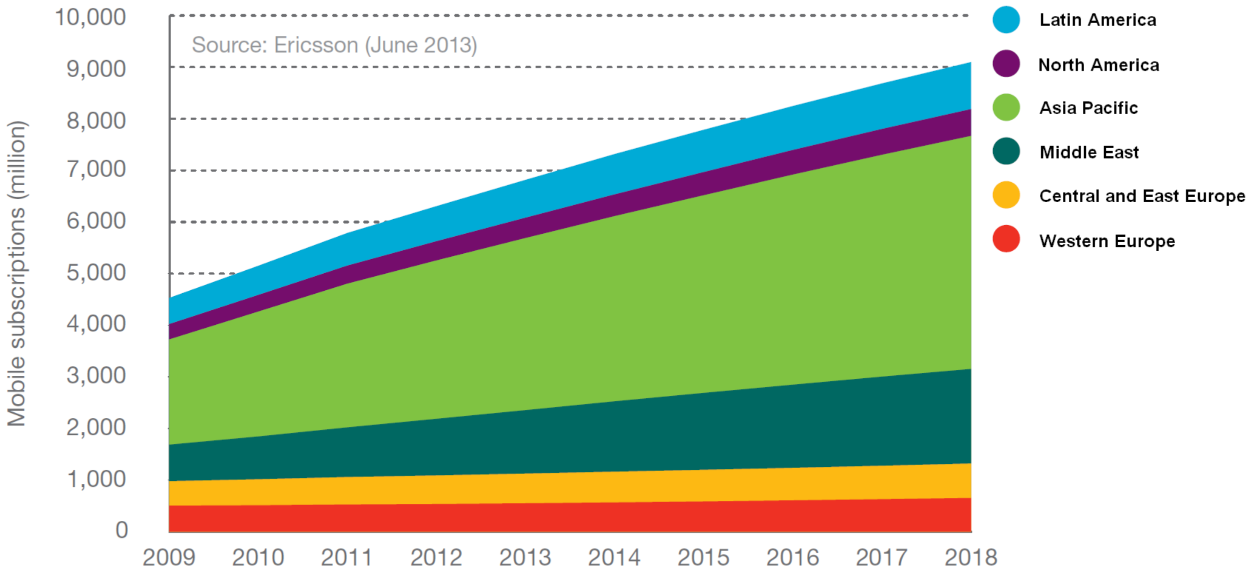


Figure 12: Mobile subscriptions by region. [15]

Now, the main question that is beginning to be discussed in the fora is: what will happen in a longer timeframe, say 10-20 years? A reasonable prediction shared in the scientific community is the following.

Today:

- about 4 billion people generate voice traffic on the order of about 20 million Bytes/month/user;
- about 20 million people generate data traffic on the order of about 4 billion Bytes/month/user.

Tomorrow:

- most probably 4 billion people will generate a global mix of data traffic, including any type of service, on the order of about 20 billion Bytes/month/user.

This could be seen as a huge chance, but also as a huge threat. Therefore, the next question that could be asked is: how could we achieve a capacity increase by a factor $\times$1000? A possible answer, taking into consideration the mobile broadband evolution, could be given by the following considerations:

- HSPA and LTE technologies can bring a capacity increase factor $\times$10 with respect to Wideband CDMA (WCDMA).
- Spectral efficiency improvements (from 1 to 2 bit/s/Hz) give a capacity increase factor $\times$2.
- More allocated spectrum gives a capacity increase factor $\times$5.

- A smaller cell radius (about 1/3 of that currently in use) gives a capacity increase factor ×10.

Overall, these factors multiplied together give a capacity increase factor around ×1000, which could match the foreseen increase of overall traffic volume..

3.2.2 Broadband service trends to 2020

The past decade has witnessed innovations which changed the global communication opportunities and people's way of life. As the bandwidths, contents and services grew, the volume of the broadband communication increased significantly and will continue to increase due to the constant information communication and entertainment flowing around the globe [8]. An important question is therefore: How much traffic is there on the Internet, and how fast is it growing? Network technologies, architectures, and business models all depend upon and influence the answer. However, there is very little solid data about what is happening on the network, and many conflicting estimates.

Internet traffic is the flow of data across the Internet. What stimulates its growth is a complex set of feedback loops operating on different time scales, involving human adoption of new services and improvements in processing, storage, and transmission technologies [9]. In spite of the widespread claims of continuing and even accelerating growth rates, Internet traffic growth appears to be decelerating [10]. The most, prominent source of IP traffic forecasts is the Cisco Visual Networking Index (VNI). Since 2006, Cisco provides an annual prediction of the expected IP traffic for 5 years into the future. Ever since it started publishing results in 2006, the CAGR has been dropping constantly with each update of the report.

In its current version, Cisco forecasts indicate that Internet traffic is to grow 27% annually until 2017, reaching 3 times the Internet traffic in 2012 by that time [11]. However, in order to extend those figures further into the future, one has to take into account that the IP traffic is growing at a sub-exponential rate [12]. This is illustrated in Figure 13, where the blue curve shows the annual growth rate (CAGR) of the total IP traffic reported in the Cisco's VNI from 2006 to 2012. The orange curve shows the regression of those values and the projection into the future. The blue curve shows the historical CAGR between years 2006 and 2012. The green curve is the resultant prediction of traffic growth until the year 2020 based on fitted curve.

The IP traffic can be further classified by the connection type (i.e., "Fixed Traffic" and "Mobile Traffic"), by the data type (e.g., Video, P2P, HTTP, Voice) and generated by either consumers or businesses.

The term "Fixed Traffic" denotes IP traffic generated by households, university populations, Internet cafés, corporate IP WAN traffic and traffic generated by traditional commercial TV services. Mobile data includes data and Internet traffic generated by handsets, notebook cards, and mobile broadband gateways.

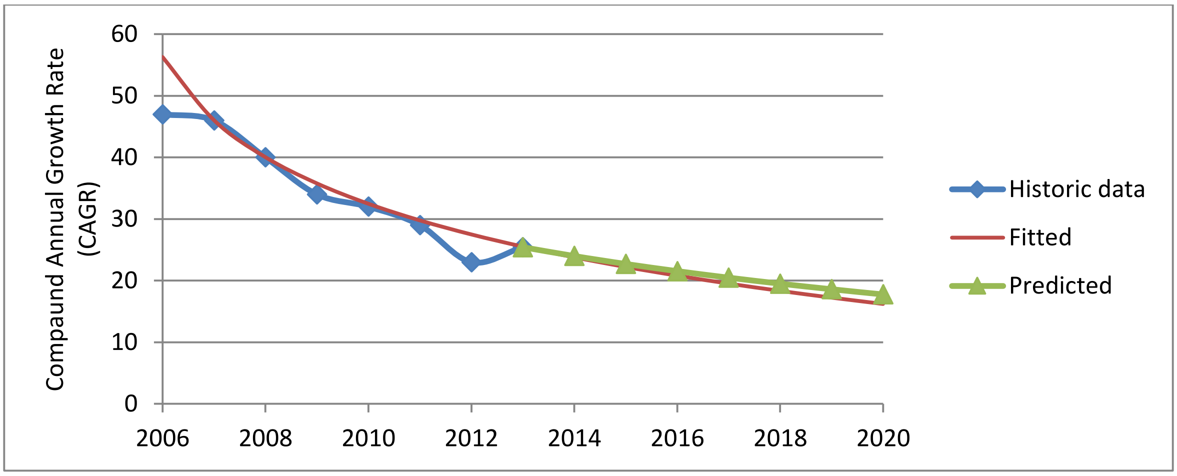


Figure 13: Prediction of the compound annual growth rate for the total IP traffic. [11]

Table 7 gives an overview of the predicted traffic volume. The table lists the results from the seven previous Cisco reports denoted as “Historic Data”. For each of the five categories, the volume and the CAGR are given. One can see that in each category, the CAGR is decreasing. A logarithmic regression of the CAGR based on the historic data is calculated in order to take the sub-exponential growth into account. Then, based on the updated CAGR, the expected traffic is predicted. Cisco report (May 2013) covering 2013-2017 is also shown in the table to enable the comparison of the two approaches [11].

By the end of 2017, global IP traffic will be around 1.4 zettabytes per year (zetta = 10^{21}). This corresponds to around 120 exabytes (EB) per month, 4.0 EB per day [11]. If that growth continues, by 2020 the Internet traffic can be 5 times what it was in 2012.

	Total Traffic		**Fixed**		**Mobile**		**Consumer**		**Business**	
	Volume	CAGR	Volume	CAGR	Volume	CAGR	Volume	CAGR	Volume	CAGR
Historic Data										
2006	4,2	47	4,2	47	0,03	130	2,6	53	1,6	29
2007	6,6	46	6,6	46	0,03	125	4,4	49	2,2	35
2008	10,1	40	10,1	40	0,03	131	7	42	3,1	32
2009	14,7	34	14,6	33	0,09	108	11,6	37	3,1	21
2010	20,2	32	19,9	30	0,24	92	16,2	34	3,9	22
2011	30,7	29	30,1	26	0,6	78	25,8	30	4,9	22

2012	43,5	23	42,6	21	0,9	66	35	23	8,5	21
Forecast										
2013	54,0	25,4	52,4	22,9	1,6	76,1	43,8	25,5	10,2	20,8
Cisco	55,5	23	53,9	20,5	1,6	66	35	23	8,5	21
2017	118,3	20,5	105,9	17,4	12,4	62,8	97,1	19,5	21,2	18,3
Cisco	120,6	17	109,4	20,5	11,1	50	98,9	19,7	21,7	18,5
2020	211	17,8	162,2	14,3	48,8	55,5	172,2	16,2	38,8	17,0

Table 7: IP Traffic Growth.

Busy-hour Internet traffic will increase by a factor of 3.5 between 2012 and 2017, while average Internet traffic will increase 2.9-fold. Busy-hour Internet traffic will reach 865Tbit/s in 2017, the equivalent of 720 million people streaming a high-definition video continuously [11]. It is expected that in the USA, the traffic per subscriber will continue to increase with 30% CAGR [13].

Metro traffic will grow nearly twice as fast as long-haul traffic from 2012 to 2017. The higher growth in metro networks is due in part to the increasingly significant role of content delivery networks, which bypass long-haul links and deliver traffic to metro and regional backbones. Total metro traffic already exceeds long-haul traffic. In 2012, total metro traffic was 1.8 times higher than long-haul traffic, and by 2017, metro traffic will be 2.4 times higher than long-haul as shown in Figure 14 [11]. To project this trend to year 2020, where amount of access and metro traffic is expected as 202 EB/month 5 folding year 2012, long-haul traffic will be 80 EB/month [17].

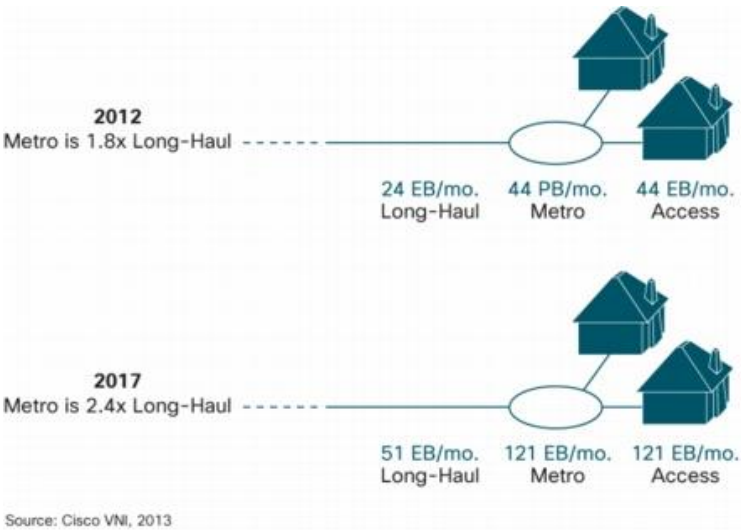


Figure 14: Metro versus Long-Haul Traffic Topology, 2012 and 2017. [11]

In the following section, the traffic predictions are broken down based on the connection type, the data type and the geographical region.

Internet traffic based on connection types

As mentioned above, Internet traffic can be categorized by the connection type: be originated from fixed devices (i.e., from PCs, TVs, tablets and other devices connected through digital subscriber line, cable, fibre, etc.) or mobile devices connected through 3G/4G networks. Based on the sub-exponential growth function, traffic (i.e., total, fixed and mobile) predictions are made until 2020 in Figure 15 [6]. Thinking of connection types, fixed access network can be further divided into wired and wireless connection types. Figure 16 shows the traffic forecast based on the connection type (i.e., wired, wi-fi, and mobile [6]). Based on the forecast for 2020 it is expected that the amount of traffic generated by wired access fixed networks will be nearly the same as the one generated by mobile access networks, which corresponds to half of the amount of Wi-Fi accessed fixed networks.

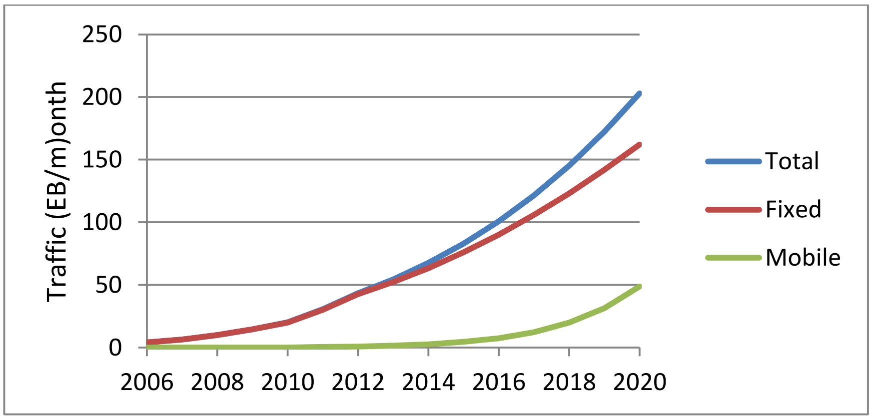


Figure 15: Predictions for the Fixed and Mobile Traffic [6]

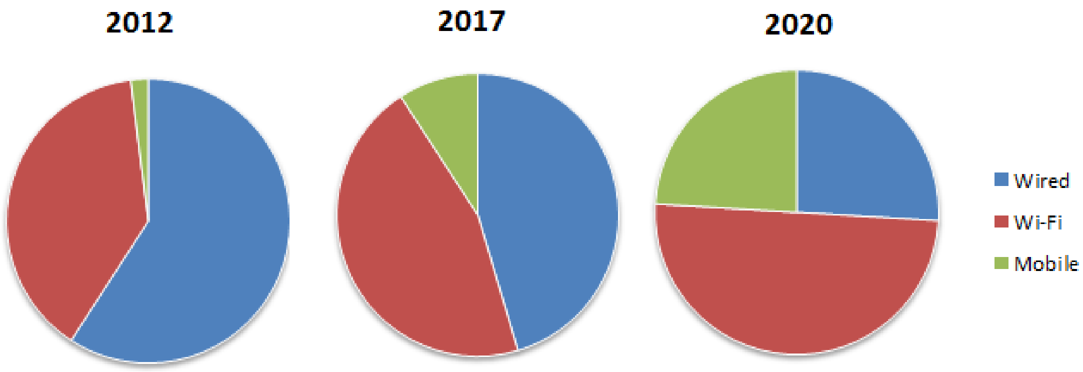


Figure 16: Traffic based on connection type. [6]

Fixed Internet traffic

Fixed Internet traffic includes Wi-Fi access traffic which is generated by mobile devices (e.g., smart phones, tablets, notebooks etc.) which offload their traffic through a fixed connection when available. IP traffic generated by wired connected end-devices comprises approximately 59% of the total IP traffic in 2012. Studies show that by 2014 IP traffic generated by end-devices connected through Wi-Fi access network is expected to surpass the wired access IP traffic. The same study implies that by 2017 wired connection type portion of Internet traffic will constitute only 39% of the total IP traffic [11]. If the forecasted growth rates continue until 2020, wired connection type portion of fixed IP traffic will constitute only 28% of the total IP traffic.

Mobile Internet traffic

Global mobile data traffic is expected to grow three times faster than fixed IP traffic – having an estimated compound annual growth rate (CAGR) of 76% and reaching 12 EB per month by 2017. Hence, the estimated 120 EB per month will include 105 EB of fixed traffic [11]. This indicates that although mobile Internet traffic is increasing rapidly, fixed Internet will continue to comprise most of the consumer Internet traffic. By 2020, mobile traffic will compete with the wired fixed traffic growing to 49 EB per month which is 25% of global traffic.

Consumer and business traffic

Traffic predictions for the consumer- and business IP traffic based on the Cisco VNI with sub-exponential growth are given in Figure 17. In 2012, the business IP traffic volume was about 20% of the total traffic. This will drop to about 15% in 2020 since the consumer traffic is growing at a slightly higher rate.

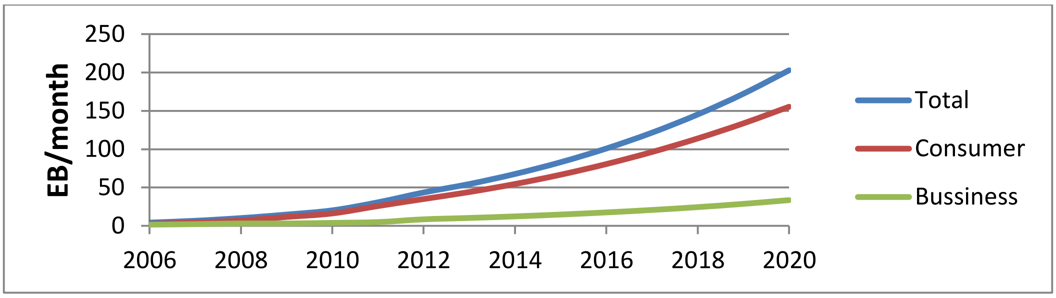


Figure 17: Predictions for the Consumer and Business Traffic.

Consumer Internet traffic based on data types

The sum of all forms of IP video (i.e., Internet video, IP VoD, video files exchanged through file sharing, video-streamed gaming, and videoconferencing) will continue to be in the range of 80-%90% of the total IP traffic. Globally, IP video traffic will account for 73% of traffic in 2017. Taking a more focused definition of Internet video that excludes file sharing and gaming, Internet video will account for 52% of consumer IP traffic in 2017 [11]. Detailed information of consumer Internet traffic is given below and the trends are shown by Figure 18.

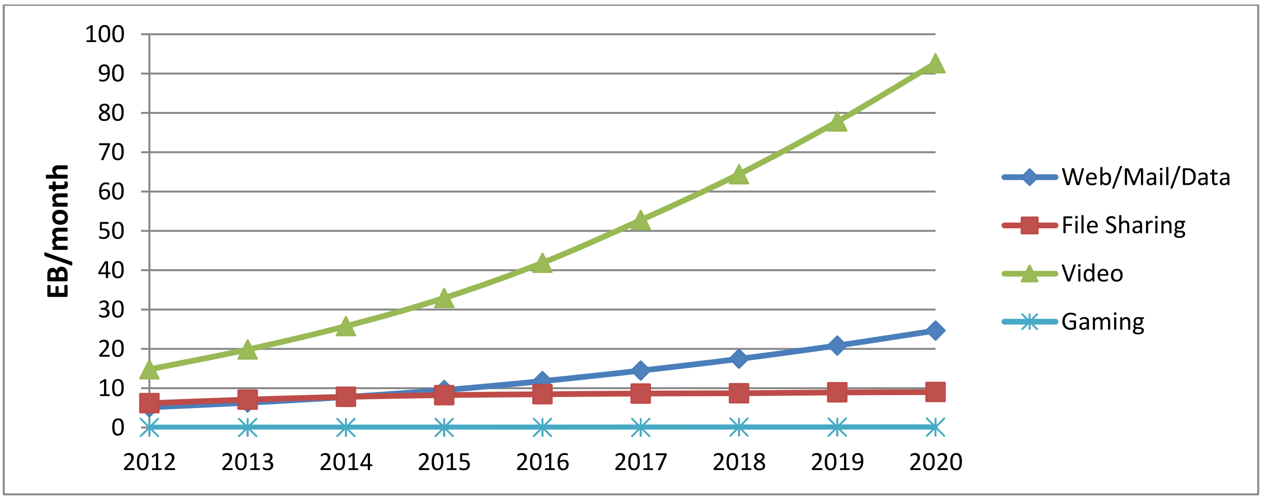


Figure 18: Consumer Internet IP traffic forecast based on data type.

Internet video traffic constituted 38% of the global IP traffic by 2012. It is expected to grow to 52% by 2017 [11]. This statistic does not include video shared via P2P file sharing. Internet video traffic is expected to increase 29% annually. The overall video content from all sources such as Internet, P2P, TV and video on demand, broadly named IP video, will continue to comprise 73% of the global consumer traffic by 2017 [11]. By 2020, it is predicted that Internet Video traffic will reach 90 EB per month which will be 60% of consumer traffic and 45% of global traffic.

File sharing comprised 23% of the consumer Internet traffic in 2012. It is expected to comprise 23% of the total consumer Internet traffic by 2017 [11]. From 2011 to 2016, file sharing traffic is expected to grow with 26% compound annual growth rate (CAGR) [11]. The trend of file sharing is included in Figure 18. In 2013, estimated file sharing consumer Internet traffic was around 7.1 EB/month. It is expected to reach 9.0 EB/month by 2020 [11].

E-Mail, data and web traffic was 19% of the total consumer Internet traffic in 2012.The traffic is expected to have a CAGR of 35% [11]. The trend of web, e-mail and data is included in Figure 18. The monthly web, e-Mail and data traffic is approximated to be 14.4 EB/month by 2017 and 24 EB/month by 2020 while it was 5.1 EB/month in 2012 [11].

In online gaming, there is no need to have a powerful graphical processor but only a low delay and a high bandwidth Internet connection. It is becoming more and more popular each day. Most of the online (or cloud) gaming platforms have been in service since 2011 or 2012 [11]. Therefore they contribute very little to the consumer traffic for now. Specifically, for now cloud gaming traffic constitutes only 0.04% of the total online and offline gaming traffic. However, it has a great potential to become one of the biggest contributors to the Internet traffic. Forecasts imply that online gaming traffic in 2020 will be 5 times higher than in 2012 [11]. Internet traffic based on Online Gaming is expected to have one of the most rapid increases for the next five years ahead, with a CAGR of 52% [11]. The projection of online gaming traffic amount towards the 2020 is growing; although its general volume and relative share is small, as shown in Figure 18.

As can be seen in Figure 18, the consumer Internet traffic is expected to increase in all data types. Among these data types video will be the dominant data type in consumer Internet traffic.

3.3 Drivers for traffic growth

The growth of the mobile traffic is huge throughout the globe and constantly changing our behaviours and our society. Nowadays, almost half of the population of the earth uses mobile communications. In the last 4 years a billion of mobile subscribers were added providing a total amount of current mobile users of 3,2 billions [16]. Anyway, there are many people who have not the possibility to access the Internet yet. Therefore, it is reasonable to think that mobile traffic will continue to grow in the future years.

According to a report of Cisco [17], in 2012 global mobile data traffic grew more than 70% 855 petabytes (PB) a month (peta = 10^{15}). This percentage of growth was different among the regions. The lowest traffic growth was experienced in Western Europe (44% of traffic growth). There are various reasons which explain the low traffic growth rate measured in such area. One of these reasons is the elimination of many unlimited data plans and the consequent introduction of tiered mobile data service plans to price the mobile services, e.g., according to the needs of the users. Moreover, in the Western Europe a slowdown in the increase of the laptops connected to the mobile network was registered. Finally, operators have promoted the offload of the mobile traffic to the fixed network, i.e., onto the Wi-Fi networks.

Conversely, the highest traffic growth rates were recorded in Asia Pacific (+95%), and in Middle East and Africa (+101%). Therefore, such emerging markets are the major contributors to the increase of the mobile traffic. This is mainly due to a decrease of the prices of devices and to the consequent higher possibility to own a mobile technology. Moreover, the development of new technologies which can be used to provide connectivity to rural areas is another reason of the traffic rise in such emerging markets.

It is important to consider that the nature of traffic growth in developed countries and in developing countries is very different [15][16][6]. In fact, Africa and Europe have a similar number of total subscribers, but this number is due to different reasons in the two regions. In Africa the growth of connections (8%) is caused by subscribers addition. Instead, in Europe this growth rate is mainly due to the increase in the number of SIM cards per subscriber. This means that many consumers own several different types of devices and use multiple SIM cards.

In general, the current traffic growth is due to many factors. The spread of wireless devices, the increase of network-attached equipment, the diffusion of video services, the introduction in the market of both public and private cloud delivery models, the faster speed of mobile connections, and finally the decreasing costs of smartphones are the most important drivers of the traffic growth. We can classify such drivers of the traffic growth in main categories:

- increase of the number of connected devices;
- increase of the available bandwidth provided to end-users;
- increase of the number of bandwidth hungry applications and services.

In the following we will discuss in details the main categories listed above which cause the increase of the overall traffic in the networks.

3.3.1 The increasing number of devices

The rise in sales of mobile devices is one of the main drivers of the traffic growth. Recently various devices that incorporate mobile network connectivity, such as smartphones, tablets, e-book readers, TV sets and gaming consoles, appeared in the market.

These new devices offer improved resolution and screen size. Such improvements caused an increase of data consumption per image and foster people to use bandwidth hungry applications, such as video streaming or video calls. Indeed, according to Ericsson [15], the mobile subscriptions are increasing for tablets, smartphones and PCs.

The strongest growth in the last years concerns smartphones sales and subscriptions. At the end of 2012, the total smartphone subscriptions reached 1.2 billion. As a matter of fact, 3.2 billion people out of 7 billion on earth currently own a mobile phone [16].

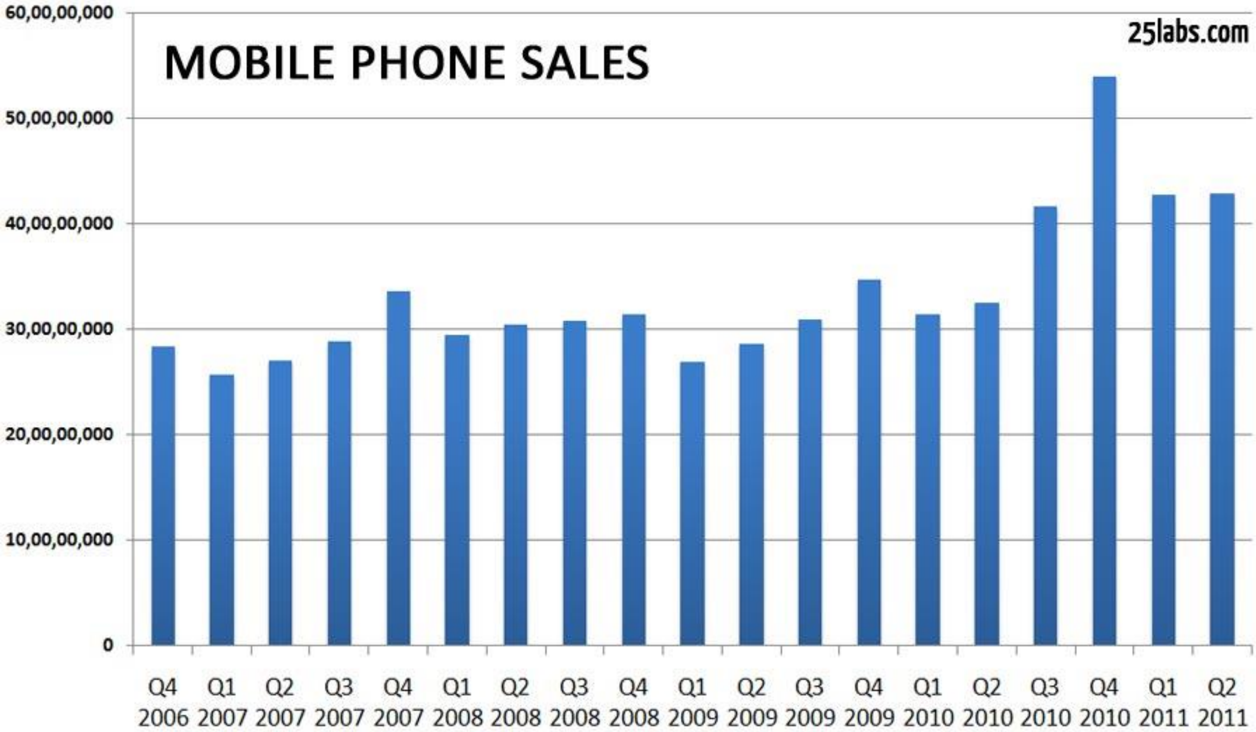


Figure 19: The sales of mobile phones worldwide. [18]

In the future, the number of smartphones will continue to increase while the number of non-smartphones will stop to grow. The number of mobile devices such as tablets, laptops, and machine-to-machine (M2M) modules will continue to grow in the next years [17].

If we consider that in developed markets a smartphone generates much more traffic than basic phones, we realize that such rise in the smartphones sales have a big impact in the traffic growth in recent and future years.

The study provided by Cisco shows [17] a forecast where the contributions to the bandwidth consumption due to different devices are reported, in present and future years. Nowadays smartphones and laptops are the main contributors to the bandwidth consumption. In the next years smartphones will become the devices which will provide the highest amount of bandwidth. A very important outcome of the forecast shown in this figure is that the amount of bytes consumed will continue to grow in the future.

Touch-screen smartphones (launched from around 2007) have been a key driver of mobile data traffic growth [14] – leading to an increase of the mobile Internet usage. By 2009 there were 95 million mobile Internet users in Europe [14]. At the same time, mobile video traffic was growing dramatically. By 2010, YouTube and other Internet services with Flash video have generated the majority of the mobile video traffic. It is predicted that video will account for 66% of mobile data traffic by 2014 [14], [15]. Media rich social network users with mobile Facebook are twice as active as the average user. In 2010, more than 75% of smartphone users accessed social network sites [14], [15].

The evolution of the devices and the expected traffic per device has been evaluated by the UMTS broadband forum in [6]. In 2020, 60% of the devices will consist of smartphones. Connected PCs and laptops (through Dongles which may also be included in the laptop) and tablets (denoted as connected devices in [6]) will make up 15% of the mix (see Figure 20). By 2020, the traffic generated by tablets and 3G/4G connected PCs and notebooks is expected to be similar to the traffic generated in 2012 by a fixed connection on Digital Subscriber Lines or CATV/Cable TV networks [6]. In Europe, in year 2020, it is expected that there will be 1.7 mobile subscriptions per inhabitant [6]. With a population of about 500 million in EU-27, that will lead to 850 million subscriptions.

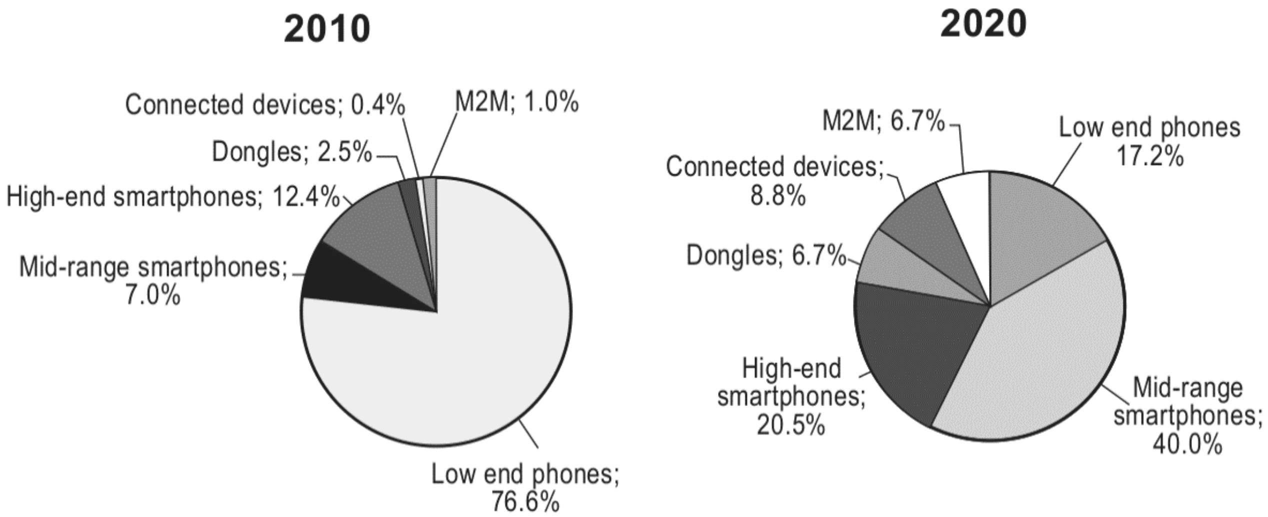


Figure 20: Worldwide mobile device mix in 2010 and 2020. [6]

By 2017, over 30% of the world's population are expected to live in metro and urban areas. These areas represent less than 1% of the Earth's total land area, yet are set to generate around 60% of mobile traffic by 2017 [15]. Broadband mobile population coverage (i.e., WCDMA, HSPA and LTE) will increase from 50% (2012) to about 85% in 2017 [15]. Figure 21 shows projection of mobile traffic generation per access technology per area in 2017.

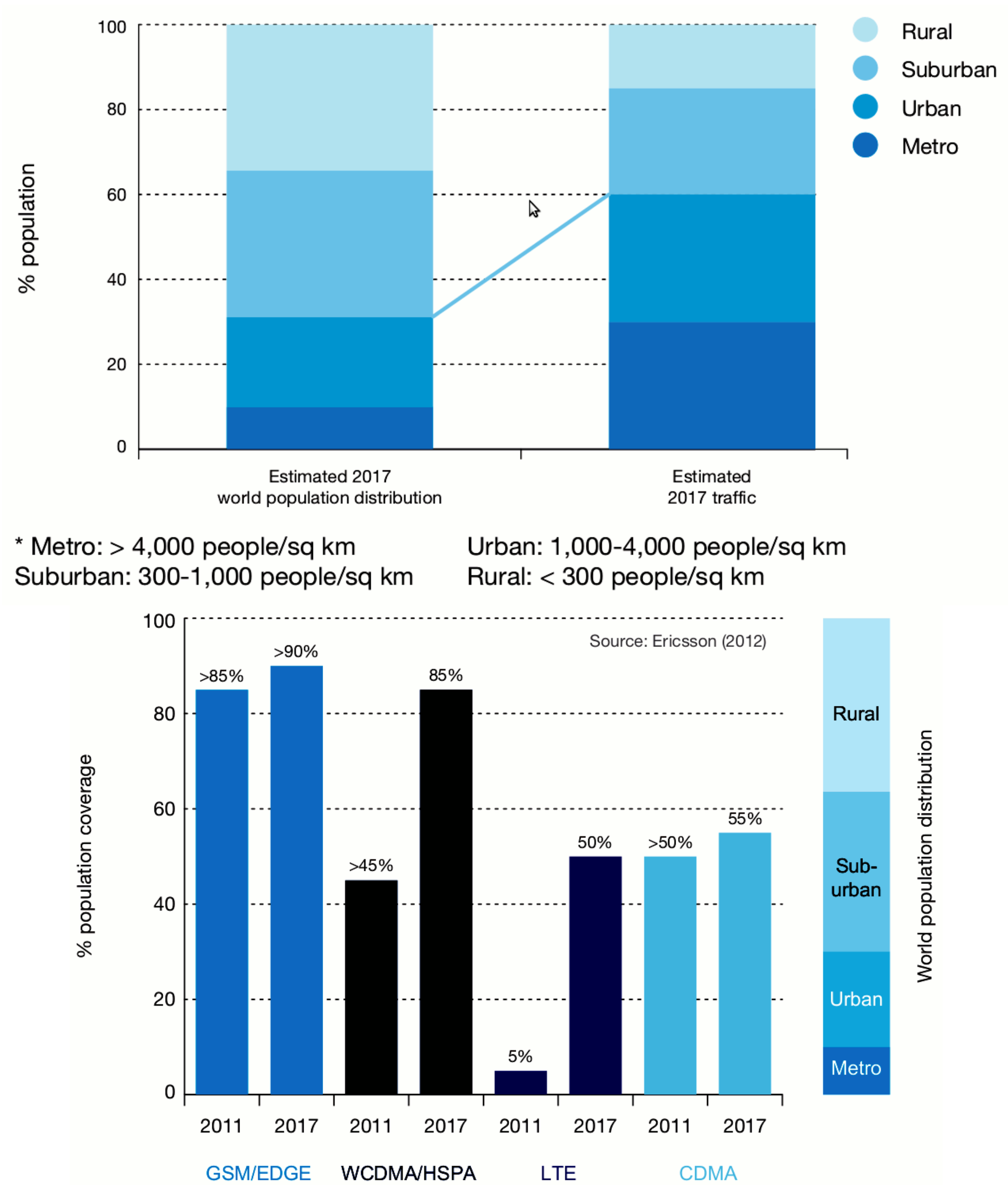


Figure 21: Mobile traffic generation per area in 2017. [15]

3.3.2 The bandwidth domain

Another driver (and enabler) for the traffic growth is the increase of connections numbers and speeds. Indeed, the augmented speed of the connections allows each user to consume more data, encouraging the use of applications which require higher bit rates (e.g., games, video streaming, video calls, and social network applications [17]).

The enhancement of the connections speed is mainly due to the increasing proportion of 4G mobile connections. The 4G connections, which include LTE and mobile WiMAX, providing higher bandwidth, lower latency, and an increased level of security, will promote the generation of a huge amount of mobile traffic [17].

Today a single 4G connection generates 19 times more traffic than another non-4G connection. This is mainly due to the fact that many of the current 4G connections are used for residential broadband access, which has in general higher average usage. Moreover, since the 4G system provides high connection speeds, it encourages the usage of bandwidth hungry applications.

3.3.3 Services and Applications

Several new mobile services and applications have been recently introduced on the market and are finding large consensus among both consumers and enterprise customers. Some of these services and applications generate a huge amount of data traffic. Examples of these applications are Cloud services, content sharing, music and video streaming, video on demand and video telephony. Among these, video streaming is one of the major contributors of bandwidth. Moreover, the number of mobile requests for video content is expected to increase [19] due to both the future enhanced technologies that will be deployed (e.g., LTE and LTE-Advanced) and the improvements to the devices (e.g., size, screen resolution, processing power). Mobile voice has been overtaken by mobile data at the end of 2009 and it is expected that the growth of mobile voice traffic will remain limited compared to the rapid and huge growth of the data traffic [15]. This trend can be noted in Figure 22, where the growth of mobile voice traffic is compared with the growth of mobile data traffic during the period from 2007 to 2013. The numbers in the figure are derived from measurements over a large base of commercial networks that cover all regions of the world.

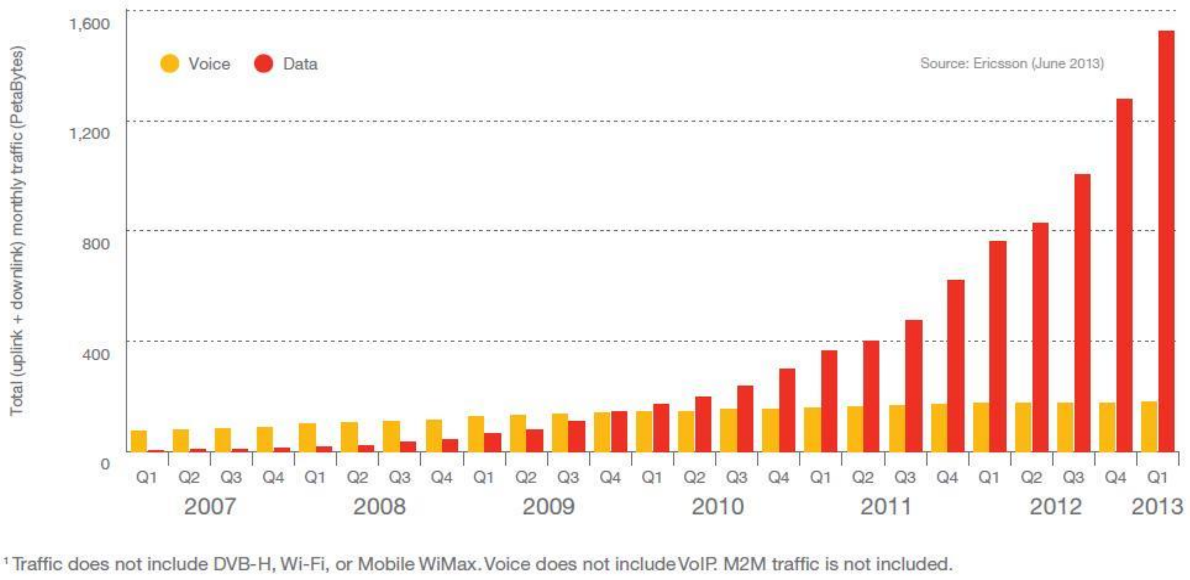


Figure 22: Global total data traffic in mobile networks, 2007-2013. [15]

Social networks applications generate significant traffic since they are widely used. The social networks applications themselves are not bandwidth hungry applications, but many times their usage is associated with video on demand utilization, thus generating a big amount of additional traffic. Figure 23 shows how the number of Facebook users through mobile devices has grown during the years.

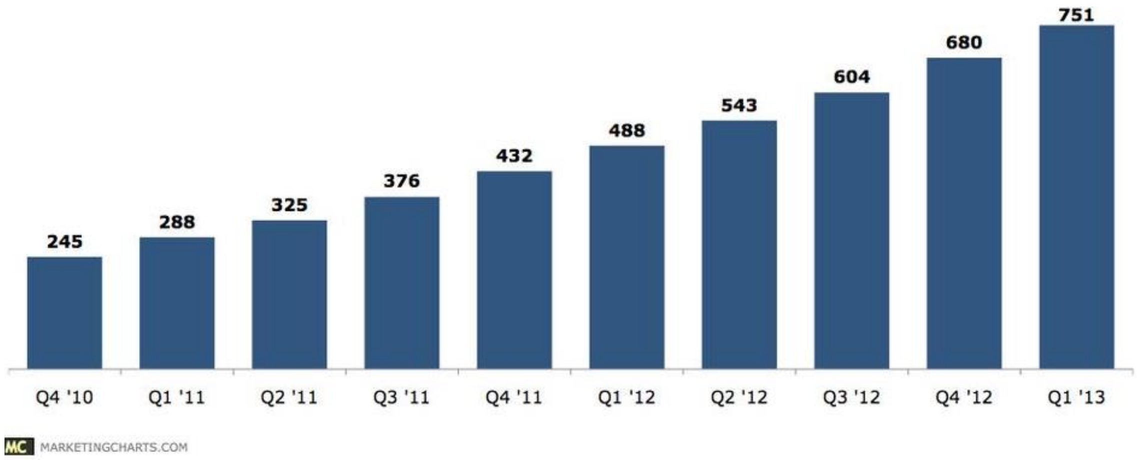


Figure 23: Faceboook's mobile monthly active users (Q4 2010 – Q1 2013). [20]

3.4 Traffic of new perspectives: CDN

Thanks to the development of multimedia technologies and high-speed networks, multimedia entertainment applications like video on demand, online songs and movies, IP telephony, Internet radio and television, and interactive games have now become popular networking applications over the Internet. To meet this demand, Content Delivery Networks (CDNs) have recently been widely deployed to deliver the contents from content providers to a large community of geographically distributed users. CDN service may be provided either by CDN service providers, such as Akamai that partners with multiple Internet service providers (ISPs), or by a big ISP itself. A CDN can achieve scalable content delivery by distributing load among its servers, serving user requests via servers that are close to the users.

In a typical CDN system, a user request is redirected to a nearby CDN server through a certain redirection mechanism, so that content delivery takes place at the edge of the network where bandwidth is abundant. This server performs admission control to either accept or block the request. If the request is accepted, the local CDN server accommodates the user request if it has the content; otherwise, it performs content routing to locate and then deliver the requested content to the user. In the existing CDN architectures, content delivery is achieved via managed and controlled servers. Resource usage depends on the popularity of the content provided and optimization can be performed on this basis. When considering an FMC scenario (described in a use case of COMBO deliverable D2.1), this architecture somewhat needs to change. One of the reasons for this is that all data has to go through the one core element of the LTE mobile core (the PGW), which is independent from the CDN location.

Cooperative caching and application-level multicast (ALM) are two important technologies in a multimedia CDN for delivering on-demand and live multimedia contents respectively. In cooperative caching, the CDN servers cache all the content cooperatively when single CDN is not capable of caching the whole content. In application level multicast, the content is delivered from the CDN server having the requested content through a multicast group formed of CDN servers and clients. With its overlay structure, CDNs shorten the paths and delays as well as stabilize the throughput.

With the development of popular video-streaming services that deliver Internet video to the TV and other device endpoints, CDNs have prevailed as a dominant method to deliver such content. Globally, 51% of all Internet traffic will cross content delivery networks in 2017, up from 34% in 2012. Globally, 65% of all Internet video traffic will cross content delivery networks in 2017, up from 53% in 2012 [11]. At 2020, traffic generated by content delivery networks will reach the volume of 90 EB/month catching the amount of video constituting 45% of global traffic.

The relative changes in volume for long-haul and metro traffic – briefly discussed through Figure 14 [11] earlier, in section 3.2.2 – are partially due to the increasing volume of CDN traffic. The prediction is that by 2017 metro traffic will be 2.4 times higher than long-haul traffic, in comparison with the 2012 status, where this multiplier was only 1.8.

The traffic of CDNs will play an important role within FMC traffic, and since it affects the traffic flow, the traffic management, and techno-economics as well, WP3, WP4 and WP5 of COMBO should consider its impact.

3.5 Signalling changes

It is supposed that the always-on data-centric nature of devices in the LTE/EPC networks can result in an explosion not only of data traffic, but signalling traffic as well [21]. Even more, it is expected that signalling traffic growth will significantly outpace mobile data traffic growth [22].

A load on network entities due to signalling traffic can be caused by, e.g.:

- frequent loss of broadband coverage that may potentially generate extremely frequent intersystem change activities [23];
- flood of registrations caused by 1) mass of mobile users attempting simultaneously to perform registration procedures such as attach or location updating and 2) restart of RAN and Core Network (CN) nodes which handle mobility management (MSC/VLR, SGSN, MME and HSS/HLR) [23];
- data traffic offloading that can also generate very frequent intersystem change activities.

Moreover, control data signalling tends to scale with the number of users, while user data volumes may scale more dependent on new services and applications [24].Thus, as the number of the LTE/EPS users is being increased, the signalling load on network entities is also growing.

One more trend related to signalling issues is to introduce a full separation between control signalling and user plane operation. The decoupling is motivated by the fact that both the control signalling operation and the user data functionality have to be implemented in optimized ways for separating these functions in the logical architecture [24]. One more fact is that the full separation of control signalling and data operation allows more flexibility in handling user data functions in a more distributed way in the networks, while at the same time allowing for a centralised deployment of the equipment handling the control signalling [24]. While most control layer decisions have been decoupled from user plane elements, the responsibility to disseminate them still relies on specific interfaces between the P-GW and S-GW, where the S-GW shares both control and user plane responsibilities.

Thus, the signalling traffic trends can be considered as open research issues and it is interesting to analyse both LTE signalling traffic that is expected to be generated towards Evolved Packet Core (EPC) and signalling changes that are required on the architectural design level to support true decoupling of control signalling from user data traffic. The former is considered in this section – the following paragraphs; whereas the decoupling of control and user planes is further described in Chapter 5.2.

Traffic signalling forecast

The always-on data-centric nature of devices in the LTE/EPC networks can results in an explosion not only of data traffic, but signalling traffic as well [21]. Some typical examples why signalling traffic can be generated are given below.

The User Equipment (UE) like smartphones are "always-on" devices that automatically attach to the network [21]. Each attached device generates signalling traffic. Besides, to conserve battery life, the smartphones can detach from the wireless connection after a short period of inactivity. Then, the UE should be reattached when data service is used again. It is obviously that the continual attach/reattach from a large number of smart devices generates considerable signalling traffic.

Besides, the applications installed in the UE create the signalling traffic. As application usage grows, signalling traffic increases [21]. Some applications, like instant messaging and social networking, are always-on services that require regular message exchange between the client and server. This continuous attach/detach/reattach process further increases the load in the control plane.

Moreover, if there should be a network problem where service is interrupted, once service is restored there will be a large amount of devices all trying to re-establishre-establish a connection automatically and simultaneously [21].

It is supposed that Diameter (a protocol used for authentication, authorization and accounting; together with policy control) signalling traffic will be dominant in LTE/EPC networks.

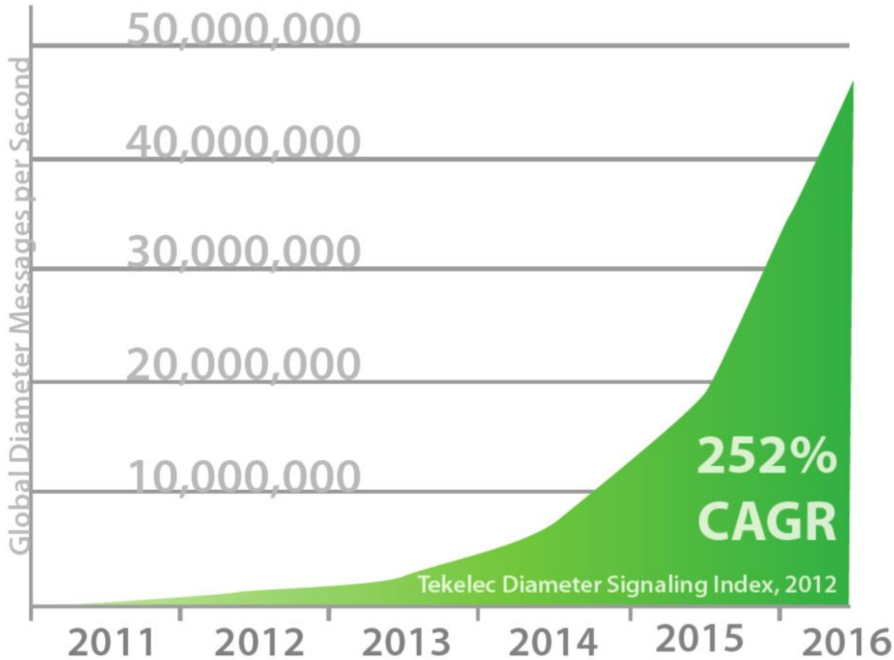


Figure 24: Forecast of global Diameter signalling traffic.

Figure 24 shows the CAGR of signalling messages required to support different services [22]. As one can see from the figure the number of Diameter messages per second (MPS) increases exponentially and is expected to reach 50M messages per second by 2016 in accordance with the forecast [22].

Control plane elements, which are mostly decoupled from the user plane path, handle Policy, Authentication, Privacy, QoS, Charging and Mobility functions [24]. Involved diameter control plane entities providing data and voice services are the HSS, EIR, OCS, and PCRF. The network entities of LTE/EPC architecture for supporting these services are illustrated in Figure 25 [22].

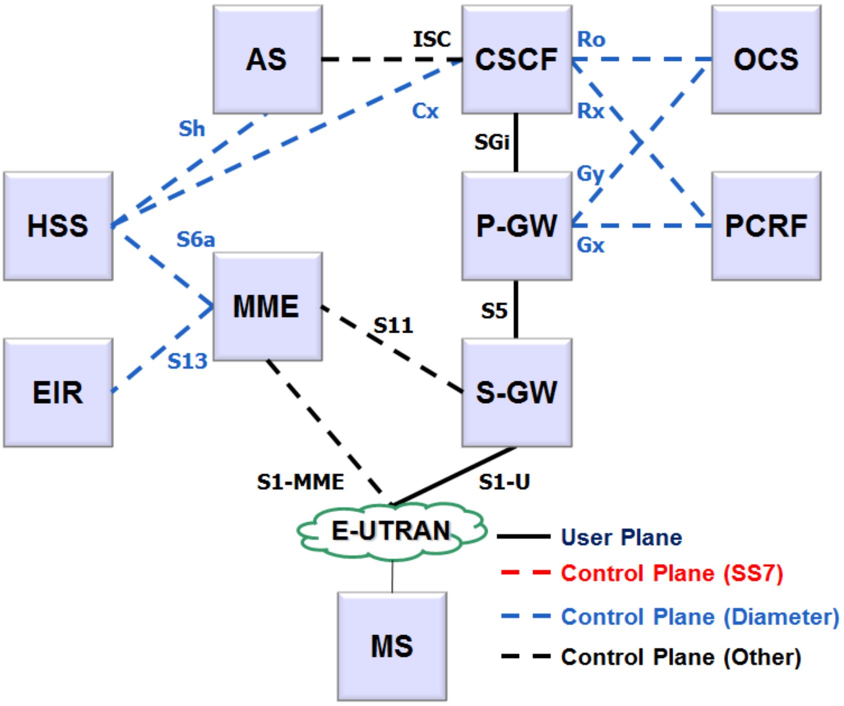


Figure 25: LTE/EPC architecture to support voice and data services.

As seen from Figure 26 [22], policy has the largest impact on signalling traffic growth since the PCRF interacts very frequent with the charging system (via the P-GW), enforcement points, etc. [22]. There can be several instances of the PCRF in the network because of scalability, redundancy and other reasons [21]. Online charging represents also essential contribution to the total Diameter signalling traffic because of the interactions between the OCS, PGW and PCRF [22]. As one can see mobility generates not so much signalling, but the relative values do not take into account LTE roaming procedures [22].

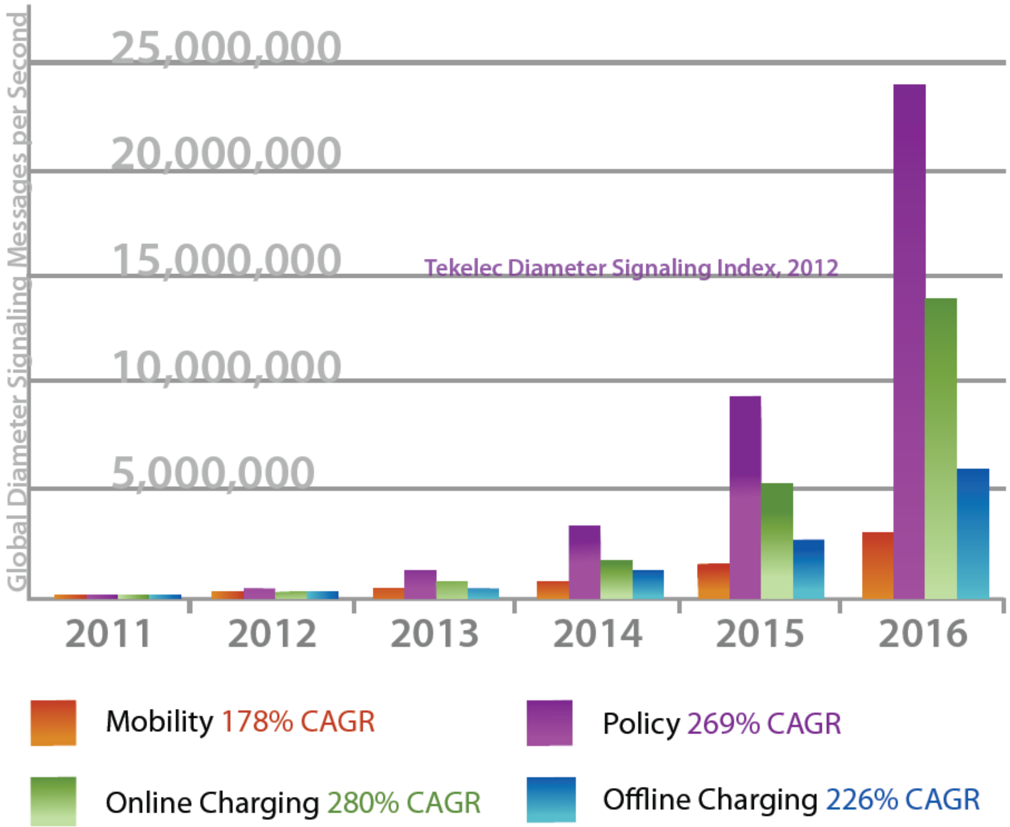


Figure 26: Forecast of global Diameter signalling traffic by message type. [22]

With regards to service type contributions to Diameter signalling growth, the report [22] states that voice and video over LTE produce the largest amount of signalling traffic.

As was mentioned above the attach/detach frequency of the UEs can essentially impact on amount of the signalling traffic, in particular, Diameter signalling traffic. We have used the Diameter traffic calculator [26] to illustrate dynamics of changes of signalling messages if attach/detach procedure is initiated each 15, 30, 45, and 60 minutes correspondingly. The methodology for calculation is based on the following events that can generate Diameter messages [25]:

- Attach of the UE to the E-UTRAN. It requires authentication of the terminal at the HSS and the EIR using S6a/S13. The HSS updates the MME with the subscriber data, the default data session is created at the S-GW and P-GW requiring policy using Gx from the PCRF and reports the initial charging event using Gy to the OCS.
- Tracking Area Update. When the device attaches, it is assigned a list of tracking areas as the registration area. When the UE moves outside the registration area a procedure is required with the MME, S-GW and P-GW to update the location of the UE and

possibly change the S-GW and/or MME that involves the use of S6a to the HSS and Gx to the PCRF.

- Service Request. The UE or the network may decide that the data session needs to be modified, e.g. to increase the QoS for a video download or reduce the QoS when the S-GW is overloaded. This requires modification of the data session at the S-GW and P-GW that involves the use of S6a to the HSS and Gx to the PCRF.
- Detach of the UE from the E-UTRAN. When the UE is powered off or the UE is not used for a prolonged period it detaches from the E-UTRAN requiring the use of the S6a, Gx and Gy.

Initial data that we use for Diameter signalling traffic evaluation is summarized in Table 8.

Number of LTE Devices (Million)	1
Annual Growth Rate of LTE Devices (%)	150%
Number of Simultaneous always-on Apps per Device	1
Annual Growth Rate of Simultaneous Apps per Device (%)	20%
Frequency of device attach/detach (min)	15, 30, 45, 60
Frequency of tracking area updates (min)	60
VoLTE Enabled (%)	0
Annual Growth Rate of VoLTE Enabled (%)	25
VoLTE BHCA per Subscriber	2
Annual Growth Rate of VoLTE BHCA per Subscriber(%)	25
Policy Enabled (%)	100

Table 8: Initial data for signalling traffic forecast.

The calculation is done for 1 Million of LTE devices. It is assumed that initially there are no LTE subscribers that use VoLTE. However, there will be a 25% annual growth of subscribers using LTE infrastructure to make voice calls. The number of attempts of the VoLTE calls in the busy hour is also supposed to grow by 25% annually.

Diameter signalling traffic forecast based on these input data depending on UE attach/detach frequency is shown in Figure 27.

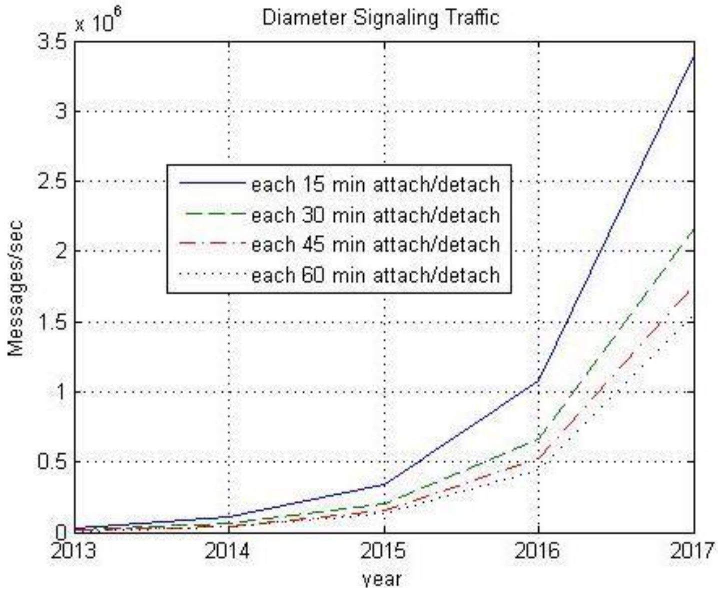


Figure 27: Diameter signalling traffic forecast.

As seen from Figure 27 the total amount of Diameter signalling messages is increased dramatically when frequency of device attach/detach to/from network is changed. In particular, if the device attach/detach frequency is changed from once per hour to one every 45 min, the increment of signalling traffic is 13%. If the attach/detach is completed each 30 min the traffic growth is around 40%. If it occurs every 15 min, the amount of signalling traffic grows more than twice in comparison with 60 min attach/detach. Therefore, the mobile network operators (MNOs) should be very careful with this parameter. For instance, if a MNO introduces offloading strategies in its network, on the one hand it can reduce essentially the amount of data traffic, but on the other hand it can lead to explosion of signalling traffic. Very frequent attach/detach process can happen, e.g. if users move by cars or trains from home to office and back and many offloading areas are distributed along the road.

4 TRAFFIC ANALYSIS AND MODELLING

4.1 Considerations on traditional traffic models for FMC

The evaluation of the network performance can be very complicated. In particular, in order to predict the performance of a network in certain traffic scenarios, it is important to be able to properly model the input traffic. The usage of traffic models allow the preparation for various outcomes – without having to test with massive amount of real users who are actually creating those various traffic scenarios that can be modelled as well.

Moreover, the proliferation of wireless technologies has allowed convergence providing aggregation of heterogenous devices and access technologies. These changes have also led to the integration of different services such as voice, video and broadband data services used over both mobile and fixed networks [27]. Technological development, market demands and deregulations made it possible to introduce the changes required to develop the aforementioned convergent environment (FMC). In order to plan the actual infrastructure, predictions on traffic characteristics are required – which can be based on the analysis of current traffic. The various FMC scenarios give additional complexity to the FMC traffic – its modelling is more complex than the superposition of standalone models. (Nevertheless, proper modelling can only be carried out once the data is already available about the systems to be modelled – it is not yet available for FMC scenarios.) Such convergence yields the need for traffic models capable of accurately characterizing the behaviour of complex network traffic.

Several models of traffic have been proposed for all kinds of network environment. Each model has a different impact on network performance evaluation. Within such models, two major focuses are the packet length distribution and the packet inter-arrival time distribution (that is, the *time correlation* of traffic).

However, modelling heterogeneous network traffic remains an open issue, where the main challenges are due to the different structures and features in the traffic behaviour for interconnected (IP based) communication networks.

If we focus on time correlation of traffic, two main categories of traffic models can be identified: *non-self-similar traffic models* (both memoryless, or Poisson-based, and with memory, or time-correlated) and *self-similar traffic models*.

Memoryless traffic models based on Poisson processes, e.g., packet arrivals or phone calls modelled as Poisson events, are the oldest models. They are based on the well-known result from stochastic processes theory that large aggregations of (even non-Poisson) traffic from independent sources tend to form a Poisson process.

Anyway, it has been observed that real network traffic may exhibit bursts, contrary to Poisson-type processes. For this reason, new models were proposed to describe bursty traffic, including appropriate time correlation.

More recently, the intriguing property of self-similarity (fractal behaviour) has been widely detected in network traffic measurements. This property seems ubiquitous and independent of how large the aggregation of traffic sources is. Its origin has been debated lengthy, but the empirical evidence is that network traffic exhibits self-similarity at all network segments, although not in all cases.

The next section summarizes those currently available traffic archives, which are the most widely used. The section afterwards provides an overview of traffic models researched and applied for fixed and mobile networks. Later on analysis and modeling methodologies are shown, which are suggested to be used even for complex scenarios that appear in FMC cases as well.

4.2 Survey of traffic archives

This section provides a summary of some of the most important online archives of Internet traffic measurements. In particular, archives are listed with information about accessibility, data type, link capacity, etc. For further information, links to websites are provided too.

4.2.1 CAIDA

Web site: http://www.caida.org/home/

These datasets contain anonymized traffic traces coming from high-speed Internet backbone links located between San Jose and Los Angeles and between Chicago and Seattle; in both cases, measures are divided in two directions. The anonymized traces are available from 2008 to 2013 and they are stored in pcap format with timestamps truncated to microseconds. The original nanosecond timestamps are provided as separate ASCII files alongside the pcap files (times files).

4.2.2 CRAWDAD

Web site: http://crawdad.cs.dartmouth.edu/data.php

CRAWDAD (Community Resource for Archiving Wireless Data At Dartmouth) is a wireless network data (including cellular and Bluetooth) available for the research community. CRAWDAD requests registration to enable the user downloading data. Available data are composed of data traces, a lot of tools for traffic management and some publications related to the data. Traces are in tcpdump format.

4.2.3 WITS

Web site: http://wand.net.nz/wits/catalogue.php

WITS (Waikato Internet Traffic Storage) brings together a multitude of different data recorded in New Zealand. Traffic traces were provided by some university (Auckland and Waikato) and some unnamed ISP. The IP addresses were anonymized using Crypto-Pan AES encryption. Archives are available for free but you can find some problems if the download is not done with an IPv6 address, that is why trying downloading with common IPv4 address may be useless. Still, there is a part of WITS database available on the site http://www.ripe.net that can be downloaded through ipv4 address.

4.2.4 MAWI Working Group Traffic Archive

Web site: http://mawi.wide.ad.jp/mawi/

This web page contains a large amount of traffic traces, most coming from trans-Pacific line (connecting US with Japan) and also some traces about IPv6. Data are daily updated and it is possible to download (without registration) full years of measures divided day-to-day.

4.2.5 UMass Trace Repository

Web site: http://traces.cs.umass.edu/index.php/Network/Network

This is a big free archive with lots of trace files measured on many different links. These traces are not sniffed from real user connections, so this is a simulation made for testing network performance. You have to pay attention because files are sorted by associated publication year, not by trace collection year. Each file has its own documentation text with date and hour of registration and explanation of the kind of link involved.

4.2.6 Universita' degli Studi di Napoli "Federico II"

Web site: http://traffic.comics.unina.it/Traces/ttraces.php

Little free archive created for academic purpose. The link observed is a link at 200 Mbps connecting the University of Napoli "Federico II" network to the rest of the Internet. These traces are in tcpdump format. Packet lengths are variable because, for each packet, full TCP headers are stored, including optional headers (e.g. MSS). Moreover, because of privacy concerns, IP addresses have been replaced preserving network membership.

4.2.7 Simple Web - Dropbox Traces

Web site: http://www.simpleweb.org/wiki/Traces

The data set was captured at 4 vantage points in 2 European countries. The first 4 files were collected from March 24, 2012 to May 5, 2012. A second dataset was collected in Campus 1 in June and July 2012. All IP addresses in the datasets are anonymized. The data were captured using Tstat, an open source monitoring tool developed at Politecnico di Torino. The Tstat setup provided a pool of unique datasets, allowing to analyse the use of cloud storage in different environments, which vary in both the access technology and the typical user habits. Further information available at: http://eprints.eemcs.utwente.nl/22286/01/imc140-drago.pdf

4.3 Overview of Traffic Models

The next sections outline the characteristics of those traffic models that consider self-similarity and those that do not. Moreover, a separate section is devoted to the mobile (or wireless) traffic modelling, due to some specific features of this type of networks.

A good survey of traffic models can be found in [28]. This includes the overview of Non-Self-Similar Traffic Models (Poisson, Pareto and Weibull Models, as well as the Markov Modulated Poisson Traffic Model). Overviews on Self-Similar Traffic Models (Self-Similarity and Long-Range Dependence, Fractional Brownian Motion) can be found in [29][31][33].

4.3.1 Traffic Models for Wireless Access Networks

With the evolution of cellular mobile system in the last decade, the problem of traffic modelling and analysis of mobile traffic over macro, micro and picocells has become a new research challenge [47].

Within mobile network systems, classical analytical models may not apply [52]. Indeed, interactions among multiple interfaces [47], picocells and femtocells environment [45][46] traffic modelling [48]-[50] and user mobility modelling [49]-[51] (including handoff processes) pose various challenges in designing integrated networks. Therefore new traffic models, especially designed for mobile traffic need to be investigated.

4.3.1.1 An Introduction to Traffic Modelling and Analysis for Wireless Networks

There is an extensive body of literature on traffic modelling and analysis for cellular networks. Because of handoffs between cells, standard models applied for decades in telephony had to be adapted. Furthermore, with different traffic types, networks with different types of cells (e.g., mixed microcells and macrocells) and the introduction of features like guard channels, and so on, the problem can be quite challenging.

In the great majority of the papers and books, Markov models are used because of the ease in applying standard queuing theory analysis and results to the problem. Even if the actual traffic characteristics may not be exactly Poisson, it is hoped that they are close enough so that the analytical or simulation results obtained by using a Poisson process model are still meaningful.

The Poisson process assumption has been challenged by some researchers (especially the handoff traffic, rather than new call traffic [58]), as a result of which newer models have been proposed that are supposedly more accurate.

One of the main thrusts of research is how to differentiate the handling of handoff traffic and new call traffic. Handoff calls are often given higher priority than new calls. In fact subscribers would find it more annoying if existing calls are dropped when handoff fails, with respect to situations where new calls are blocked [59][60].

The complexity of the standard queuing system model grows exponentially as the number of states increases, i.e., as the number of channels and cells increases. Approaches to simplifying the teletraffic analysis problem to something more computable have generally focused in simplifications to the model itself, or on ways to simplify the computations for a given model.

Of the attempts to simplify the model itself, making many assumptions related to homogeneity and loose coupling, some researchers have attempted to obtain useful information about a whole network by using a one-cell model. However, the accuracy of single-cell models is very questionable for sophisticated wireless networks that employ channel borrowing, dynamic guard channels, etc. Furthermore, it is not useful for multi-layer networks.

4.3.1.2 Telephone Traffic Modelling in Mobile Cellular Networks

In this and next sections call-related models are often referenced. Most of these models can be and are applied for data traffic as well. In there, instead of "calls", their data-networking synonym: communication "sessions" should be considered.

In classic telephone traffic theory, developed for wired networks, call (in data networks: session) arrivals to a local exchange are usually modelled as a Poisson process, at least over short observation intervals to assume stationary arrival rate, since the user population served by the exchange is very large and with negligible correlation among users. This assumption of memoryless traffic has been often retained also in presence of mobile users: in literature, incoming calls in cellular networks are mostly modelled as a Poisson process, with both call holding time and interarrival time assumed with negative exponential distribution.

Nevertheless, as discussed in [43], it has been argued that this Poisson assumption might not be valid in wireless cellular networks.

A further empirical study on real GSM telephone traffic data was reported in [44]. Answered call holding and interarrival times were found to be best modelled by the lognormal-3 function, rather than by the Poisson negative exponential distribution.

In summary, *several studies contradicted the ubiquitous likelihood of the classic Poisson model for telephone traffic in cellular networks and suggested that call arrivals may be significantly time-correlated, due for example to access congestion, user mobility and possible correlation between nearby users.*

However, we note that the Poisson traffic model is still assumed in almost all works, mainly for the sake of simplicity, when cellular network performance is evaluated. Questions may arise, therefore, on the practical relevance of this simplifying assumption.

In [43], authors investigated possible time-correlation of both originated and terminated answered call arrivals in sets of real GSM telephone traffic data, provided by an Italian operator. After examination of such empirical data, the authors found that call arrivals proved excellent consistency, by MAVAR analysis and χ^2-test evaluation, with a *non-homogeneous Poisson model with diurnal variable rate*.

Results presented in [43] confirm, at least to the limited extent of those empirical data, that the Poisson model may be still adequate to describe realistically telephone traffic in cellular networks, unless focusing specifically on particular issues like small user population, access congestion and very frequent handovers.

4.3.1.3 Models based on Poisson distributed call arrival with handoff processes

In [48], a traffic model for cellular mobile telephone systems is presented, where handoff procedures are considered.

In the system model, the new call origination rate is uniformly distributed over the mobile service area. It is also assumed that a very large population of mobiles is served, and the average call origination rate is independent of the number of calls in progress.

4.3.1.4 Poisson-arrival-location model

In the Poisson-arrival-location model (PALM) [53] customer arrivals are modelled by a non-homogeneous Poisson process and move independently through a general location state space according to a location stochastic process.

PALM assumes that different users do not interact and that the network has no capacity constraints in term of the number of radio channels available for calls. Such model can be used to represent both a time-dependent behaviour and users mobility in wireless communication networks.

4.3.1.5 Models based on general distributed Poisson call arrival for general distributed handoff processes

One of the main differences between telephone fixed public switched transport networks (PSTN) and mobile cellular networks is the presence of the handoff traffic [47]. Handoff calls, in several works e.g., [48][50][54], are assumed to be memoryless Poisson processes. However, in microcellular systems, a mobile user can request handoff multiple times at various cell boundaries altering the traffic behaviour.

Authors of [47] and [55] consider that the Poisson assumption may not apply for multiple handoffs of a call, and suggest a two-moment approach suitable for both traffic streams of micro and picocells. Each of the two moments are represented by their mean and variance. Their studies assume fixed channel allocation scheme.

The two models in [56], called Probability Generating Function (PGF) and Binomial Moment Generating Function (BMGF), are more complex. They are based on semi-Markov analysis with no closed form solution and have complex algebraic solutions.

4.3.1.6 Traffic Model for calls in integrated femto-macro cellular network

A tractable traffic model for the offload from macrocell to femtocell is a research area that received little attention in the research community. The large and dense scale deployment of femtocells makes the handover issues more challenging. A well designed femtocell-macrocell integrated network can divert heavy traffic from congested macro cellular network to femtocell network.

Chowdhury and Jang [57] studied a traffic model based on macrocell-femtocell integrated network. Call inter-arrival process is assumed to be Poisson. Femtocells are randomly deployed within the macrocell coverage area.

In such model, during the handover of a call both macrocell-to-macrocell and femtocell-to-femtocell handover calls/sessions can occur. Moreover, also direct calls/sessions can arrive to the macrocell during the considered handover. A femtocell may not accept the call/session in case of poor signal to noise interference ratio (SNIR).

4.4 Traffic Analysis Methodology

In this study, we aimed primarily at detecting the degree of self-similarity or long-range dependency (LRD) of network traffic, because of its large impact on network resource planning. This is true in both fixed and mobile access networks, whenever buffers or link capacities need to be designed to meet given performance requirements. However, it would be interesting to understand if traffic in new FMC networks will exhibit peculiar behaviours in terms of self-similarity or not.In brief, self-similarity of the traffic yields for over-planning of the resources; whereas no self-similarity and no LRD allows more resource-effective planning. This is because traffic bursts are less probable – and more easily predictable in the latter case.

We recall that, by definition, LRD of a process is defined by an asymptotic power-law decrease of its autocovariance or equivalently power spectral density (PSD) functions [30][31][32]. For example, in the Fourier frequency domain, the PSD of a LRD process *Y*(*t*) follows asymptotically

$$S_Y(f) \sim c_2 |f|^{-\gamma} \quad \text{for } f \to 0\,,\, 0 < \gamma < 1 \qquad (1)$$

In other words, LRD consists in a power-law behaviour of certain second-order statistics versus the duration τ of the observation interval in the time domain, or the Fourier frequency *f* in the frequency domain.

In short, detecting the degree of self-similarity or long-range dependency in measured traffic data means to estimate model parameters *H* and γ, assuming a LRD model based on a power-law PSD of traffic data, with

$$\gamma = 2H - 1. \qquad (2)$$

Several techniques exist to estimate *H* and γ of data series, supposed self-similar or LRD, both in the time domain (e.g., variance-time plot) and in the frequency domain (e.g., periodogram) [30][32][35], which are all based on measuring the slope of a linear fit in a log-log plot.

A class of more advanced techniques is based on wavelet analysis [32][33][34][36] [38]. Among wavelet-based techniques, the so-called *log scale diagram* (LD) is one of the most important [32][37]. It analyses data over an interval of time scales (octaves), ranging from 1 (finest detail) to a longest scale given by data finite length. Also in this case, by observing the diagram slope, *H* and γ are estimated.

To estimate LRD parameters, we use instead the method based on the Modified Allan Variance (MAVAR), which was proposed only recently and was demonstrated to be more accurate, with finer resolution and simpler than such wavelet methods [41][42].

4.4.1 The Modified Allan Variance

The complete description of the Modified Allan Variance (MAVAR), and the brief summary of its properties can be found at [41][42].

Given an infinite sequence $\{x_k\}$ of samples of $x(t)$ with sampling period τ_0, MAVAR is defined as

$$\mathrm{Mod}\,\sigma_y^2(\tau) = \frac{1}{2n^2\tau_0^2}\left\langle \left[\frac{1}{n}\sum_{j=1}^{n}\left(x_{j+2n} - 2x_{j+n} + x_j\right)\right]^2 \right\rangle = \\ = \frac{1}{2}\left\langle \left[\frac{1}{n}\sum_{j=1}^{n}\left(y_{j+n} - y_j\right)\right]^2 \right\rangle \qquad (3)$$

where <·> denotes infinite-time averaging, $\tau = n\tau_0$ is the observation interval and y_k is the average value of $y(t) = x'(t)$ over interval τ beginning at t_k, i.e. the k-th sample of the first difference of $\{x_k\}$ with lag τ:

$$y_k(\tau) = \frac{1}{\tau}\int_{t_k}^{t_k+\tau} y(t)dt = \frac{x_{k+n} - x_k}{n\tau_0} \qquad (4).$$

Thus, MAVAR is a kind of variance of the first difference of $\{y_k\}$ or of the second difference of $\{x_k\}$ (note: differences multiplied by τ). In very brief, it differs from the unmodified Allan variance in the additional internal average over n adjacent samples: for $n = 1$ ($\tau = \tau_0$), the two variances coincide.

4.4.2 Using MAVAR for Estimating the Hurst Parameter

Let us consider a LRD process $x(t)$ with PSD and Hurst parameter $1/2 < H < 1$. Then, MAVAR follows $\sim\tau^{\mu}$ (ideally for $n\to\infty$) with $\mu = 2H-4$. In brief, the following procedure is suggested to estimate H:

- 1) compute MAVAR by using the ITU-T standard estimator [39], based on $\{x_k\}$, for integer values $1 \le n < N/(M+1)$ (we use a geometric progression of ratio 1.1, i.e. 24 values/decade, for finest rendering of trend);
- 2) by least-square linear regression, estimate the average slope μ of MAVAR in a log-log plot for n>4 and excluding highest values of n, where confidence is lowest;
- 3) if $-3 < \mu < -2$ (i.e., $-1 < \alpha < 0$, $0 < \gamma < 1$), get the estimate of the Hurst parameter as

$$H = \mu/2 + 2 \qquad (5).$$

Under the more general hypothesis of power-law PSD [40], then up to P slopes μ_i can be identified ($-3 \leq \mu_i < 2M-2$) to yield the estimates $\alpha_i = -3-\mu_i$ ($-1-2M < \alpha_i \leq 0$) of the P components of f^{α_i} noise.

Some care should be exercised against non-stationary terms in data analysed (e.g., big steps, slow trends), which cause slope changes that may be erroneously ascribed to random power-law noise. On the other hand, polynomial drifts are cancelled, unless their order is greater than M. Thus, the order M can be conveniently adjusted. In [37] (Sec. III.B.4), similarly, it is suggested to increase the number of vanishing moments until the H estimate converges to a stable value, thus indicating that all smooth trends have been cancelled.

A key issue is to determine the confidence of these estimates and whether they are unbiased or not. In [37] (Sec. III.C), this problem is studied for the H estimator based on wavelet decomposition. Provided that the number of the vanishing moments is chosen appropriately, the estimator is proven to be unbiased (or with low bias on finite data sets). Closed forms of the variance and confidence intervals of this estimator are derived as well, although under a number of simplifying assumptions. Since MAVAR can be rephrased in terms of appropriate wavelets, it can be argued that similar results may be valid also for the estimator proposed herein.

Nevertheless, deriving exact expressions for the confidence intervals of H and α_i estimates is not immediate. Exact computation of confidence intervals of MAVAR estimates is tedious and even depends on the spectrum of the underlying noise. Therefore, the evaluation of confidence intervals of estimates of H and α_i results even more complex, depending also on the algorithm used to estimate the average slope of curves and on the interval on which this is carried out.

4.5 Traffic analysis and complex modelling methodology at a macro-level

4.5.1 The cycle of model-based traffic generation

The following methodology aims at providing a model for complex control plane as well as data plane traffic at the edge of the mobile as well as the fixed access network. This model is going to be used within the traffic generators at practical laboratory experiments within WP6 of COMBO.

This methodology will allow to model the traffic behaviour of hundreds of thousands of users as they attach to the network, initiate user-related communication, and move within the network while producing and receiving traffic.

A major task of the methodology is traffic **analysis**, carried out in order to find proper **models** and parameters that fit this behaviour. The models are implemented in the form of a traffic generator, which produces synthetic traffic whose statistical parameters match the **observed** real-life patterns. After **verification** and fine-tuning of the model, traffic generators can be **deployed**, and operated based on the verified traffic models. This can help revealing the performance limitations of the core of FMC.

Figure 28 illustrates this methodology through a general cycle of model-based traffic generation: observation, analysis, model creation, implementation and finally verification and deployment.

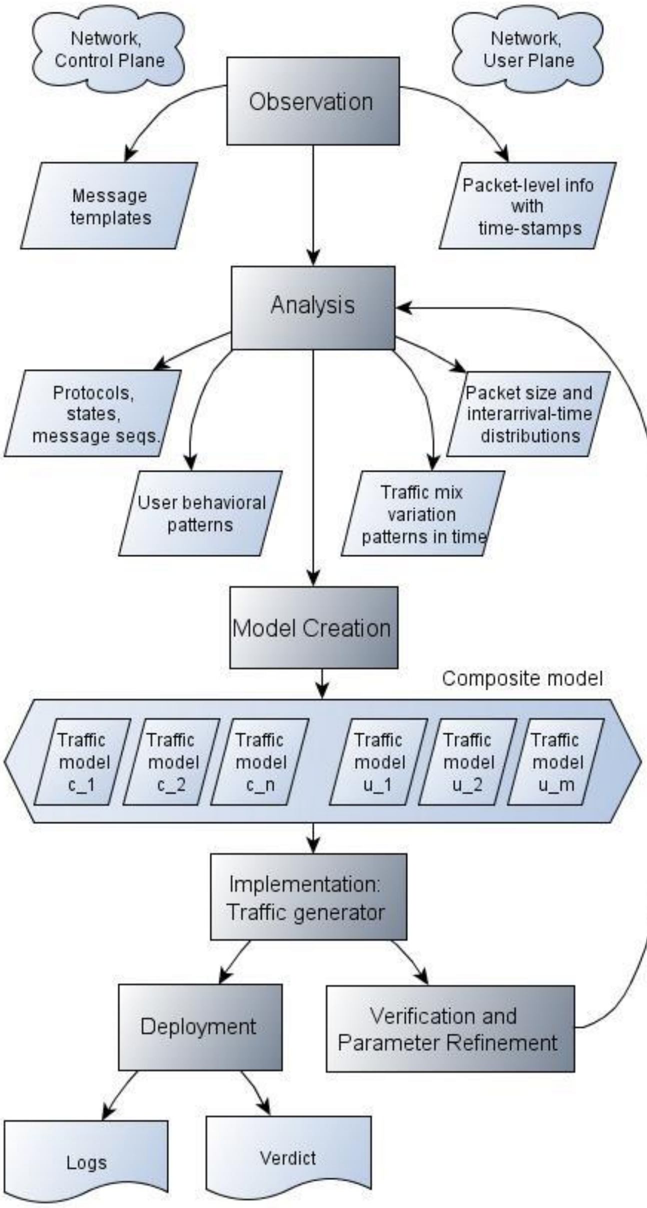


Figure 28: Tasks and their results in the methodology. Shaded grey rectangles illustrate procedures; parallelograms represent intermediate data.

Observation: Capturing of real-life traffic data in an operational network. Traces are collected from both the control plane and the user plane.

Analysis is performed on the collected data. Types of various network activities are identified. For each activity, relevant features and statistical parameters are identified, and their values are extracted from the data.

Model creation: Based on the relevant parameters and message samples models are built, which account for the various activities and traffic patterns observed in the network.

Implementation: The models are realized as a traffic generator. The device simulates the operation of a specific network segment through the parameters of the implemented models.

Verification and refinement: The statistical properties of the synthetic traffic are matched against those observed in real patterns. The models' parameters are refined in an iterative process in order to improve the prediction accuracy of the models.

Deployment: Once the models are considered sufficiently accurate, the traffic generator can be deployed in a pilot network (or live network) to carry out complex load testing tasks. At this point the models may be operated outside the previously observed realistic parameter range. Thus the device can be used to simulate extreme network activities or boundary conditions, which would otherwise be difficult or impossible to produce in a real-life setup.

Due to their distinct characteristics, control and user planes need to be addressed separately throughout the model creation process.

4.5.2 Modelling of the Control plane for traffic generation

Control plane messages are captured bit-by-bit – at each reference point of the network where control plane traffic is to be modelled. In the analysis phase, control message sequences and protocol state transitions are identified and stored. Subscriber mobility patterns are analysed and their relevant parameters are identified.

The models created for the control plane are mainly based on protocol specifications: message sequences and state transitions need to conform to the standards and vendor-specific extensions. Subscriber mobility models, on the other hand, can as well employ statistical parameters.

4.5.3 Modelling of the User plane traffic

User plane traffic capture is a process of collecting data packets sent over the network.

The analysis needs to identify categories of user activities (such as voice calls, video streaming, web browsing or email traffic) and define relevant parameters which characterize the activities (e.g. packet delay and jitter for VoIP and video; expected value and variance of packet sizes for email download).

Typically different sets of parameters are identified for different types of observed activities.

4.5.4 The Composite traffic model

In order to model the traffic in actual networks, a collection of different traffic models need to be elaborated, each one accounting for a particular type of user activity. That is, different models need to be used for characterizing, e.g., voice calls, video streaming, web browsing or email traffic. Similarly, on the control plane different models are needed for e.g. describing subscriber attach demand or mobility within the network.

The specific models are then combined into a composite model; its structure is depicted in Figure 29). The composite model characterizes the different types of subscriber activities on the control plane and tells us how particular subscriber activities contribute to the overall observed traffic on the user plane. In this sense the composite model is a super-model, whose parameters are the constituting models themselves.

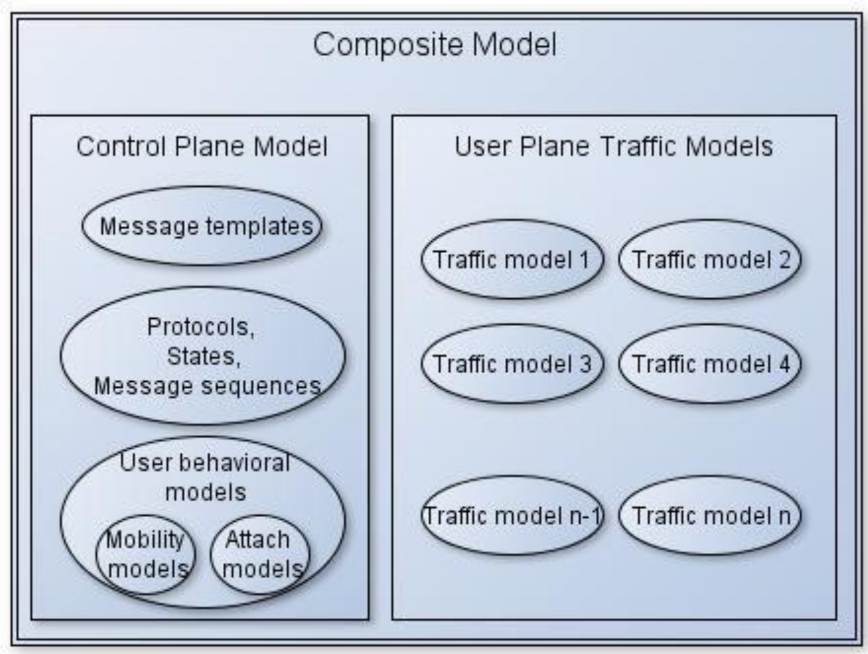


Figure 29: Elements of the composite activity model.
Different types of models describe different activities on the control and user plane.

In the following subsections the process of data capture, analysis, model creation, implementation, verification and deployment are detailed.

4.5.5 Data capture

The motivation behind traffic capture in the described method is to provide data for analysis and model creation. Two key features of the capture process are *losslessness* and *accurate time-stamping*, which are addressed below. Capture is performed at multiple network interfaces of the control and the user plane.

Control plane data capture

The collected data are practically bit-by-bit captures of control plane traffic. The capture contains actual protocol messages along with mobility information. These control messages can be collected into a set of templates, used by the implemented traffic generator in the network testing phase.

Losslessness is a key requirement here. Missing even one control message may lead to missing, e.g., a temporary identifier update, which in turn may lead to loss of a whole communication session.

In general, accurate time stamping is not of paramount importance for the control messages, as long as the original order of the messages is preserved. Time stamping of the **mobility** control messages, on the other hand, carry valuable information for model creation. Accurate time stamping is necessary in order to build a mobility model from the statistical properties of how location changes appear within the live network.

Note that control plane data account for a relatively small portion of the total traffic - the counter example being, e.g., machine-to-machine communication.

User plane data capture

The actual contents of the user packets are not relevant from the model creation point of view. However, in-band signalling in the user plane needs to be captured and stored accurately, just like in the case of control plane messages.

In the case of user plane packet capture, losslessness is a less strict requirement. Actually it may be satisfactory to capture packet headers only. Whether or not headers only would suffice depends on the planned depth of the analysis, and the granularity of the modelling. As an example: traffic generation with random payload does not require payload fingerprint analysis. It is completely satisfactory to chop such packets during their capture, and merely analyse their time-stamp and the 5-tuple of “from-IP”, “to-IP”, “from-port”, “to-port”, and “protocol”.

The requirement for the accurate time-stamping is also determined by the purpose of the modelling. When the model includes packet inter-arrival time modelling (or even taking long-range dependence into account), the theoretically smallest packet inter-arrival time determines the required time-stamping accuracy for proper capture. (It is 672ns for 1 Gbit/s Ethernet, and 6.72ns for 100 Gbit/s Ethernet.)

4.5.6 Data analysis

The purpose of data analysis is to categorize activities in the network and identify relevant parameters for the various activities in order to build models from them. We would also like to find typical and non-typical traffic patterns so as to use them for generating varying, real-life-like synthetic traffic. The user and the control plane data should be analysed separately.

Protocol analysis in the control plane

The control plane of the mobile core is responsible for call and session management, mobility management, policy control and charging (when referencing mobile-related behaviour, we describe them within the LTE terminology). For FMC traffic the same applies; although mobility management does not exist as such for fixed traffic.

Signalling on the control plane uses the message-oriented, reliable SCTP transport protocol. This ensures that the communicating devices always receive complete signalling messages in correct order.

The actual task of protocol analysis is to

- identify the dialogs to be simulated (include ie., the Initial Attach, Periodic Location Update, Normal Location Update and PDP (Packet Data Protocol) Context Activation procedures),
- collect message templates for the dialogs,
- identify relevant message parameters.

Some subscriber activities can be characterized by statistical parameters. Such are the probability of changing location within the network, activating a PDP context or detaching from the network.

Traffic analysis in the user plane

Categorization of network activities is important for successful model creation. Data traffic is a result of parallel user activities: users are simultaneously engaged in voice calls, view video streams, browse the web, read emails, and so on.

Different user activities produce different traffic patterns. Each traffic pattern can be characterized by various parameters. These parameters are chosen so that they help building an effective traffic model for the particular traffic pattern.

From this point of view traffic analysis is a complex deconvolution problem. In the aggregate traffic flow of the user plane one needs to identify the various user activities and extract the relevant parameters, which characterize the traffic pattern produced by the particular type of activity.

Deep packet inspection at the user plane alone may not be sufficient to differentiate types of user traffic. In this case analysis of the control plane messages can give us further clue for correctly identifying user plane traffic.

User plane packets account for a significant portion of the overall traffic volume. Relevant parameters (such as packet type, size, source and destination addresses) are extracted from the data flow. The actual contents of the user plane packets can be discarded.

4.5.7 Model creation

The composite model described in this section incorporates building blocks from earlier published modelling methods, briefly summarized in the following references.

Chandrasekaran [28] gives an overview of the various traffic models. The Cisco paper [61] on VoIP traffic patterns compares various traffic models for voice calls, including Erlang B and C, Extended Erlang B, Engset, Poisson, EART/EARC, Neal-Wilkerson, Crommelin, Binomial and Delay.

The UMTS Forum Report44 [6] is built on a traffic model which distinguishes service categories (video/audio streaming, mobile gaming, etc.), device categories (smartphones, tablets, connected embedded devices, etc.) and various subscriber activity patterns. These components contribute to the overall traffic models through parameters such as traffic per service/device, device mix, upload/download direction, period of the day/week/year, and others. The traffic model selection criteria are based on the number of sources (finite, infinite), call arrival patterns (smoothed, peaked, random), holding times (exponential, constant) and what happens to blocked calls (cleared, held, delayed, retied).

Aalto et al. [62] compared various scheduling algorithms from link delay and fairness aspects, and found that scheduling algorithms in the access network have their impact on the observed traffic in the core network.

The composite model for traffic generation

This section describes the creation methodology for the composite traffic model, which consists of independent, classical traffic models. The parameters of the composite model include:

- total number of simulated subscribers in the system,
- number of subscribers taking part in each activity (i.e., number of subscribers for each traffic model),
- triggering events/probabilities to change any parameter in any of the traffic models,
- the rate at which subscribers start/stop particular activities (i.e., the change rate in the number of subscribers simulated by a particular traffic model, such as attach/detach rate).

In the composite model we assume that one subscriber is engaged in one activity at a time. That is, a simulated subscriber is used in only one traffic model at a time. In the pool of simulated subscribers, each subscriber is assigned to a given traffic model for a set period of time, in an on-off manner. Figure 30 depicts an example of how the user traffic models contribute to the composite model in time.

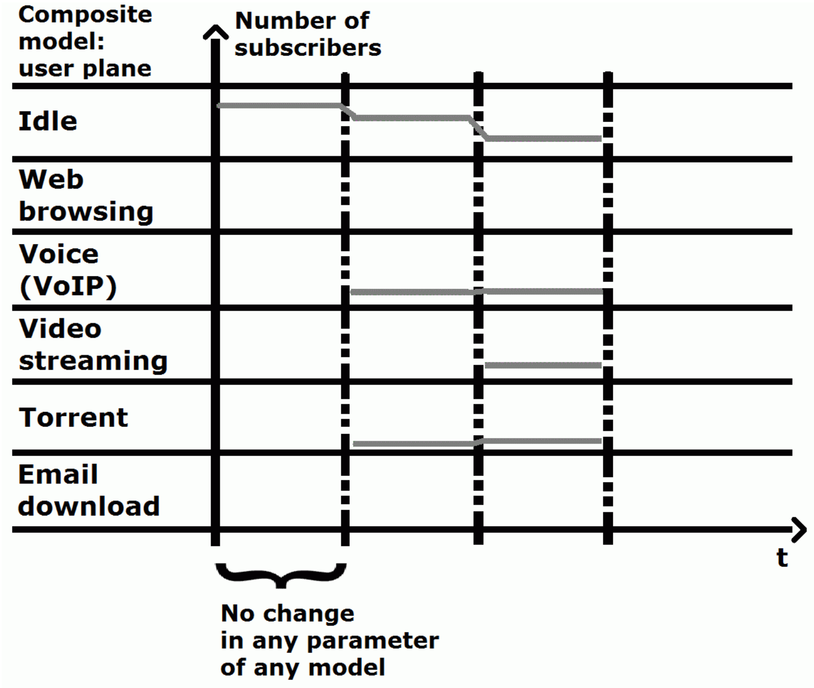


Figure 30: Superposition of user traffic models in time.

In order to simulate a limited number of subscribers, we found it practical to introduce an *Idle* traffic model. This model acts as the subscriber pool, which holds available users for the various activities simulated by the models.

The traffic model parameters are considered as target values only, in the context of other model parameters. The target may not be met if fewer subscribers are simulated than required by a model or the traffic generator's hardware and software resource limitations are reached, or the resources of the Network Under Test are exhausted. In a load testing scenario the latter may actually be a desired event - provided that the aim is to discover the limitations of the system.

4.5.8 **Implementation**

In the control plane messages, the network-, service- and user-specific parameters are filled out as required. In practice, a relatively huge part of these messages never change: many parameters in call-setup, session or mobility management are set to exactly the same value. Network- and service-specific parameters (e.g., point codes, service capabilities, expected QoS parameters) have no or limited variance; user-related parameters (e.g., endpoint identifiers, temporary codes), on the other hand, vary a lot.

From this viewpoint it is possible to build a protocol message pool in which each kind of dialog is represented by a set of message templates. In these templates there are

- parameters filled out with *fixed values*,

- variable parameters, whose values are *chosen from a range* (based on certain rules),
- variable parameters *matching the protocol logic* (i.e., temporary identifiers, sequences, etc.).

In the user plane, individual dummy packets are generated as defined by the composite model. Actual packet lengths, and inter-arrival times are also derived from the composite model in use.

4.5.9 **Verification and refinement**

The general purpose of making a model is to give predictions. In our case, after the composite model is implemented, the traffic generator device is deployed and its output is matched against the earlier-captured real-life traffic.

In the control plane, the generated traffic needs to match the captured traffic as set out by the protocol standards. The statistical properties of subscriber mobility patterns need to match that of the captured patterns.

In the user plane, the statistical properties of the generated traffic need to match those of the captured data - provided that the same subscriber activity patterns (same traffic mix) are used. Until these requirements are matching, the model parameters need refinement through further traffic analysis and fine-tuning of the parameters.

4.5.10 **Deployment**

Network testing in an active way is carried out by simulating nodes that

- send control-plane protocol messages to the Network Under Test (NUT),
- keep track of the status of dialogs, transactions and contexts,
- handle user-plane traffic sent to/received from the NUT.

Depending on the type of network (segment) we actually deploy such a model-based traffic generator. The control plane protocols – including their internal logic – as well as the user plane protocols has to be made properly available in the traffic generator and tuned to the NUT.

4.6 Traffic analysis results

In this section, some results of MAVAR analysis over three different data sets of real IP traffic sequences, based on the procedure described in section 4.4, are presented.

The purpose of this analysis is to provide support for modelling FMC networks. Although analysis results are not available from such scenarios yet (since they are not deployed), aggregated traffic traces that are sourced from fixed as well as mobile segments can be analysed. Current traffic analysis aims at understanding whether the aggregated traffic is “averaging out” in volume, thus showing little self-similarity, or even the aggregated traffic is showing self-similar behaviour in various time-scales. The latter would mean that relatively complex models are suggested to be used for aggregated FMC traffic; and over-provisioning is suggested for the nodes involved.

4.6.1 Traffic measured by AITIA (10 Gb/s link SGSN - RNC pool)

The first sequence $\{x_k\}$ (bytes per time unit, [bytes/t.u.]) was derived by processing the traffic trace provided by AITIA, measured on a 10 Gbit/s link towards the Serving Gprs Support Node (SGSN) from the Radio Network Controller (RNC) pool. This traffic is highly aggregated indeed, including traffic of all RNCs connected to the SGSN.

Figure 31 presents the sequence $\{x_k\}$ of $N \cong$ 30 000 samples, acquired with time unit τ_0= 50 µs over a measurement interval $T \cong 1.5$ s. Figure 32 plots the histogram of the sequence. Figure 33 depicts the $\mathrm{Mod}\sigma_y^2(\tau)$ of $\{x_k\}$ in a log-log scale. Here, MAVAR exhibits a regular slope $\mu \cong$ -2,7115 estimated by least-square linear regression (excluding small and large values of τ). Therefore, the traffic trace seems to obey a simple power law [40] with $\alpha \cong 0.28$, which is a kind of LRD process with $H \cong 0.64$.

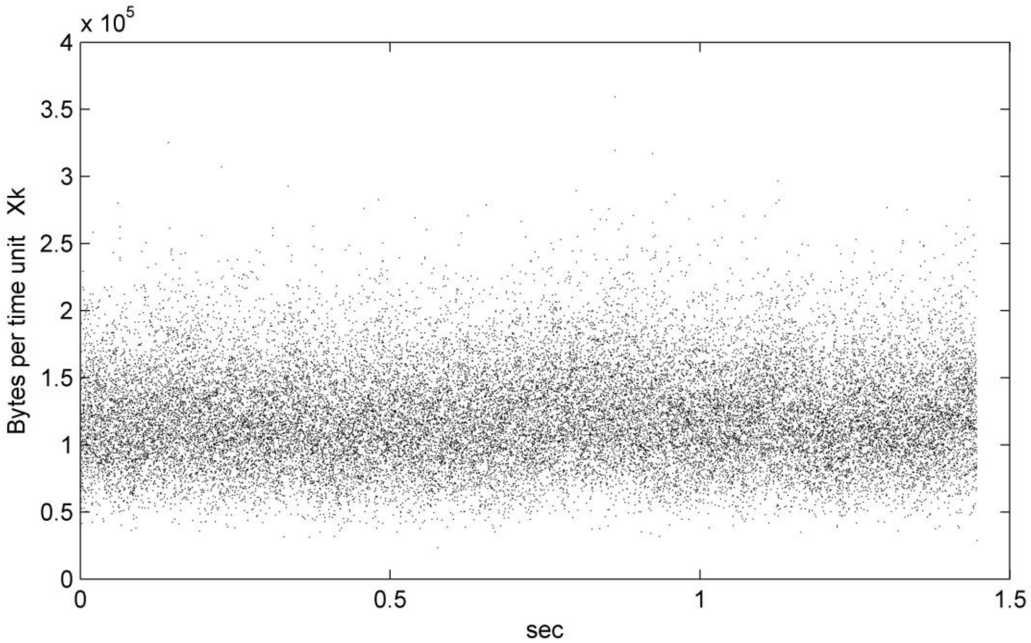


Figure 31: Traffic sequence $\{x_k\}$ AITIA 10 Gbit/s link (N = 30 000, τ_0 = 50 µs, T = 1.5 s).

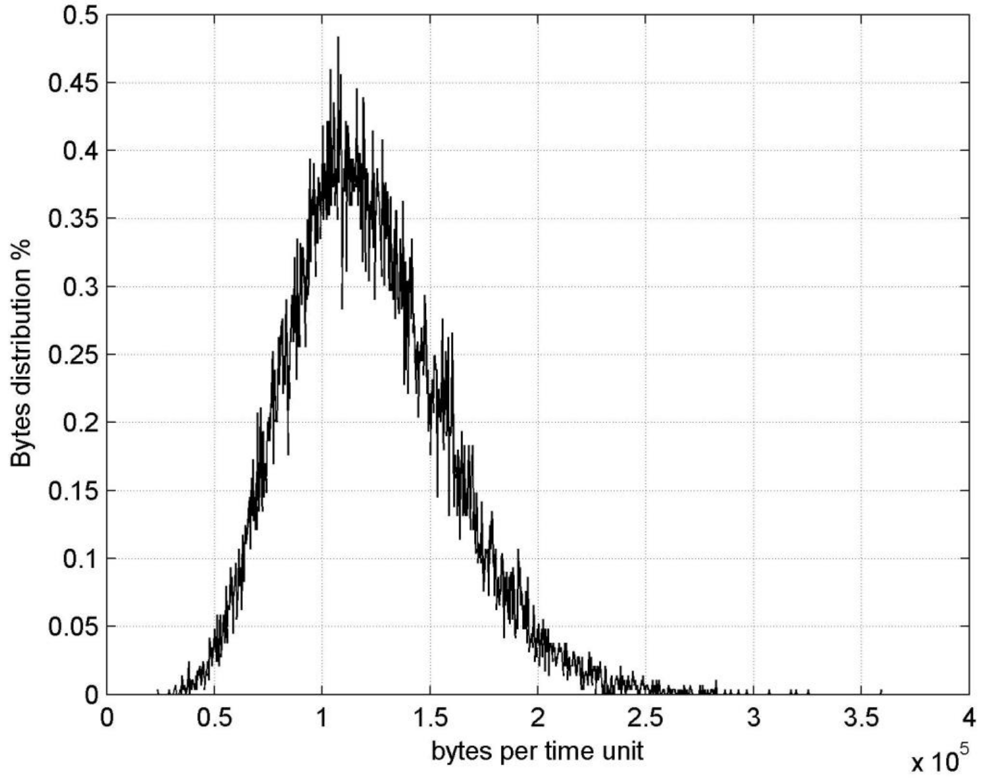


Figure 32: : Normalized histogram of {x_k} AITIA 10 Gbit/s link (N = 30 000, τ_0 = 50 µs, T = 1.5 s).

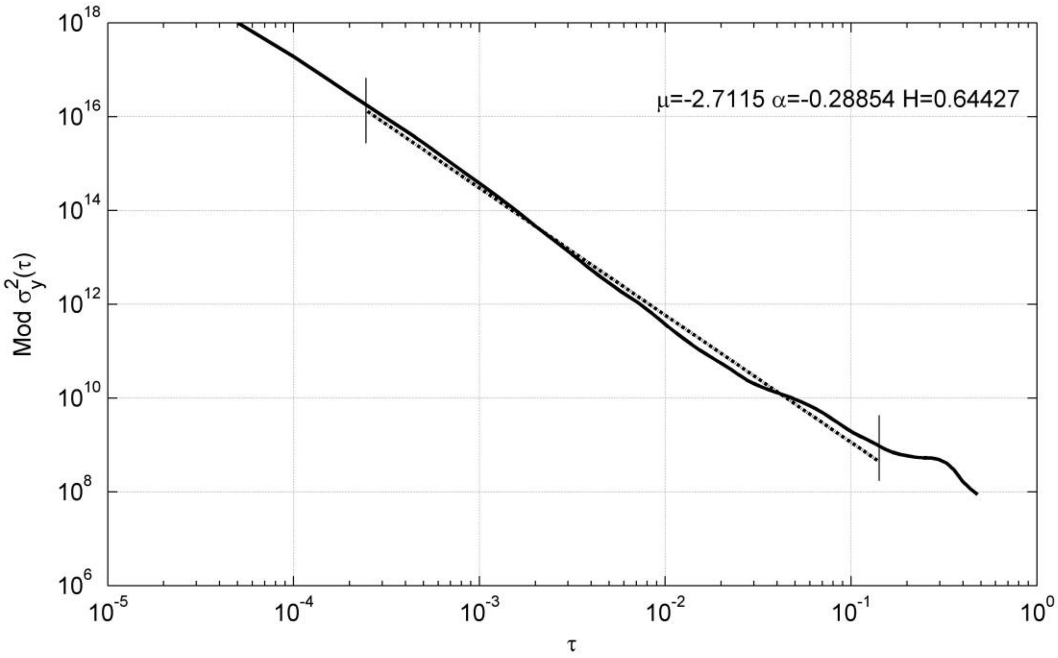


Figure 33: $\mathrm{Mod}\,\sigma_y^2(\tau)$ of traffic sequence {x_k} AITIA 10 Gbit/s link (N = 30 000, τ_0 = 50 μs, T = 1.5 s).

4.6.2 Traces from the CAIDA project repository (10 Gbit/s link Seattle-Chicago)

The second set of real IP traffic series {x_k} [bytes/t.u.] was captured on a 10 Gbit/s link connecting Seattle with Chicago in the CAIDA UCSD Project [63]. Both series are made of N=360 000 samples, acquired with time unit τ_0=10 ms over a measurement interval T=3600 s, between 11:59-12:58 in two different days (day 1: May 29, 2013 and day 2: June 20, 2013).

Figures 34 and 36 depict the traffic sequences $\{x_k\}$ [bytes/t.u.] of day 1 and day 2 respectively. In figure 36 at $t \cong 500$ s a step-like behaviour can be appreciated, however MAVAR is not significantly affected by such non-stationary trend [42] in estimating the Hurst parameter of random components. Figure 35 and 37 present the histogram of the sequences $\{x_k\}$ corresponding to day 1 and day 2, respectively.

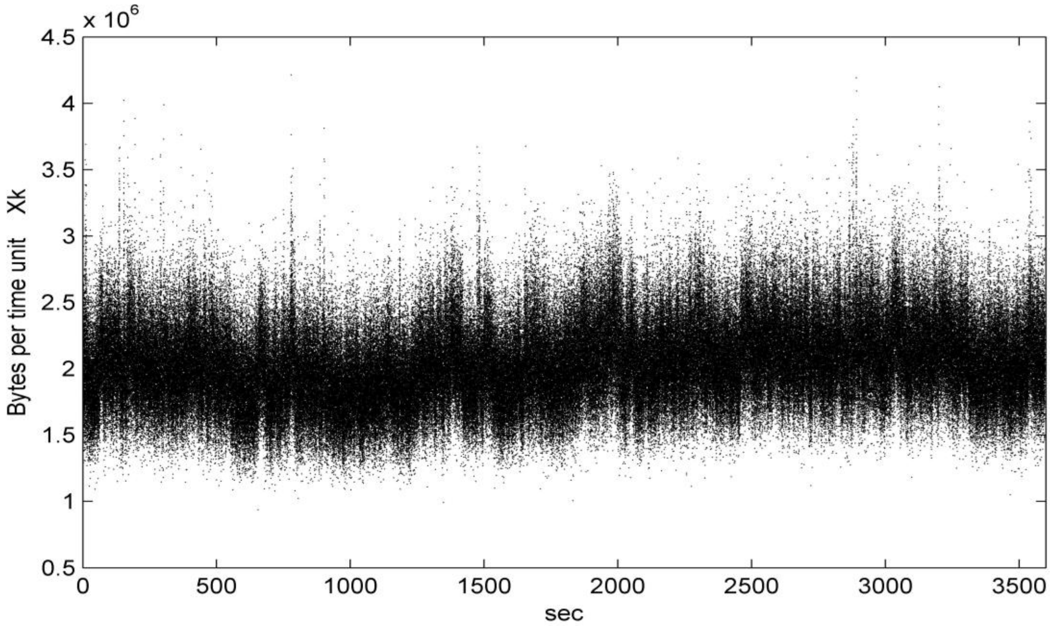


Figure 34: Traffic sequence $\{x_k\}$, CAIDA, 10 Gbit/s link Seattle-Chicago, day 1 (N = 360 000, τ_0 = 10 ms, T = 3600 s).

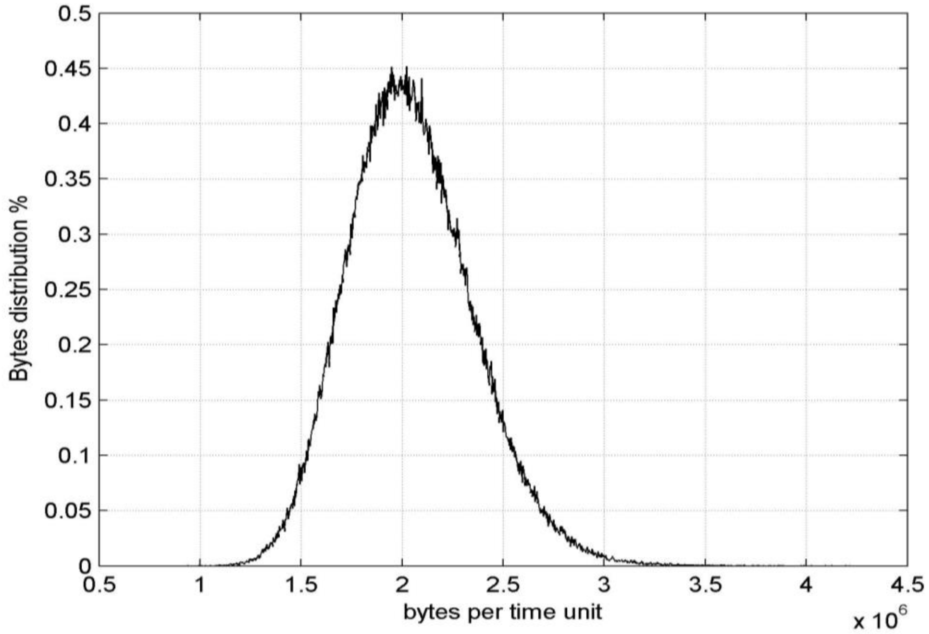


Figure 35: Normalized histogram of $\{x_k\}$, CAIDA, 10 Gbit/s link Seattle-Chicago, day 1 (N = 360 000, τ_0 = 10 ms, T = 3600 s).

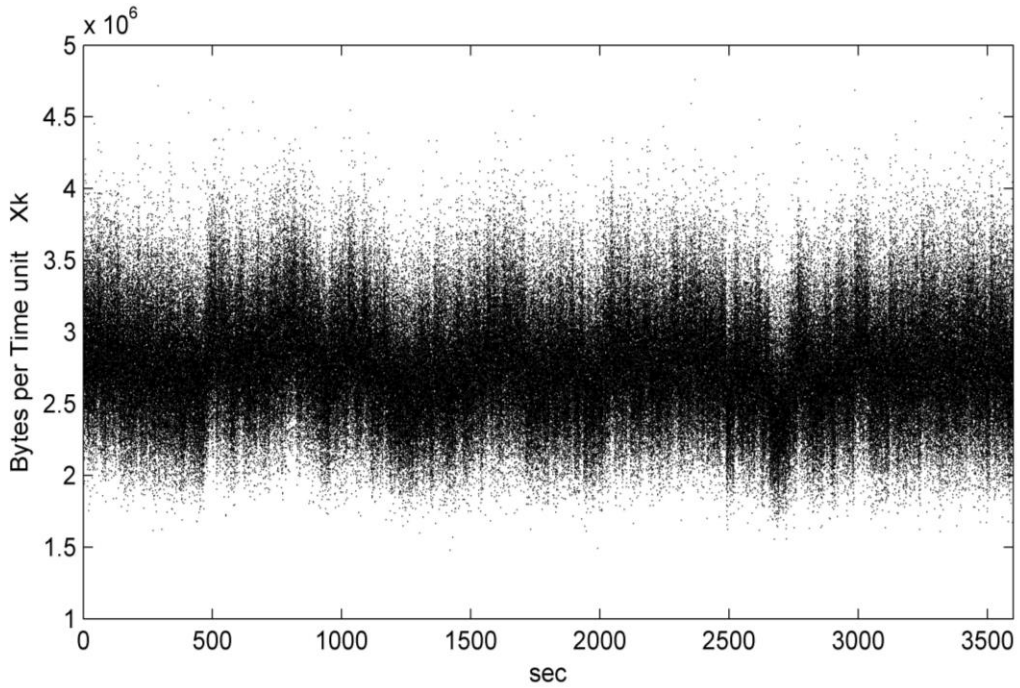


Figure 36: Traffic sequence $\{x_k\}$, CAIDA, 10 Gbit/s link Seattle-Chicago, day 2 (N = 360 000, τ_0 = 10 ms, T = 3600 s).

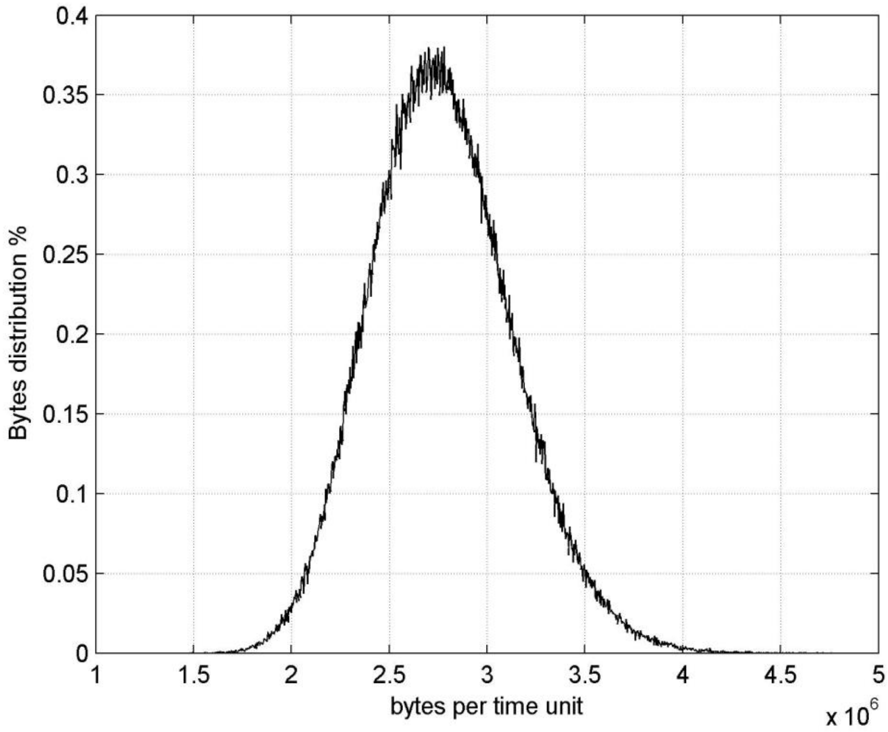


Figure 37: Normalized histogram of $\{x_k\}$, CAIDA, 10 Gbit/s link Seattle-Chicago, day 2 (N = 360 000, τ_0 = 10 ms, T = 3600 s).

In Figure 38 the MAVAR diagrams in log-log scale of the sequences $\{x_k\}$ of day 1 and day 2 are presented. Curves are almost identical in the two days, only with a slightly difference for the largest values of τ, were MAVAR loses confidence. Excluding the lowest and highest values of τ, a nearly straight line behaviour is observed in both sequences with slope μ_1 = -2.12 and μ_2 = -2.11 (obtained by least square regression) of day 1 and day 2 respectively. Therefore, the traffic traces seem to obey simple power laws [40] with $\alpha_1 \cong$ -0.88 for day 1 and $\alpha_2 \cong$ -0.89 for day 2, corresponding to LRD processes with $H_1 \cong$ 0.94 and $H_2 \cong$ 0.95.

Then, the 1h overall time span of the traffic trace CAIDA 10 Gbit/s link day 2 was divided in distinct subsequent time intervals T = {600 s, 100 s, 60 s, 10 s}. For each sweep of the overall time span, MAVAR was computed for each single segment of data (Figure 39 presents the specific case of day 2 using a sweep of 6 intervals of T = 600 s). A summary of the values of μ, α and H estimated by MAVAR diagrams are presented in Table 9.

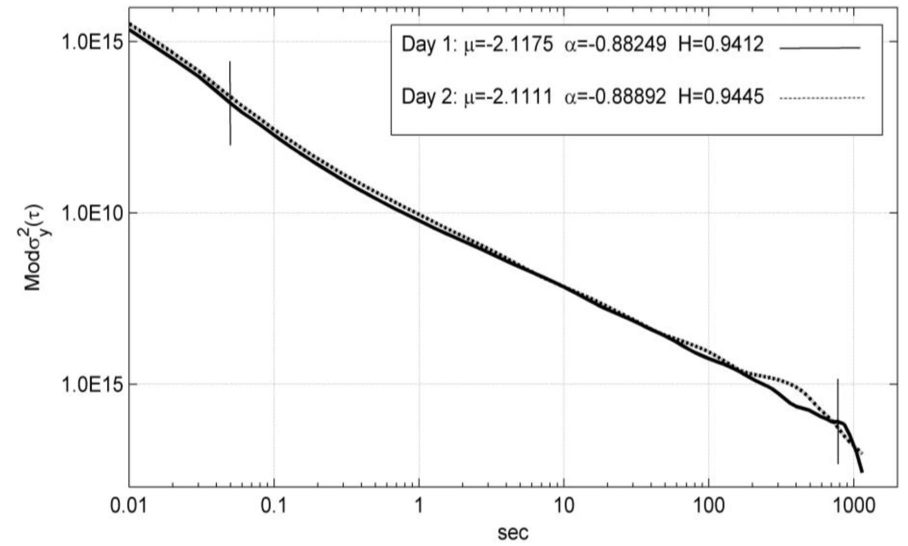


Figure 38: $\mathrm{Mod}\,\sigma_y^2(\tau)$ of sequences {x$_k$} , CAIDA, 10 Gbit/s link Seattle-Chicago, from day 1 and day 2 (N = 360 000, τ_0 =10 ms, T = 3600 s).

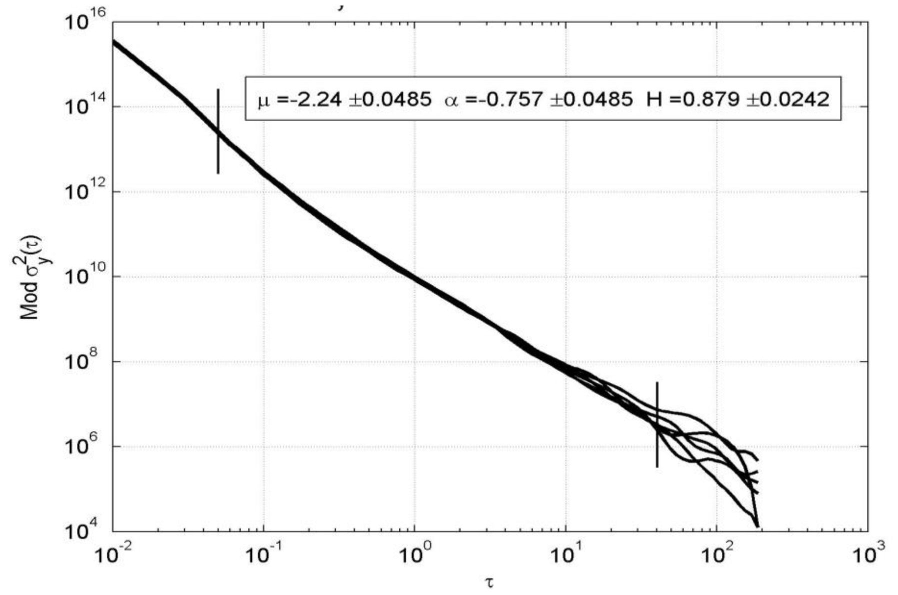


Figure 39: $\mathrm{Mod}\,\sigma_y^2(\tau)$ of 6 sequences {x$_k$} , CAIDA, 10 Gbit/s link Seattle-Chicago, day 2 (sweep of 1 hour 6 x T=600 s, N = 60 000, τ_0 = 10 ms).

T	*μ*	*α*	*H*
3600 s	-2.11	-0.889	0.941
600 s*	-2.24±0.0485	-0.757±0.0485	0.879±0.0242
100 s*	2.37±0.0631	-0.634±0.0631	0.817±0.0315
60 s*	-2.41±0.0899	-0.594±0.0899	0.797±0.0449
10 s*	-2.47±0.16	-0.264±0.16	0.632±0.0801

Table 9: Summary of values of μ, α and H estimated by MAVAR analysis of the CAIDA/Seattle-Chicago archive. *Values presented with mean and standard deviation.

4.6.3 Traces from the CAIDA Project Repository (OC48 Peering Link, T = 1h)

The third set of real IP traffic series $\{x_k\}$ [bytes/t.u.] was captured on a west coast OC48 peering link for a large ISP in the CAIDA Project 2002 [64]. The dataset is made of N=360000 samples, acquired with time unit τ_0=10 ms over a measurement interval T=3600 s, at time 09:00-09:59 on the 14 August 2002.The traffic trace seems to obey a simple power law [40] with $\alpha \cong$ -0,69, presenting LRD with $H \cong$ 0,85.

Then, the 1h overall time span of the traffic trace CAIDA 2,5 Gb/s peering link was divided in distinct subsequent time intervals T = {600 s, 100 s, 60 s, 10 s}. For each sweep of the overall time span, MAVAR was computed for each single segment of data. A summary of the values of μ, α and H estimated by MAVAR diagrams are presented in Table 10.

T	*μ*	*α*	*H*
3600 s	-2,301	-0,699	0,849
600 s*	-2,44±0,045	-0,562±0,045	0,781±0,023
100 s*	-2,6±0,071	-0,404±0,071	0,702±0,035
60 s*	-2,71±0,086	-0,294±0,086	0,647±0,0432
10 s*	-2,83±0,169	-0,174±0,169	0,587±0,084

Table 10: Summary of values of μ, α and H estimated by MAVAR analysis of CAIDA archives / OC48 Peering Link, T=1h. *Values presented with mean and standard deviation.

4.6.4 Traces from the CAIDA project repository (OC48 peering link, T = 3h)

The third set of real IP traffic series $\{x_k\}$ [bytes/t.u.] was captured on a west coast OC48 peering link for a large ISP in the CAIDA Project 2002 [64]. The dataset is made of $N = 360\,000$ samples, acquired with time unit $\tau_0 = 10$ ms over a measurement interval $T = 3600$ s, at time 09:00-09:59 on August 14, 2002.The traffic trace seems to obey a simple power law [40] with $\alpha \cong -0.69$, presenting LRD with $H \cong 0.85$.

Then, the 1h overall time span of the traffic trace CAIDA 2.5 Gbit/s peering link was divided in distinct subsequent time intervals $T = \{600$ s, 100 s, 60 s, 10 s$\}$. For each sweep of the overall time span, MAVAR was computed for each single segment of data. A summary of the values of μ, α and H estimated by MAVAR diagrams are presented in Table 11.

T	*μ*	*α*	*H*
10800 s	-2.009	-0.991	0.995
3600 s*	-2.17 ± 0.239	-0.829 ± 0.239	0.849 ± 0.12
1800 s*	-2.27 ± 0.221	-0.729 ± 0.221	0.865 ± 0.111
600 s*	-2.4 ± 0.149	-0.598 ± 0.149	0.799 ± 0.075
100 s*	-2.56 ± 0.113	-0.437 ± 0.113	0.718 ± 0.056
60 s*	-2.62 ± 0.117	-0.377 ± 0.117	0.689 ± 0.058
10 s*	-2.78 ± 0.184	-0.22 ± 0.184	0.61 ± 0.092

Table 11: Summary of values of μ, α and H estimated by MAVAR analysis of CAIDA archives / OC48 Peering Link, T = 3h. *Values presented as average ± standard deviation.

4.6.5 A comparison of α values estimated over time

In order to present how the estimated values of α depend on the time span of the traffic traces, Figure 40 plots the mean and standard deviation of the values of α estimated on the same traffic trace of length 3600 s, but on data subsegments for increasing values of T, as estimated by MAVAR diagrams. Similar analysis was carried out on all the previously introduced data sets – which lead to the same concluding remarks.

We see that, the longer is the time span of the data segment, the higher is the LRD that is detected (higher values of H, or, equivalently, lower values of α).

Moreover, in order to detect possible trends of α over time, plots of values of $\alpha(t)$ estimated by MAVAR on data subsegments of 60 seconds over three different traffic sequences $\{x_k\}$ are presented in Figure 41. Similar analysis was carried out on all the previously introduced data sets – which lead to the same concluding remarks.

In all cases, no peculiar trend is evident. Only some random variation, as expected. It would be interesting to investigate further on this direction, by estimating α (or equivalently *H*) over longer measurement time intervals, for example along days. As traffic average values may exhibit diurnal trends, also the LRD might exhibit slow periodic trends.

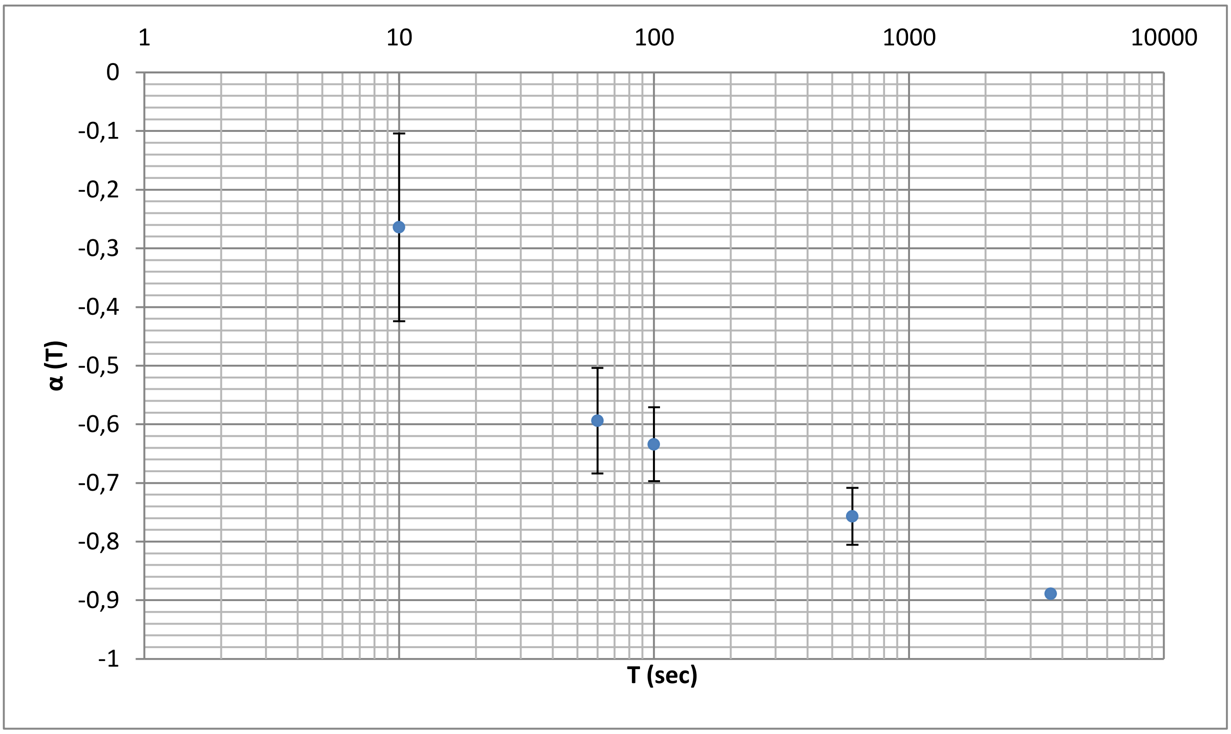


Figure 40: Plot of α estimated by the MAVAR analysis results over the 1 hour traffic sequence $\{x_k\}$ corresponding to the Seattle-Chicago 10 Gbit/s link, on day 2. (Mean and standard deviation for T < 3600 s). T in log scale

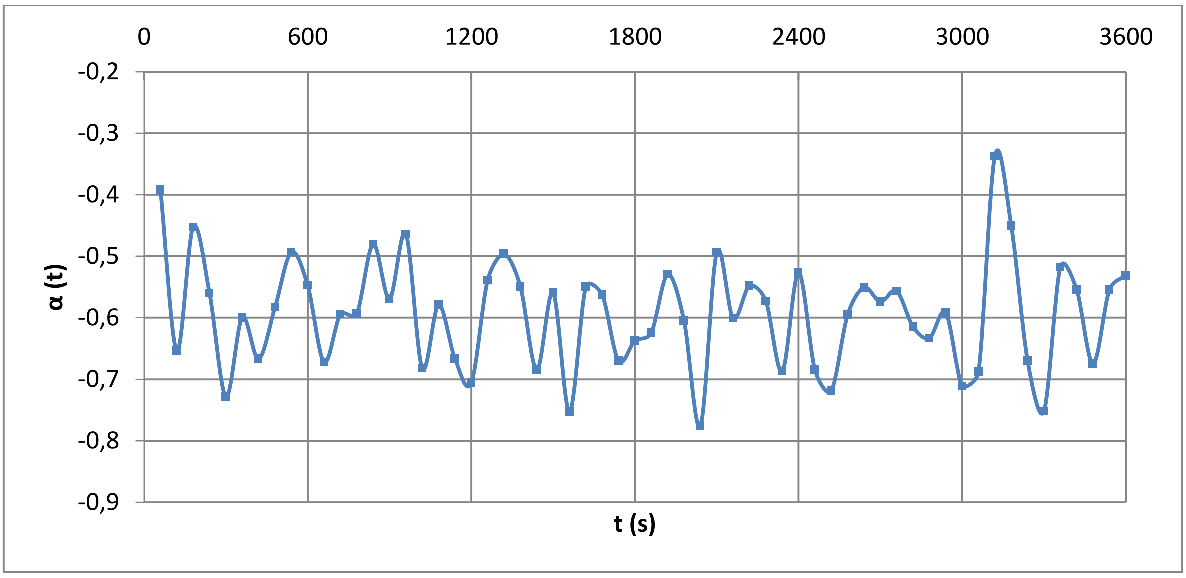


Figure 41. Plot of α(t) estimated from 60 computations of MAVAR every 60 seconds (T = 60s) over the 1 hour traffic sequence $\{x_k\}$ corresponding to the Seattle-Chicago 10 Gb/s link, on day 2.

4.6.6 Some conclusions about self-similarity analysis

The analysis results presented were computed on data measured

- on a 10 Gbit/s link towards the SGSN from the RNC pool;
- on a 10 Gbit/s link connecting Seattle with Chicago in the CAIDA UCSD Project;
- on a west coast OC48 (2.5 Gbit/s) peering link for a large ISP in the CAIDA Project 2002.

In all cases, traffic measured is highly aggregated from various sources, including mobile cellular networks and fixed networks. As a matter of fact, today there are no FMC networks yet available for traffic measurements. However, the measurement results provided represent various cases where fixed and mobile traffic are aggregated.

In all cases presented here, as well as in all others we studied but did not report here for the sake of brevity, significant values of self-similarity were measured, except on shortest time scales.

Even in the AITIA traffic data, which span only 1.5 s of measurement interval, the traffic proves to be not purely Poisson (with zero time-correlation). Here, the parameters measured have been $\alpha \cong 0.28$ and $H \cong 0.64$, rather far from the pure Poisson traffic where $\alpha = 0$.

In other traffic traces, which span longer time intervals up to a few hour, the traffic exhibits much higher self-similarity, with values of α ranging from -0.25 to -0.9 (and H from 0.6 to almost 1). Such values denote a strong self-similarity of traffic, under all cases of traffic aggregation considered.

Such high values of self-similarity call for significant over-provisioning of network resources (buffers, link capacity), compared to the ideal case where all traffic is supposed to be Poisson.

Interestingly, no significant slow trends in the values of α and H were detected. As further development, we believe that it would be interesting to analyse if such trends can be identified over periods spanning at least some days.

These results do affect FMC network planning - both on the node level and the network level. Those earlier assumptions about the aggregated traffic “average out” - hiding high bursts and peaks - cannot be supported, unfortunately. The main message here is that these high levels of self-similarity do call for over-provisioning.

On the other hand, when creating traffic models for nodes or network segments that handle mixed fix and mobile traffic, self-similarity should be considered as well. Experiments using pure Poisson traffic models for traffic volumes and arrivals would lead to modelling results not representing self-similarity. This means that the conclusions of such modelling would lead to fake safety for capacity planning.

4.7 How the targeted use cases affect traffic model selection

The targeted use cases are analysed in three main groups.

4.7.1 FMC access for mobile devices using Wi-Fi

In a user-initiated **mobile to Wi-Fi offloading** scenario, we assume that not all services are equally likely to be offloaded. Bandwidth-demanding connections are more likely to be offloaded to the Wi-Fi/fixed interface. On the other hand, for voice and audio communication, users may prefer connections with guaranteed QoS parameters such as low delay variation. Moreover, available bandwidth may actually encourage subscribers to start using services they would not use otherwise.

At this point we do not know, to what extent handover affects quality of experience. If an offloading procedure is noticeable or even annoying to the users, then that would deter them from voluntarily initiating an offloading procedure.

We also need to consider a specific situation which leads to **oscillating user behaviour**: When the QoS starts to deteriorate, the user initiates a handover to another network; when he or she finds that using the other interface does not give satisfactory results, after a while he/she switches back to the original interface to see if the connection improved meanwhile. This may go on for an extended period of time. Such behaviour could be circumvented if the user device regularly polled the available interfaces for QoS parameters and reports them to the user. Interface selection can also be initiated by the network if such parameters are available.

As for the traffic mix of the dual attachment, we anticipate that the combined traffic mix is not necessarily a linear combination of the mobile and Wi-Fi traffic patterns. We can equally anticipate increased traffic volume due to increased overhead on the control plane, as well as decreased overall traffic volume because the users need to share their attention among multiple activities – made possible by dual attachment. The actual traffic mix need to be estimated from observations from FMC pilot systems and early deployment.

4.7.2 Large traffic variations between public, residential and business areas

The traffic models need to take into account that different services are used in the public, residential and business areas (e.g. more social networking in public and residential, while less in business areas). The time of the day also affects the types of services used (e.g., video on demand or online gaming more in the evening hours).

We expect to observe larger traffic variation in the dynamically switched access/aggregation network, whereas less traffic variation in the core network, because the traffic variations among the access networks are evened out in the core. In order to achieve a realistic model of the network, the traffic mix needs to be measured multiple times and separately in the aggregation network and the core network.

At this initial phase of the project, we cannot anticipate if user experience on e.g. the day/night reconfiguration has an effect on the user preference of using particular services. For example, if an employee needs to stay in the office late in order to have a VoIP call with an overseas customer, would network reconfiguration affect the QoS of the network? What would be the employee's quality of experience?

Would automated tasks such as software updates, which are normally scheduled for the night/weekend, be affected by the reconfiguration schemes? If yes, then our models would also need to give account of such changes in the modelled traffic patterns.

4.7.3 Converged access and aggregation network cases

In a converged scenario we anticipate the widespread deployment of small cells and nanocells. Such cells differ in their traffic load characteristics from larger cells: smaller cells assume fewer users in the cell at a given time, which in turn determines the distribution of active services and offered traffic.

Smaller cells imply that more of them would be needed. This results in the increased number of control messages in the core network for procedures such as paging, context management or handover. That is, control message statistics also determined by the network architecture.

5 Evolution of traffic growth considering FMC scenarios

5.1 Offloading scenarios with WLAN femtocells

5.1.1 Introduction

Nowadays there is enormous growth of the number of a new generation of mobile devices like various smart phones (e.g., iPhone, Android-based, etc.), laptops, netbooks, etc. in the market. At the same time, mobile networks operators are incorporating actively Internet applications and services for the mobile devices. There are thousands of web data applications and services available now (e.g., YouTube, Facebook, Spotify, IM, mobile TV, etc.) that are becoming extremely popular in the mobile user environment. According to the Cisco VNI Global Mobile Data Traffic Forecast [65], overall mobile traffic is expected double every year from 2013 onwards.

As a result of these two factors, there is an explosion of both data and signalling traffic towards the core network of Mobile Network Operators (MNOs). As a consequence, congestion situations can arise in the network. Thus, solutions to avoid unnecessary traffic load at network nodes are needed. One such solution is to apply a traffic offloading mechanism by means of femtocells or WLAN. It can solve macro core network capacity crunch avoiding future upgrades of the network infrastructure. The factors motivating the offloading process in the network of the MNOs are mentioned in Figure 42.

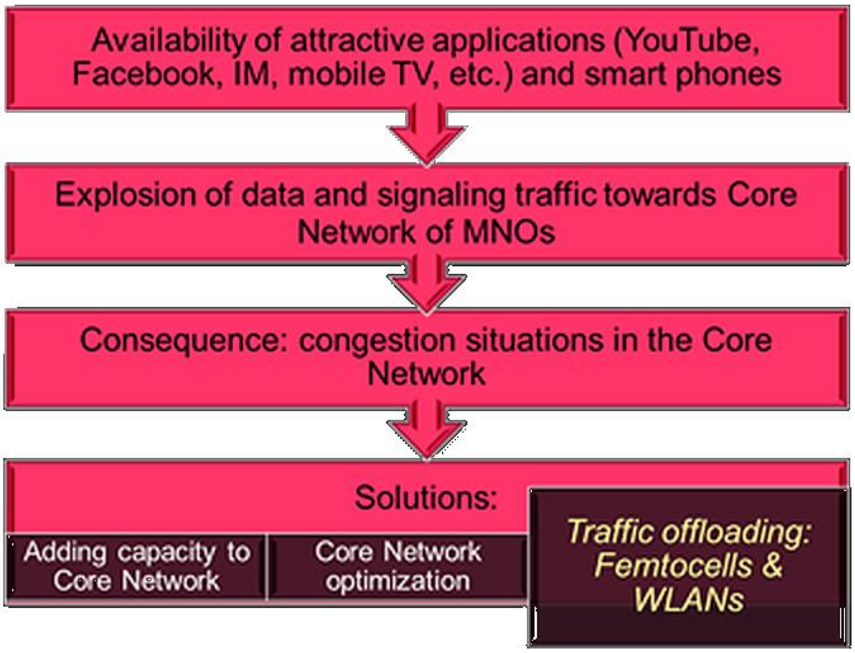


Figure 42: Factors motivating offloading.

It is interesting to consider the traffic offloading process from the viewpoint of a MNO perspective. In this context it is important for the MNO to care about a traffic load that goes to its core network (e.g., Evolved Packet Core, EPC) after implementing offloading technique. What is needed for the MNO is to know what was before offloading and what happens after in the context of traffic parameters to adjust with them dimensioning characteristics of its network, e.g., to evaluate required system capacity.

A view of a network implementing offloading is presented in Figure 43. It contains some offloading areas within the MNO network coverage. Thus, with the advent of femtocells, WLANs or Integrated Femto-WLAN solutions [69], the network of the operator could be offloaded when the user is at home or at an enterprise, as her/his traffic would be routed through the corresponding alternative ways. When a user is out of offloading areas traffic goes by traditional way through the macro network of the operator. However, definitely there will be also users who prefer for some reasons to have connections to network services over macro network even if they are inside offloading coverage.

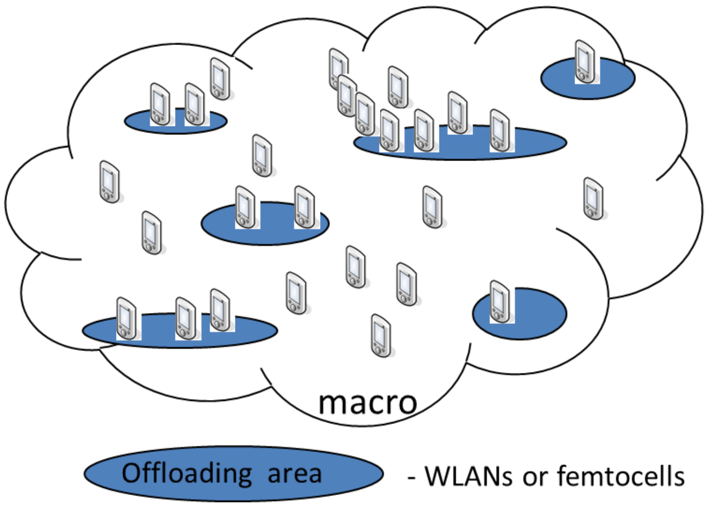


Figure 43: A view of a network implementing offloading.

To describe performance of mobile networks implementing offloading an appropriate analytical model is needed. In the next subsection we propose a model for traffic of a single source that eventually reaches the network of the operator after offloading.

5.1.2 Traffic model for offloading

In the same way as in previous literature, the activity of a single source can be modelled as a strictly alternating ON/OFF process [70]. In the model, we introduce the effect of offloading over such source by means of an additional strictly alternating ON/OFF process. The resulting traffic sent to the network of the MNO in a regular way (i.e., the non-offloaded traffic) from a single source is then modelled as the product of the two above processes. Therefore, the aggregation of several such resulting processes are still needed to be handled by the operator in the conventional way (the one to be used for the purposes of network dimensioning). Below we introduce the details and notation of all above processes.

Model of user activity

The activity of a single source/user is modelled as a strictly alternating ON/OFF process, where ON periods are independent and identically distributed (i.i.d.), OFF periods are i.i.d., and ON and OFF periods are mutually independent. Furthermore, previous measurements have also shown that such periods follow heavy-tailed distributions (e.g., due to file size distributions, web pages) [71] (and references therein). This is also assumed in our model.

Therefore, the process Y (t) is a stationary binary time series {Y(t), t ≥ 0} such that [70]:

$$Y(t) = \begin{cases} 1 & \text{activity period} \\ 0 & \text{idle period} \end{cases} \qquad (6)$$

A representation of such a process is presented in Figure 44. During the activity period, the user is transmitting and receiving packets. On the other hand, no traffic is exchanged with any network during idle periods, e.g., user reading time of downloaded content. Notice that Y(t) only describes source/user behaviour and is independent from the network through which the traffic is sent, i.e., no offloading considerations have been made in this process.

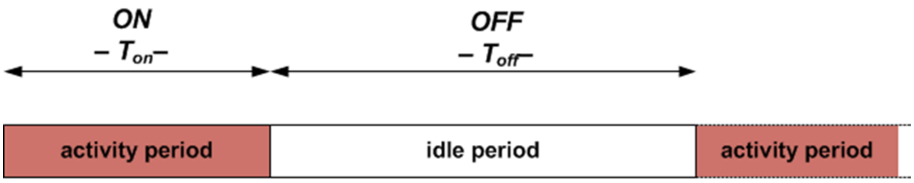


Figure 44: User activity is modelled as a strictly alternating ON/OFF process, Y(t).

Model of offloading periods

Previous measurements carried out in real networks have shown that smartphone connection and disconnection periods to offloading areas follow heavy-tailed distributions [75]. This observation led us to also model offloading periods for a single source as strictly alternating ON/OFF process. In the same way as before, we also assume that ON periods are i.i.d., OFF periods are i.i.d., and ON and OFF periods are independent. Therefore, the process X(t) is a stationary binary time series {X(t), t ≥ 0} such that

$$X(t) = \begin{cases} 1 & \text{user flow sent through MNO network} \\ 0 & \text{user flow sent via offloading network} \end{cases} \qquad (7)$$

A representation of such a process is presented in Figure 45. During ON periods, the traffic generated by the source (if any) would be sent towards the network of the MNO in a regular way (i.e., traffic is not offloaded). On the other hand, during OFF periods all traffic generated by the source would be routed through the offloading network (e.g., smartphone under the coverage of a femtocell or Wi-Fi).

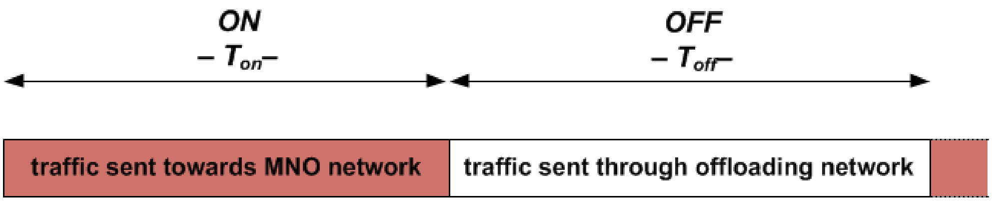


Figure 45: Offloading periods are modelled as a strictly alternating ON/OFF process, X(t).

Model of non-offloaded traffic from a single source

The traffic generated by a single source that is treated in a conventional way by the MNO (i.e., non-offloaded traffic) can be modelled as the product of the previously defined processes. That is,

$$Z(t)=X(t)Y(t) \quad (8).$$

Figure 46 represents such a process, which is also a strictly alternating ON/OFF process, with ON and OFF periods following heavy-tailed distributions and whose characteristic parameters can be derived from those of the original ones. During ON periods, the traffic being generated by the source (i.e., user activity in ON state) is forwarded to the network of the MNO as usual. On the other hand, during OFF periods, either there is no activity from the source or traffic is being sent through the offloading network (e.g., through Wi-Fi, femtocells).

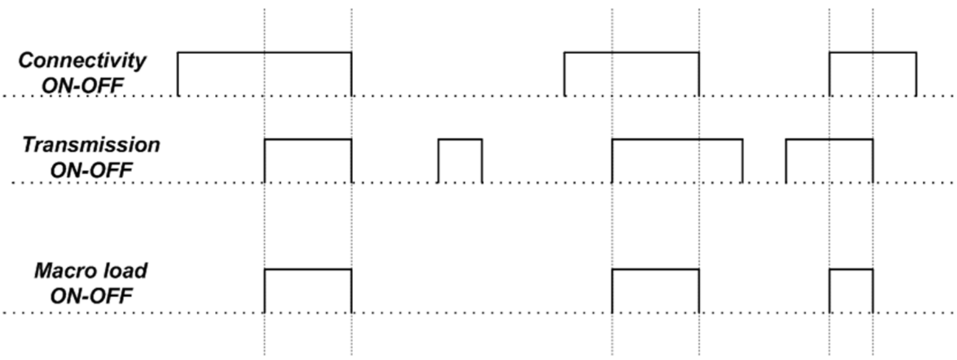


Figure 46: Non-offloaded traffic from a single source is modelled
as the product of two strictly alternating ON/OFF processes, Z(t)=X(t)Y(t).

High-level view of the modelling procedure

A high level view of steps for the modelling procedure is as follows. Since, as explained above, Z(t) has ON and OFF periods that follow heavy tailed durations, this process is long-range dependent. Therefore, the aggregation of several such processes is self-similar and can be characterized by means of the Hurst parameter (H). This can be done by means of the same techniques presented in [70]. In turn, parameter H can be derived from the parameters characterizing the heavy-tailed behaviour of the ON and OFF periods of Z(t). Besides, such parameters can be obtained from the original processes X(t) and Y(t). Therefore, once the parameters of the original processes are known, by following the above steps, we can characterize the behaviour of the aggregated non-offloaded traffic, and hence, the resources needed in the network of the MNO to serve it.

5.1.3 Applicability of the model

There are various ways in which offloading could be deployed in a network. Currently, the most popular one is the use of Wi-Fi (instead of 3G), when available, for all the traffic exchanged by the user [72]. With the advent of femtocells, some parts of the network of the operator could also be offloaded when the user is at home or at an enterprise, as her/his traffic would be routed through the femtocell. The proposed model is focused on the traffic that eventually reaches the network of the operator in a regular way (i.e., the non-offloaded traffic). That is, the MNO can compare data traffic (its volume and characteristics) that went through the network before and after implementing offloading. Then, it is possible to evaluate the benefits that offloading generates in the network of the MNO in terms of resource consumption (network capacity). In this sense, the traffic pattern generated towards the network of the MNO in the two above scenarios (i.e., Wi-Fi or femtocells) is the same. That is, all traffic generated by the user when outside coverage of a Wi-Fi access point (or a femtocell) is sent to the network of the MNO as usual. On the other hand, traffic is offloaded when under the coverage of the access point. If a user within the offloading coverage still prefers to use macro connectivity the traffic is treated as non-offloaded. Therefore, the model captures this traffic pattern disregarding the technology in use for offloading, and in this sense, it is agnostic from the specific technology in use.

There are many offloading techniques (e.g., LIPA, SIPTO, I-WLAN, MAPCON, IFOM [72],[73],[74]) that are being defined by 3GPP nowadays. In some techniques offloading (e.g. SIPTO) is done on a PDN connectivity basis, other techniques (e.g. IFOM) realise offloading on a per-flow basis, i.e., some flows are offloaded and some other flows are not when under the coverage of a Wi-Fi access point or a femtocell. In a more generic case, there may be intermediate nodes in which offloading decisions are taken (e.g., the gateway of a company towards the network of the MNO). However, in terms of traffic pattern generated towards the network of the MNO, the resulting flow would behave in the same way as in previous cases. Therefore, the proposed model might still be used for those flows that are offloaded, no matter where the offloading point is in the network (e.g., terminal, femtocell/access point, intermediate node). Hence, the total aggregated traffic toward the network of the MNO would result from the aggregation of offloaded flows (output of the model) with regular flows that are not offloaded at all.

5.2 Signaling traffic evolution with control and user plane decoupling

5.2.1 Towards a full decoupling of control signalling from user data traffic

Since it is supposed that in the proposed FMC reference architecture the commonly distributed elements of mobile core network (e.g., S-GW, PDN-GW) will be separated from the commonly centralised elements of the mobile core (e.g., MME, PCRF), the idea is to consider in this section some design aspects of the mobile network architecture to decouple *completely* signalling operation from user plane functionality and estimate how many signalling changes it requires on the control plane level to manage connectivity and mobility procedures.

One of the EPS architecture aims is to introduce a clear separation between control and user plane operation. Control plane elements, which are mostly decoupled from the user plane path, handle authentication, privacy, QoS and mobility functions. The decoupling is motivated by several factors [21].

One of these factors is that control data signalling tends to scale with the number of users, while user data volumes may scale more dependent on new services and applications. The second factor is that the full separation of control signalling and data operation allows more flexibility in handling user data functions in a more distributed way in the networks, while at the same time allowing for a centralised deployment of the equipment handling the control signalling [21].

Moreover, without full decoupling of control and user planes, multihoming support that is required for deployment of advanced 3GPP concepts where multiple interfaces are used simultaneously (e.g., IFOM, MAPCON, S1-flex [21], [22], [23]) is an issue. It requires the duplication of user plane operation that should be coordinated appropriately with control plane functionality. If the decoupling between two planes is not designed in its full extension, the complexity of the process is significantly increased.

The decoupling between the two planes is not realized completely up to now even though there is significant progress in this direction. In particular, when observing the Control Plane structure of the EPS network one can notice that the architecture builds around the tandem MME/S-GW that is located at the centre of the whole system. While MME only has control plane functionality, the S-GW shares both control and user plane responsibilities. In particular, the S-GW acts as a transit point for the signalling exchange between the MME and the PDN-GW. Interestingly, while most control layer decisions have been decoupled from user plane elements, the responsibility to disseminate them relies on specific interfaces between the PDN-GW and S-GW. In this sense, the future evolution path should be done towards a full separation of control signalling from user data traffic.

Thus, the basic proposal here is to move all control signalling responsibilities of the S-GW to the control plane (the MME in particular) and, as a consequence, establish additional interfaces between control and user plane elements. Specifically, a design principle that it is worthwhile to propose is to establish an interface between the MME and the PDN-GW that substitutes the role of S5/S8 interface for control plane information (see Figure 47).

As a consequence the MME acquires the role of being the sole responsible for the establishment and maintenance of the user plane path during the attach process and during mobility events. To achieve this goal there is the need to define an additional interface between the MME and the PDN-GW that will be used to exchange all those signalling messages that are currently exchanged through a combination of S11 and S5/S8 interfaces. Even more, the S11 interface needs to be extended to accommodate those signalling messages that are directly exchanged between the PDN-GW and the S-GW.

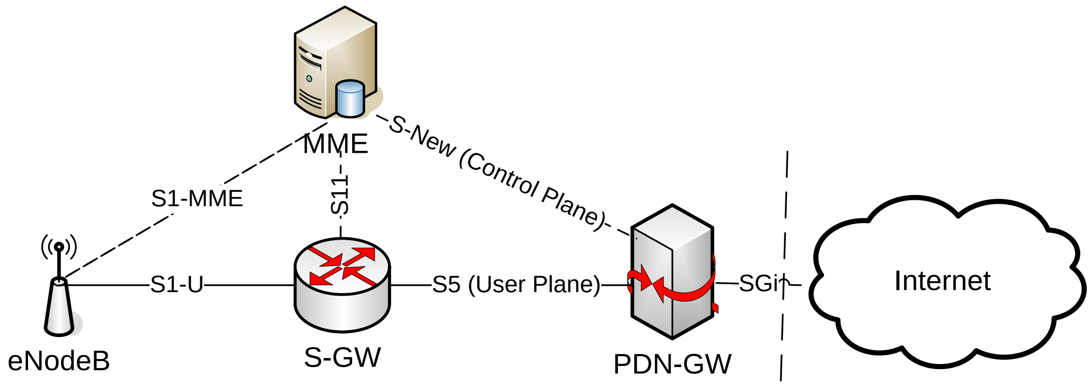


Figure 47: New interface for full decoupling of control signalling from the user data traffic.

With a complete decoupling, the formation of multiple user plane paths becomes simpler (among other advantages it can bring), as it is reduced to informing the certain user plane nodes (e.g., the eNodeBs, the PDN-GW) about the multiple options that they have to forward data by the special entity (MME) that handles all control plane functionality as illustrated in Figure 48.

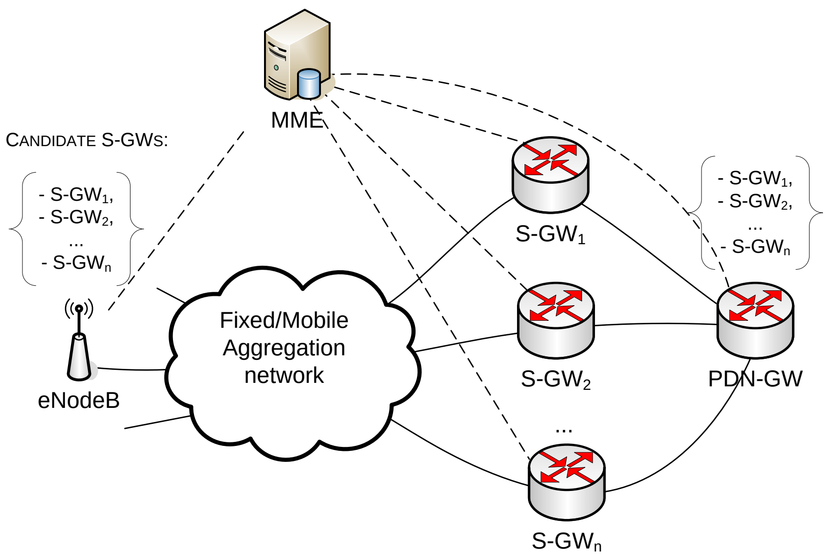


Figure 48: Multiple data path formation with full decoupling of control signalling from the user data traffic.

It could be realized by two ways. As the first way, the MME can send to the eNodeB a signalling message (e.g., Initial Context Setup Request/Attach Accept [24]) including addresses of several candidate S-GWs for user plane instead of one as it is currently realized. As the second way, the MME can send to the eNodeB the same signalling message, but several times, with a different S-GW address. Similarly, the PDN-GW can receive the same information about candidate S-GWs, directly from the MME responsible of this particular User-Plane path.

The feasibility to move the messages exchanged between the PDN-GW and the MME in the current EPS architecture to the new proposed interface is analysed in next Section.

5.2.2 Analysis of signalling changes related to a full decoupling of control and user planes

This section analyses the feasibility of full separation of control signalling operation from the user plane functionality when managing connectivity and mobility procedures. In particular, we review how the singling messages currently exchanged between the MME and the PDN-GW over the S-GW (using S11 and S5 interfaces) can be redistributed taking into account the new interface called conditionally S-New to release the S-GW from control responsibilities.

Within the scope of this analysis, we review two basic EPS procedures [24], namely, the attach procedure and the X2-based handover (without S-GW relocation).

The attach procedure with the S-New interface

Figure 49 presents a flow diagram with control messages related to default bearer establishment within the attach procedure when there is the S-New interface between the MME and the PDN-GW.

As one can see from the diagram, the Create Session Request message (initiated by the MME) that is currently released between the S-GW and the PDN-GW (crossed-out lines in Figure 49) can be sent directly between the MME and the PDN-GW (dash-and-dot line). Among other parameters the message can inform the PDN-GW about MME's Tunnel Endpoint Identifier (TEID) for sending control messages on the S-New interface towards the MME.

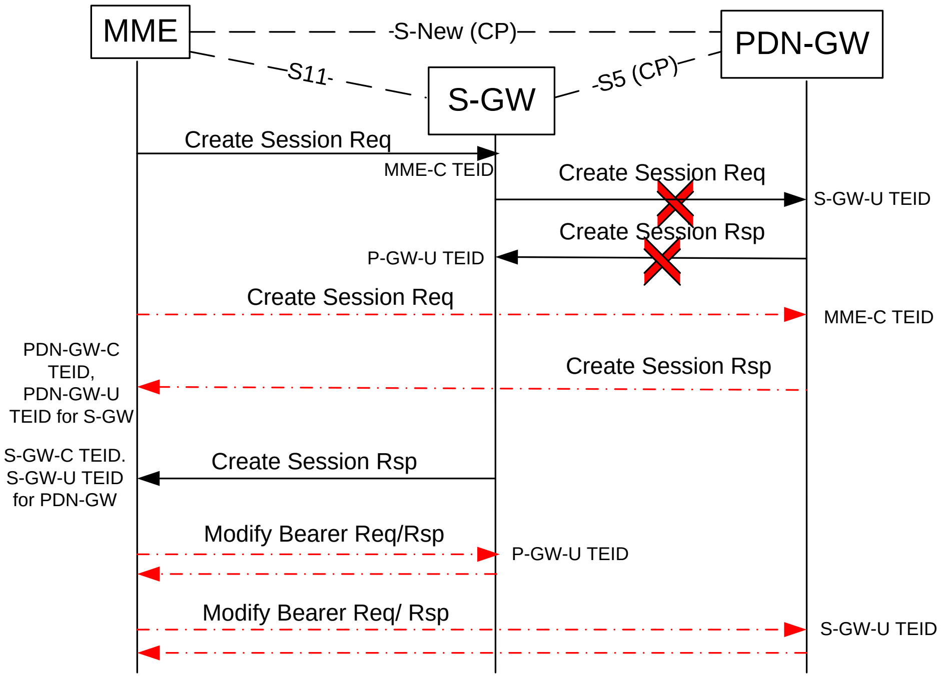


Figure 49: Default bearer establishment with the S-New interface.

Once the PDN-GW receives this request, it sends the Create Session Response with some parameters including the PDN-GW's TEID for sending control messages on the S-New interface from the MME to the PDN-GW. Besides, the message contains the PDN-GW's TEID for user plane. In turn, with the Create Session Response the S-GW informs MME about the S-GW's TEID for user plane.

Now the MME is aware of the PDN-GW's TEID and S-GW's TEID for user plane and it should forward this information to the S-GW and the PDN-GW for sending PDUs on S5. It can be realized, for instance, by means of sending the Modify Bearer Request (or Create Session Request as an option) from the MME to the S-GW and from the MME to the PDN-GW with the PDN-GW's TEID and S-GW's TEID, respectively. The S-GW and the PDN-GW should acknowledge by sending the corresponding response messages to the MME. Then, the S-GW and the PDN-GW can exchange data with each other in both directions.

Therefore, for default bearer establishment with the S-New interface we can release two signalling messages from S5 interface, but we need to add four messages to the S-New interface and two messages to the S11 interface.

Figure 50 presents a flow diagram with control messages related to dedicated bearer activation within the attach procedure when there is the S-New interface between the MME and the PDN-GW.

With the S-New interface, the PDN-GW can send the Create Bearer Request message directly to the MME without the S-GW involvement. The message should contain the PDN-GW's TEID for user plane. The MME forwards this information to the S-GW.

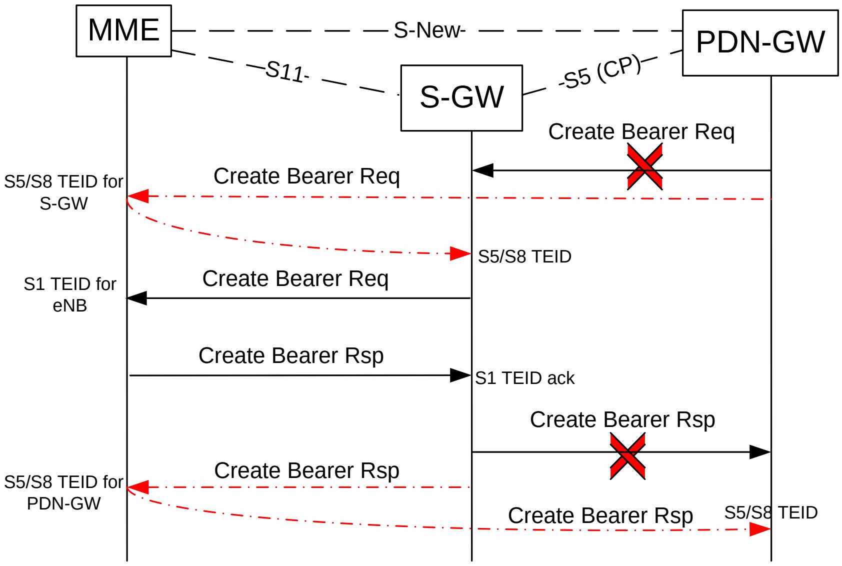


Figure 50: Dedicated bearer establishment with the S-New interface.

Once the S-GW receives this message, it sends the Create Bearer Response to the MME with the S-GW's TEID for user plane. The MME should propagate this message towards the PDN-GW.

As a result, for the dedicated bearer activation process with the S-New interface we can release two signalling messages from S5 interface, but instead, we need to add two messages to the S-New interface and extend the S11 interface by the two messages.

Figure 51 presents a flow diagram with control messages related to the bearer modification procedure when there is the S-New interface between the MME and the PDN-GW.

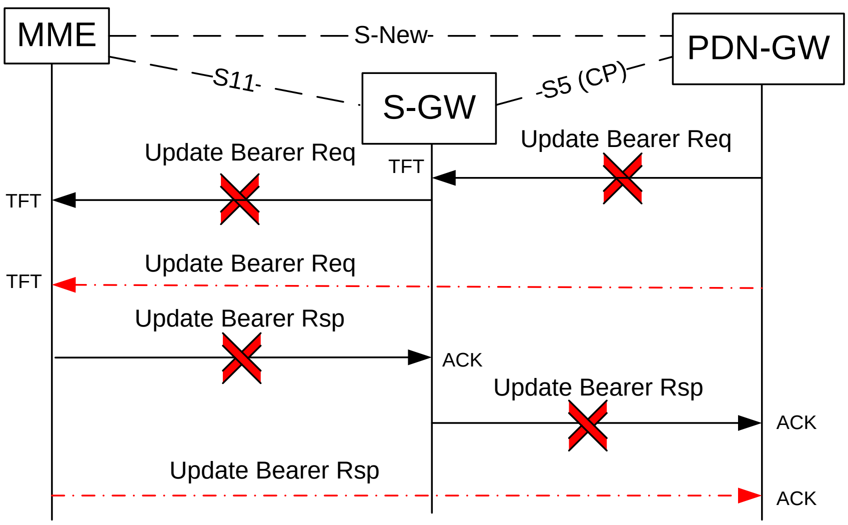


Figure 51: Bearer modification procedure with the S-New interface.

The PDN-GW can send the Update Bearer Request with the generated Traffic Flow Template (TFT) and updated QoS parameters directly to the MME without the use of the S-GW as a middle point. Upon reception of the message the MME issues the Update Bearer Response acknowledging the bearer modification.

Thus, for the bearer modification process with the S-New interface we can release two signalling messages from S5 interface and two signalling messages from the S11 interface adding two messages to the S-New interface.

Handover with the S-New interface

Figure 52 presents a flow diagram with control messages related to X2-based handover for the proposed architecture with the S-New interface between the MME and the PDN-GW.

With the S-New interface, the MME can send the Modify Bearer Request message simultaneously to the S-GW and the PDN-GW with different context information. The message sent to the S-GW contains the eNodeB address and TEID for downlink user plane. The Modify Bearer Request sent to the PDN-GW contains the information required by the PDN-GW, for instance, user location IE, UE Time Zone IE, serving network IE, etc. [24].

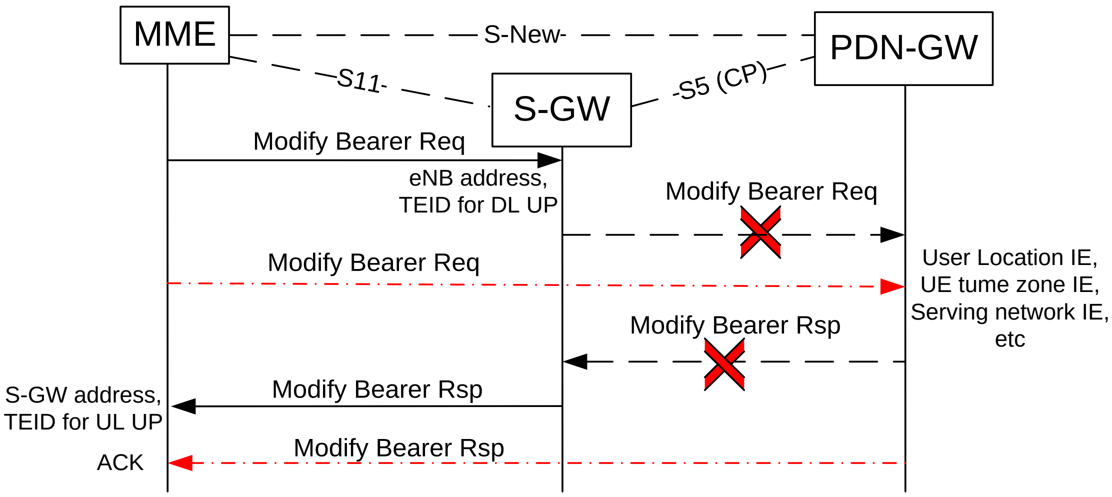


Figure 52: X2-based handover with the S-New interface.

The S-GW returns a Modify Bearer Response to the MME with the S-GW address and TEID for uplink user plane. The PDN-GW acknowledges the MME with the Modify Bearer Response. Thus, for the handover process we can release two signalling messages from S5 interface moving them to the S-New one.

Control messages reallocation in the new architecture

A summary of control messages reallocation related to the attach and mobile procedures considered above is presented in Table 12.

Procedure	Signalling messages		
	S-New interface	*S11 (extension)*	*S5 (Released)*
Default bearer establishment	Create Session Req/Rsp, Modify Bearer Req/Rsp	Modify Bearer Req/Rsp	Create Session Req/Rsp
Dedicated bearer activation	Create Session Req/Rsp	Create Session Req/Rsp	Create Bearer Req/Rsp
Bearer modification	Update Bearer Req/ Rsp	-	Update Bearer Req/Rsp, Update Bearer Req/Rsp (S11 released)
X2-based handover	Modify Bearer Req/ Rsp	-	Modify Bearer Req/Rsp

Table 12: Signalling reallocation to release S5 (Control Plane)

The procedures can be supported by means of regular signalling message exchange through the new proposed interface with some extension of S11 interface functionality to release the S5 interface from control plane (CP) operation. For instance, in the considered procedures (see Table 12), ten signalling messages are sent through the new interface and four additional messages are sent through the S11 interface (from the MME to the S-GW) to release ten control messages that are currently exchanged through a combination of S5 and S11 interfaces.

Thus, we have considered some high-level aspects related to a network architecture design enhancement towards a full decoupling of control signalling from user data traffic and signalling changes it can require. In particular, we revealed how signalling exchange related to both the attach and mobility procedures can be supported by means of the new interface between the MME and the PDN-GW with some extension of the functionality provided through S11 interface. As a result of the analysis, we conclude that much less changes than one could initially have expected are needed to apply the proposed design principle for network architecture evolution. It enables a basis for effective deployment of advanced 3GPP concepts related to multihoming and offloading scenarios (e.g., S1-flex, IFOM, MAPCON [21], [22], [23], [27]).

5.2.3 Summary on decoupling control and user plane signalling in the mobile architecture

The goal of the section, as stated at the beginning, is to consider design aspects of the mobile network architecture to decouple control from user plane and estimate how much of signalling changes it requires on the control plane level to manage connectivity and mobility procedures. Decoupling control and user plane could be needed in future FMC scenarios when some functions are move from the mobile core network to other fixed network areas (e.g. use case 6 in D2.1). This exercise allows to know that it is possible to do that and what the modifications needed are.

This approach affects the standard signalling and some modifications have been proposed in the section (for example signalling reallocation suggestions summarized by Table 12). The number of new signalling messages has been counted and the number of removed signalling messages has been computed as well.

The important finding is that without making deliberate changes in the control message flow, some elements of the architecture (especially S-GW, MME, PCRF) will become overwhelmed with control-dialogues. This is mainly due to the fact that the policy-, mobility- and charging-related messages scale up with the number of subscribers, and that the subscribers do prefer to stay mobile with their equipment. The overload of the aforementioned core elements is a high risk, which yields for changes – actually: decoupling – of how control and user plane traffic is handled.

However, a complete analysis to understand whether the total signalling is reduced had not been done in this section – this question is out of the scope of the section itself and it is subject to further studies

5.3 FMC traffic model for aggregation networks

5.3.1 Introduction

FMC is already a reality in the aggregation domain of today's networks, since fixed and mobile traffic is aggregated via the same infrastructure (see Figure 54). However, since locations of fixed and mobile network functions have been traditionally optimised separately, a joint optimisation can leverage synergies by identifying optimum locations for network functions as envisioned by use case 6. This necessitates a combined approach to traffic modelling in order to determine fixed and mobile traffic flows in the aggregation network domain.

The FMC traffic modelling approach documented in this chapter focuses on, but is not restricted to, Germany and can be easily generalised. Figure 53 depicts our modelling approach in terms of collected data which has been used in the modelling process, and can serve as guideline in future FMC modelling attempts.

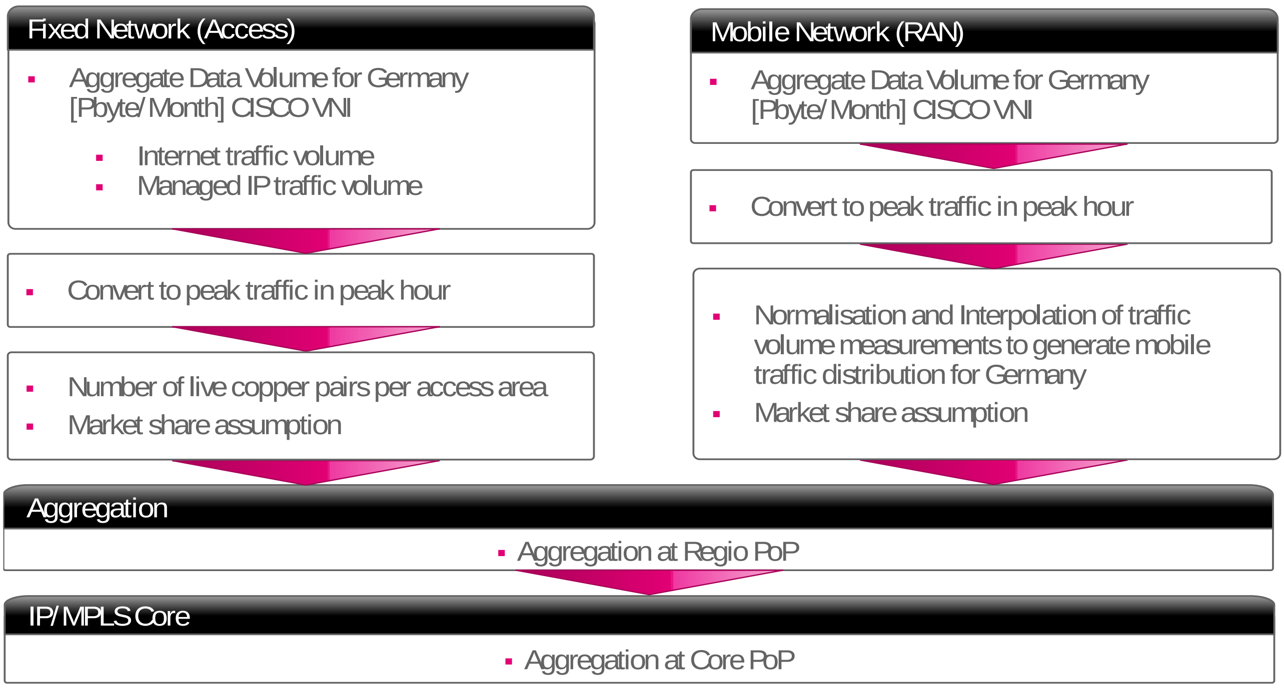


Figure 53: FMC modelling methodology.

To build on a common and openly accessible foundation, the CISCO Visual Networking Index (VNI) has been chosen to provide the forecast data as input for the traffic model. To identify region or core PoP locations, struggling to keep up with corresponding throughput requirements, we will use the *peak traffic in peak hour* as metric. This metric enables us to dimension aggregation and core networks, because traffic contributions of a high number of independent customers are multiplexed together rendering traffic fluctuations on short time scales irrelevant. However, at the edge of the network significantly less customers are aggregated, necessitating a more refined statistical model operating on shorter time scales to cope with inherent traffic burstiness. Since the data provided by the CISCO VNI is available as *traffic volumes per month* it has to be converted to peak traffic in peak hour as indicated in Figure 53.

Aggregated peak traffic in peak hour has to be distributed across all access areas according to certain metric. In case of the fixed network, this metric is the number of copper pairs available for xDSL connections per fixed network access area, which gives a reasonable indication of the regionally distributed traffic. Unfortunately, in case of the mobile traffic, this distribution is harder to come by, which is due to the space varying nature of mobile traffic. Here we used traffic volume measurements as well as estimates for cell coverage areas to interpolate a Germany-wide mobile traffic distribution. To operate on a common geographical metric, this distribution is tied to fixed network access areas. Fixed and mobile traffic components are scaled by a market share assumption. Finally the fixed and mobile traffic is aggregated together via a common fixed/mobile aggregation network (Figure 54 shows such an example) consisting of region and core PoPs.

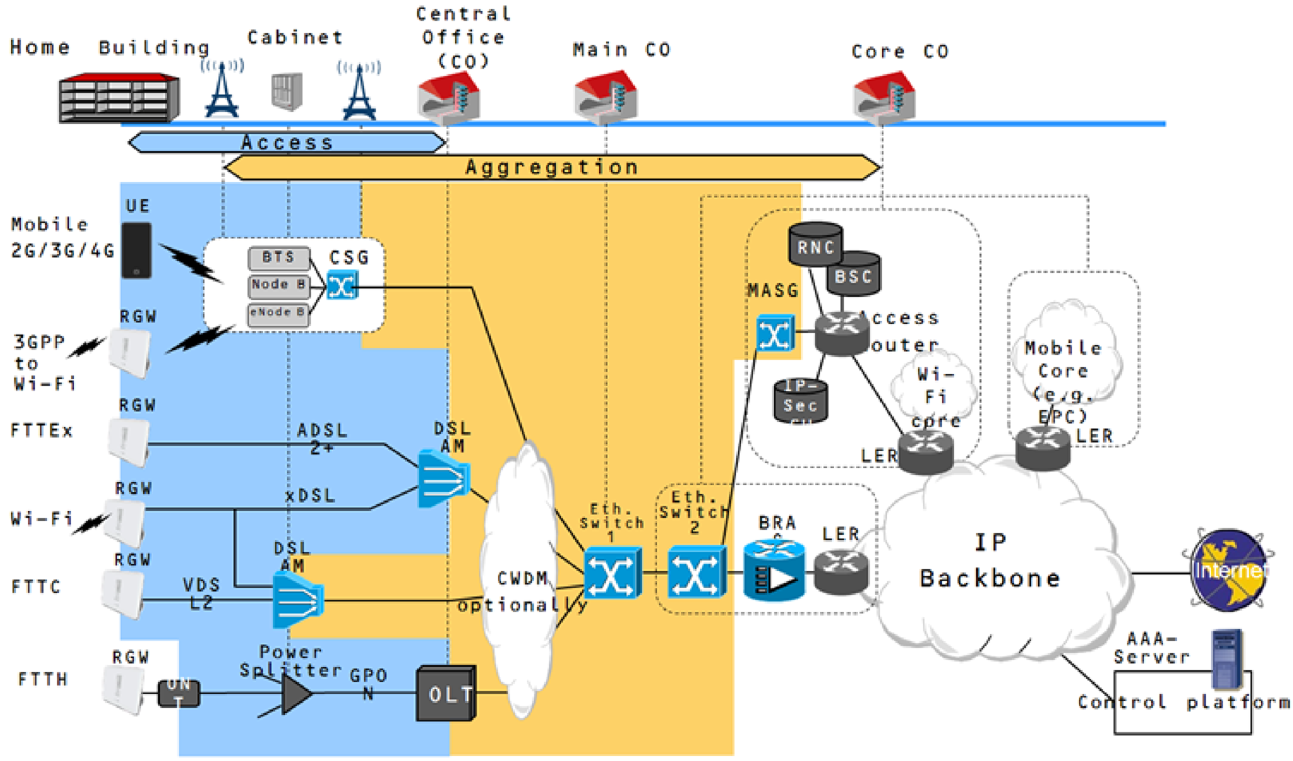


Figure 54: Example of FMC in current Networks.

5.3.2 Traffic volume forecast

The traffic model introduced here is based on traffic forecast data from CISCO VNI 2012 [65], which forecasts the IP traffic evolution up until the year 2017. The study distinguishes between fixed and mobile traffic volume per month, where the fixed traffic share is again subdivided into Internet and managed IP (largely IPTV). Furthermore a basket of applications is introduced covering online gaming, web and other data, and video and file sharing for consumer and business customers. Figure 55 and Figure 56 show fixed network traffic volume predictions for Germany from 2012 to 2017. The dominant fixed network traffic share (Internet as well as managed IP) is traffic attributed to video streams, which is expected to become even more important in the future. In terms of fixed network Internet traffic, web and other data, as well as file sharing traffic, constitute important proportions of the overall traffic, but do not fuel the growth. Business traffic accounts for roughly 20% of the overall IP Internet traffic and for 43% of managed IP traffic. Note that the share of business traffic declines to 28% in 2017, even though the absolute traffic is increasing slightly.

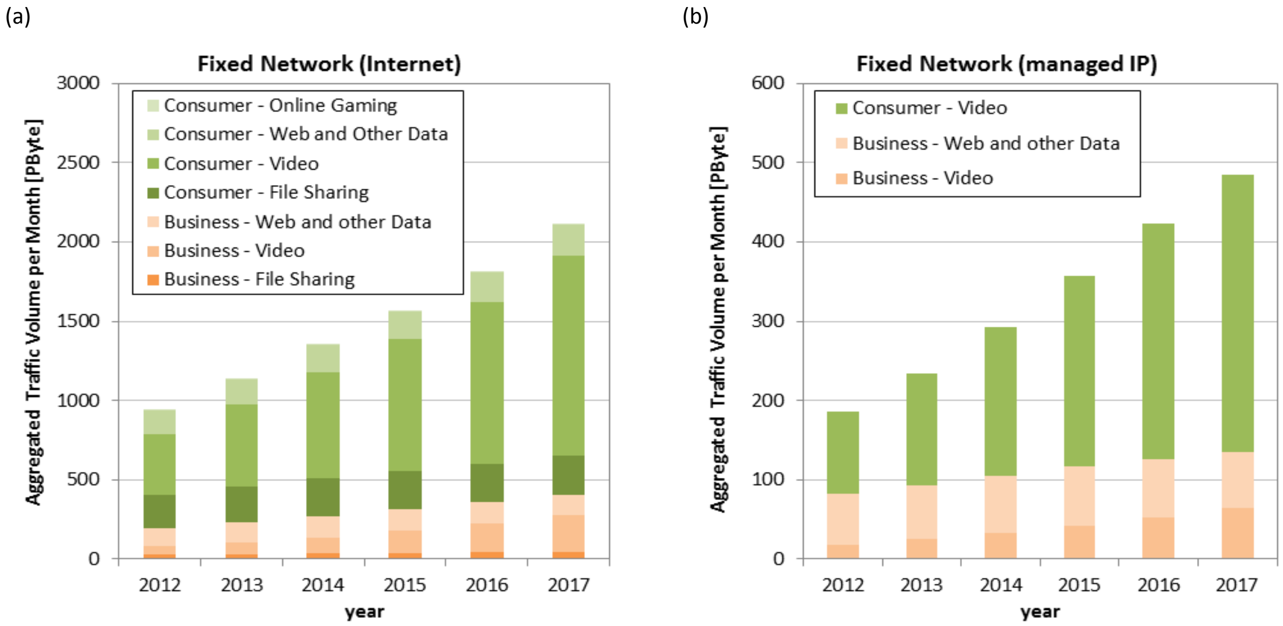


Figure 55: Fixed network Internet traffic volume (a) and managed IP traffic volume (b) for Germany projected until 2017.

Figure 56 depicts mobile network traffic volume in Germany from 2012 to 2017. Here, video and web and other data services are the most important drivers for traffic growth, while file sharing accounts for 6% in 2012 (declining to 2% in 2017). Similarly to the fixed network Internet traffic, business traffic accounts for roughly 20% of the overall traffic in mobile networks as well.

Note that shares of each application and customer classes are given for completeness only. For the sake of reduced complexity, we will restrict ourselves in the following to the categories *Internet* and *managed IP* for *fixed network traffic* as well as *mobile network traffic* without any subdivisions.

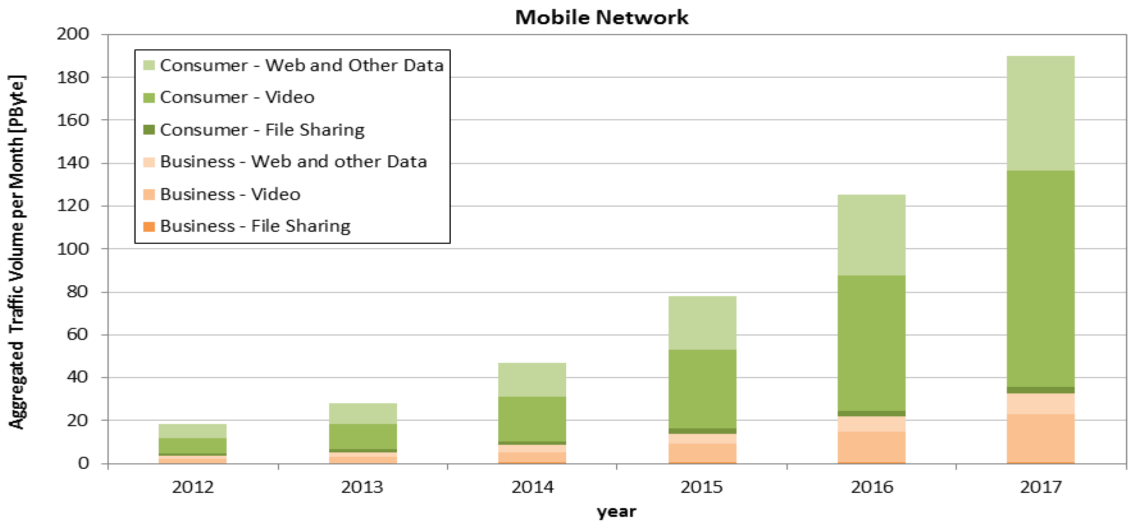


Figure 56: mobile traffic volume for Germany projected until 2017.

Figure 57 directly compares the growth of fixed and mobile network IP traffic volume according to CISCO VNI for Germany. It can be seen, that mobile traffic is expected to grow much faster than fixed network traffic with a compound annual growth rate (CAGR) of more than 50%, whereas fixed network Internet and managed IP traffic grow only with around 20-25% over the next years. Consequently, mobile traffic which corresponded in 2012 to only 1.6% of the fixed network traffic volume is expected to rise to 7.3% in 2017. This development, which is certainly due to the expected proliferation of smartphones exploiting 4G and future 5G mobile networks, underlines the growing importance of FMC traffic models, since the growing mobile traffic has to be backhauled and aggregated via fixed network infrastructure.

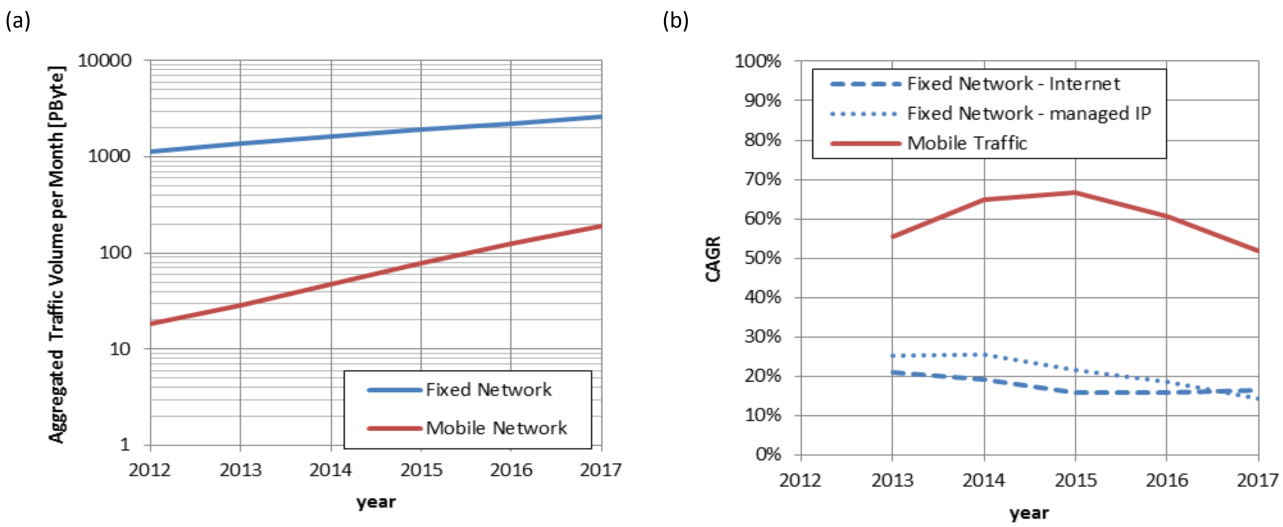


Figure 57: Traffic Volume per month for fixed and mobile traffic (a) and, compound annual growth rate (CAGR) for Internet and managed IP traffic (b) in the fixed network and mobile traffic according to CISCO VNI for Germany.

The peak traffic in peak hour is generally defined as the average traffic during the busiest hour of the day. To convert aggregate traffic volume per month to peak traffic in peak hour the following formula has been used:

$$\text{Peak Traffic in Peak Hour} = \frac{\text{Traffic Volume per month} \cdot 8^{\text{bit}}/_{\text{Byte}}}{17 \cdot 30.44\text{d} \cdot 3600\text{s}} \quad (9)$$

assuming that 1/17 of the overall diurnal traffic is accumulated during the peak hour as depicted in Figure 58.

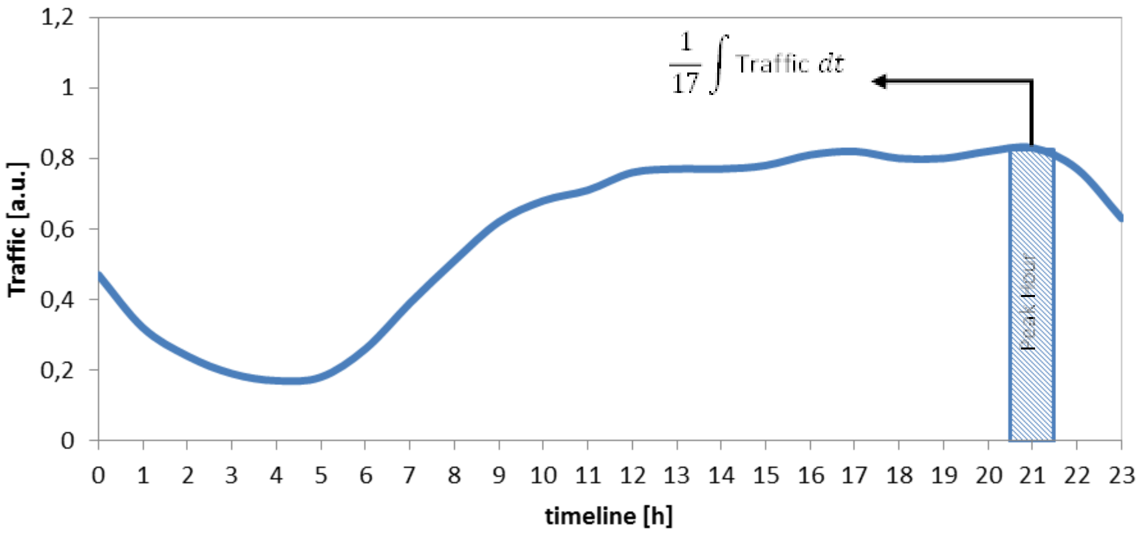


Figure 58: Typical load curve for a remote access router, with 1/17 of the diurnal traffic being accumulated in the peak hour.

5.3.3 Regionalisation of aggregated traffic

Aggregated mobile traffic described in the previous section has been regionalized by using per-cell traffic volume measurements, estimations of coverage areas for particular cells as well as geo-referenced fixed network service area boundaries. In the following the methodology is explained based on an example access area.

Initially, the spatial overlap between cell coverage and each fixed network access area is determined as shown in Figure 59a. Subsequently, the traffic volume measured for each cell is distributed according to the overlap between cell coverage and access area (see Figure 59b).

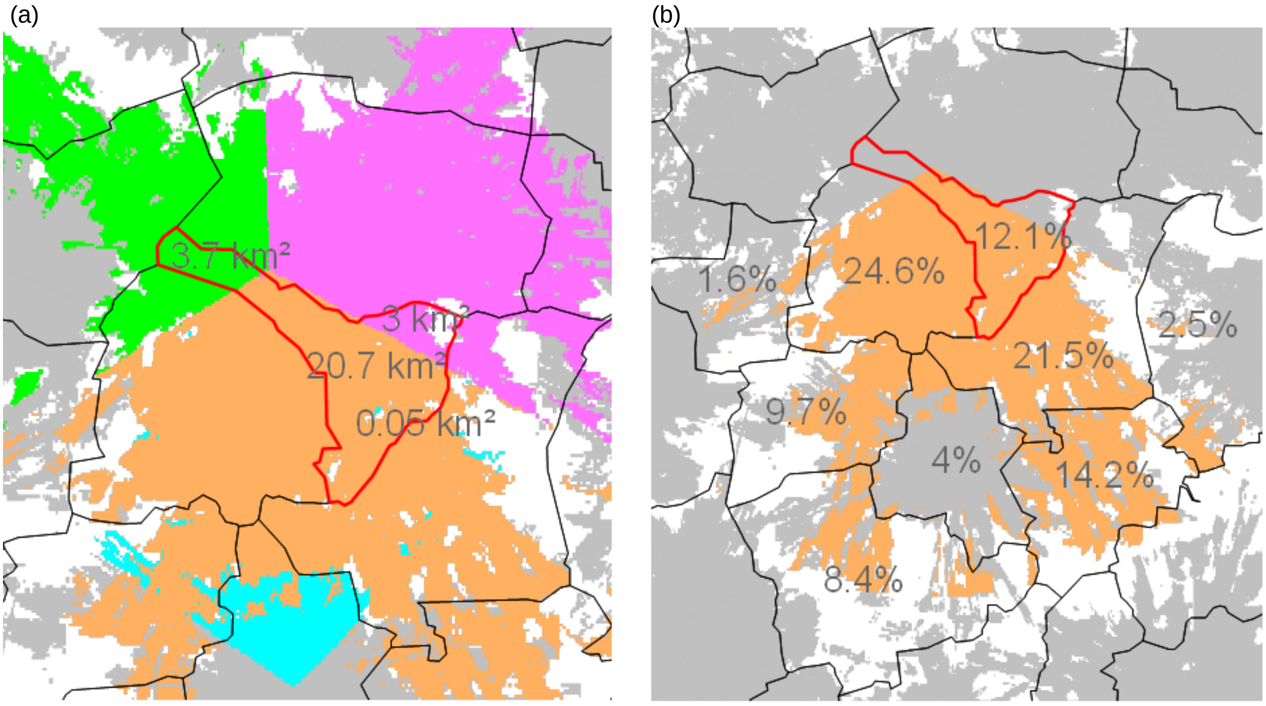


Figure 59: Spatial overlap between multiple mobile cells and an exemplary fixed network access area (a) and, traffic of each cell is distributed to access areas, covered by the corresponding cell (b).

As a next step, it is possible to sum up data volumes from multiple cells and determine the overall volume originating in the corresponding access areas as depicted inFigure 60. However, there are white spots which are not covered by any mobile cell (Figure 60 upper right hand corner of access area). In this case, we still assign a certain data volume to this currently unserved area under the assumption that 100% of each access area will be covered in the future to improve customer service. If more than 60% of the corresponding access area is currently covered by existing cells, data volume contributions from unserved areas are interpolated according to the data volume density in the served area. Otherwise (less than 60% of the access area is covered), uncovered white spots are interpolated according to the mean data volume density of access areas of the same cluster (dense urban, urban, rural). To achieve this, all fixed network service areas have been clustered according to their population density (population/area):

$$\text{Traffic density} = \frac{\text{Sum of UMTS } \textbf{data volume} \text{ of all access areas within cluster}}{\text{Sum of UMTS } \textbf{coverage area} \text{ of all access areas within cluster}} \quad (10).$$

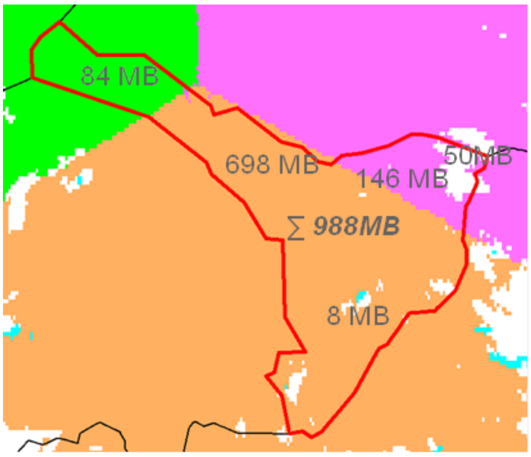
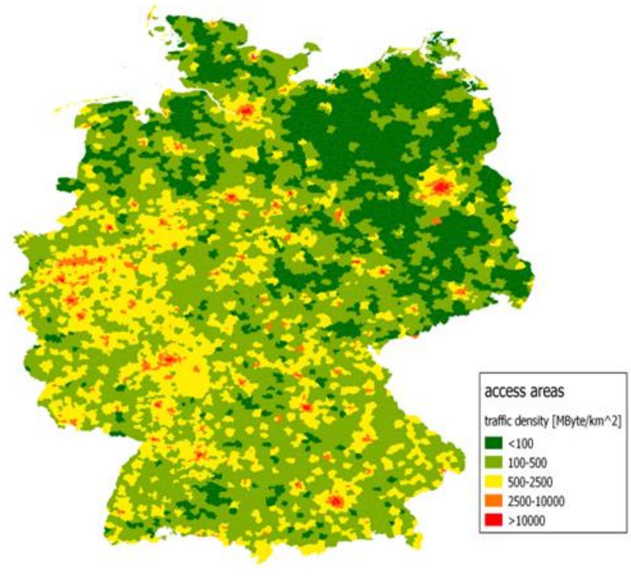


Figure 60: Data volumes originating in example access area.

Figure 61a shows the result of the regionalization in terms of the varying traffic density across Germany. Due to the lack of customer mobility, fixed network traffic per access area can be regionalized by using the number of live copper pairs per access area in Germany as shown in Figure 61b.

(a)



(b)

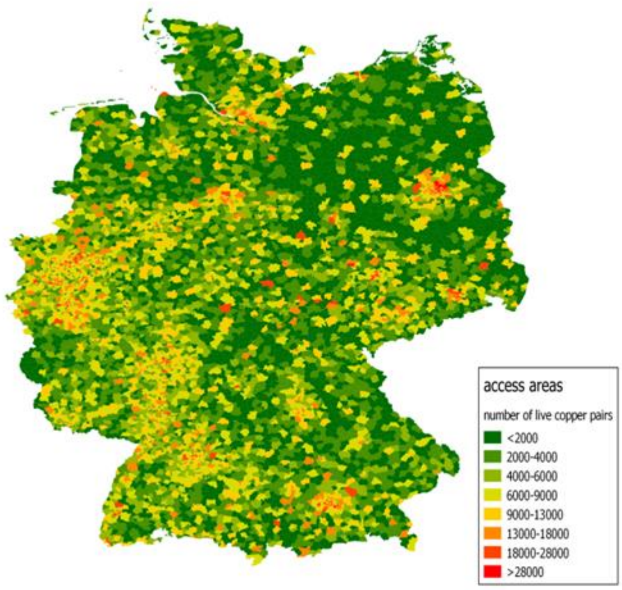


Figure 61: Base to regionalize forecasted traffic for (a) mobile network [traffic density per access area] and (b) fixed network [number of live copper pairs].

After assigning each access area its corresponding peak traffic in peak hour share, the value is scaled by the Deutsche Telekom market share which is 33.15% for fixed network and 31.2% for the mobile network as published by the German regulating authority responsible for the telecommunication sector [76].

Fixed network traffic originating in the access areas is initially concentrated in local exchanges and then aggregated at region PoPs as well as core PoPs as depicted in Figure 54. On the mobile network side traffic is originating from the antennas, concentrated via a common fixed/mobile aggregation network (region and core PoPs) and transported to the mobile core network. Figure 62 features a potential fixed/mobile aggregation network topology with more than 800 region PoPs as well as 18 core PoPs at 9 locations. The present topology is designed to aggregate large traffic volumes which are assumed to be related to the population density.

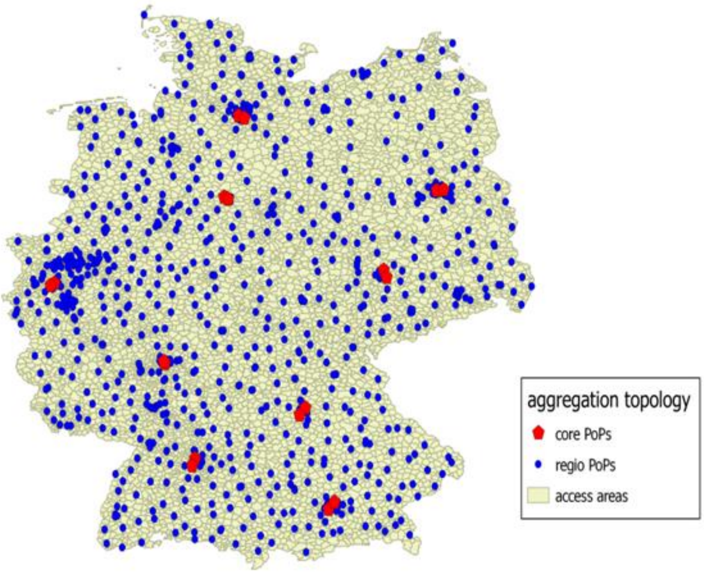


Figure 62: Potential Germany-wide aggregation topology with region and core PoPs.

5.3.4 Traffic Evolution at Region and Core PoPs

Figure 63 shows the projected rise in IP traffic of the busiest region and core PoPs for fixed and mobile network traffic between 2012 and 2017. The busiest region PoP does not exceed a peak traffic of 15.1 Gbit/s fixed network traffic and 1.3 Gbit/s mobile network traffic by the end of 2017. In 2017 the fixed network traffic at the busiest region PoP will not even have doubled compared to 2013, whereas mobile traffic at the busiest PoP will increase by more than a factor of 6 during the same time span.

Considering the core network: the busiest core PoP will experience fixed network traffic of 390 Gbit/s in 2017, while the busiest mobile network core PoP increases to 21.3 Gbit/s (the entire mobile network traffic in 2012 amounts to 24 Gbit/s). Again, mobile traffic at the busiest core PoP increases by more than a factor of 6 between 2013 and 2017, while fixed network traffic not even doubles. Note that, the fixed network traffic will probably be lower in reality, since IPTV traffic (the dominant part of managed IP traffic – see Figure 55) is multicast traffic leading to a multicast gain in core and aggregation networks.

(a) (b)

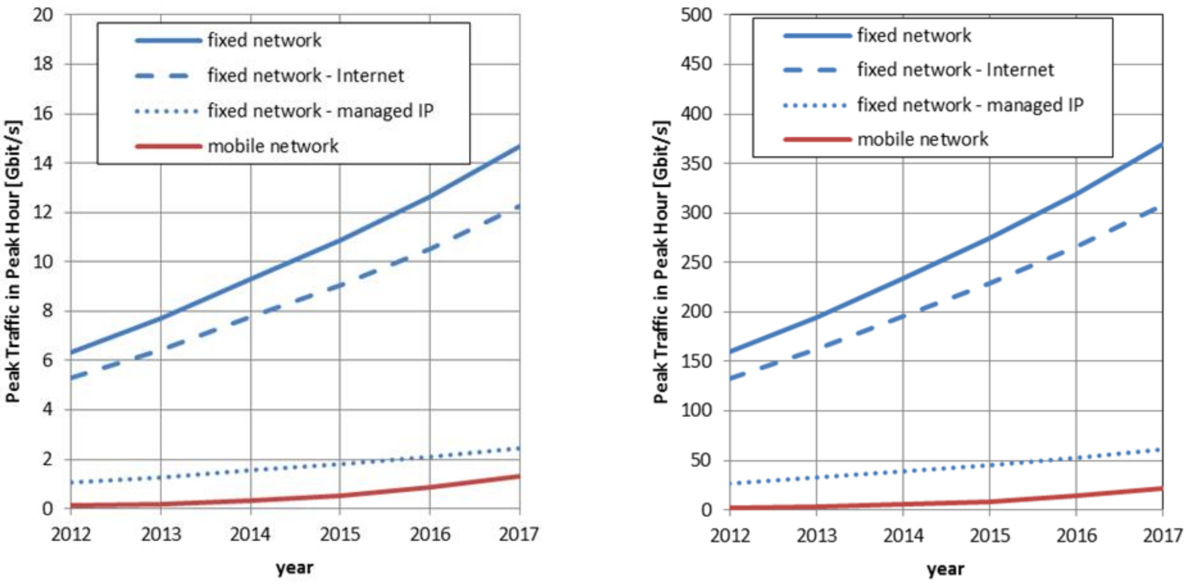


Figure 63: Peak traffic in peak hour of the busiest region (a) and core PoPs (b) for fixed and mobile network traffic.

To cover potential variations in traffic growth a progressive and conservative case have been introduced as shown in Figure 64. In the progressive case, the traffic is assumed to grow twice as fast compared to the CISCO predictions, while in the conservative case the traffic grows at half the CISCO rate. As a consequence, the fixed network traffic doubles in 2017 when comparing the progressive case to the CISCO predictions, while the conservative case leads to 35% less fixed network traffic. In the mobile network the development is more extreme due to a generally higher growth rate (see Figure 57 b): the progressive case leads to an increase by a factor of 5 in 2017 compared to the CISCO predictions, while the conservative traffic assumption leads to a 65% reduction in mobile network traffic, again compared to the CISCO case.

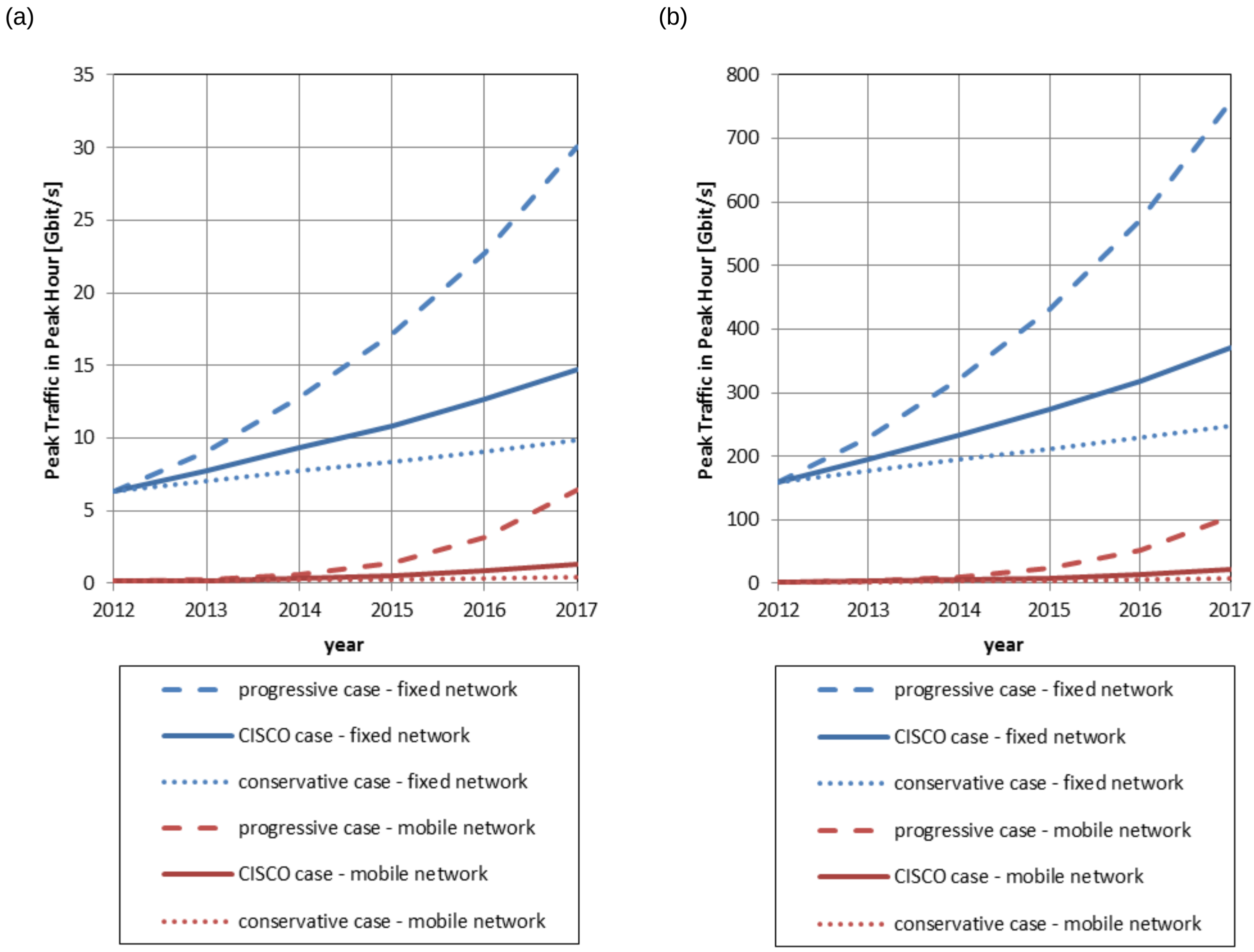


Figure 64: Peak traffic in peak hour of the busiest region (a) and core PoPs (b) for progressive, CISCO and conservative cases.

Figure 65 depicts the cumulative distribution of peak traffic for region and core PoP in the year 2017 of the CISCO case. The average region PoP deals with 4.1 Gbit/s of peak traffic in the peak hour, compared to the average core PoP, which has to deal with 205 Gbit/s. While less than 3% of the region PoPs exhibit a traffic of more than 10 Gbit/s, the upper 10% of the core PoPs process 390 Gbit/s. The median for region PoPs lies at 3.4 Gbit/s and for core PoPs at 194 Gbit/s.

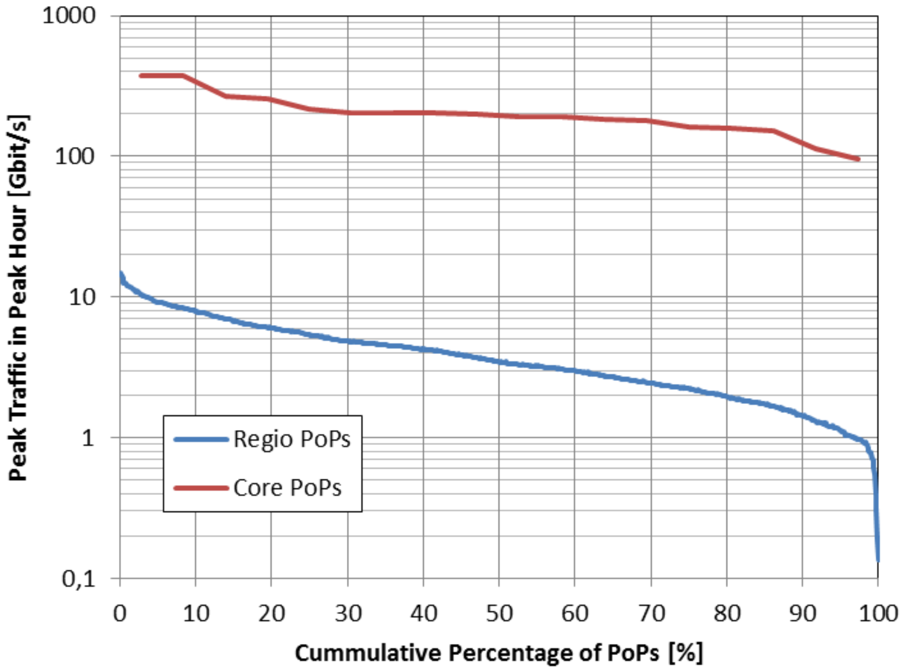


Figure 65: Cumulative distribution of peak traffic in peak hour at region and core PoPs in 2017 (CISCO case).

5.3.5 Summary of FMC traffic model for aggregation networks

It was found that mobile traffic, though growing exponentially, is still contributing only a small fraction to the overall traffic (2012: 1.6%, 2017: 7.3%) in the average traffic growth scenario. However, under a progressive traffic growth assumption, mobile traffic at the busiest core PoP will grow to 104Gbit/s, which corresponds to 16.6% of the traffic being processed at that particular PoP. The average regional PoP traffic in 2017 amounts to 4Gbit/s and the average traffic at a core PoP to 200Gbit/s, again assuming average traffic growth over the next years.

This clearly indicates that traffic growth due to mobile traffic alone is potentially not a driver to change the aggregation network structure or even re-direct traffic from mobile users closer to the edge of the network affecting the network architecture.

6 CONCLUSIONS

This deliverable serves as a reference point of traffic-related considerations within COMBO. It provides a snapshot of currently visible trends for the ways traffic changes, and provides modelling instruments for revealing the complex patterns of traffic. These trends – and their timing – affect architectural plans, capacity plans, decisions for offloading scenarios, performance management considerations, and techno-economic studies.

The main drivers for traffic growth are well-known: the increasing number of broadband subscribers and devices, the need for high definition, streaming media, the introduction of new services and applications, and the effect of these factors on bandwidth demands. Various sources predict a 1000-times traffic increase within a 20 years span – if not earlier.

This document contains traffic analysis results from various parts of the network and service domains of various operators. It does not attempt to cover the countless combination of analysis possibilities at all domains. The aim was to show traffic analysis instrumentation at work, answering questions at various domains of the architecture (both fixed and mobile, aggregation and core), and providing methodologies for modelling the traffic – based on the analysis results.

There were, however some important comparative traffic analysis studies carried out. The analysis of real Internet traffic at Orange and Telefónica shows that busy hour(s) are in the same period: 7-9PM, and that there are also some common application usage characteristics appear in the same time on fixed and mobile networks. Video streaming, as an example, is commonly used at the same time on fixed and mobile networks, hence its more optimized handling can be considered during FMC scenarios. There are indeed differences between the Orange and Telefónica usage profiles: the analysis results the show differences in the daily life of the subscribers. There are small variations on these results depending on the area, usage culture, and other factors.
Usually these traffic profiles are very similar, with occasional differences. This means that general traffic models can be set up; although the variables of the models have to be chosen properly, to allow the necessary tuning.

Even though real FMC traffic archives are not available, the measurement results provided from public traffic archives represent various cases where fixed and mobile traffic are highly aggregated. In all the traffic traces significant values of self-similarity were measured. Therefore, complex models (considering self-similarity) are suggested to be used for aggregated FMC traffic and over-provisioning is suggested for the involved nodes.

The consumer Internet traffic is expected to increase in all data types. Video will be the dominant data type (and its volume will grow faster than others'), with regular web-access, data access and mail traffic still being important. File sharing is not predicted to grow significantly; and the volume of gaming traffic (which operates wilh small packets) remains low.
Data traffic from mobile access will grow faster than from fixed access; it is predicted to reach over 200 ExaBytes per month for the 2020 horizon, becoming four times higher than it was in 2013.Fixed, mobile and wifi access will be there in 2020 and we have considered that in some use cases: UC1 and UC4 - this is in-line with the work done in Task 2.1 within COMBO.

Metro traffic will increase in greater pace than long-haul - this means that performance issues must be tackled earlier in that segment. The data exchange related to machine-to-machine communication is expected to grow significantly; although the machines are not expected to become major contributors of the ovrall mobile traffic within the 2020 horizon. The main responsible for the mobile traffic growth will be the smartphones; laptops and tablets also play an important role.
An important finding on mobile control traffic is that it will grow exponentially, due to smaller location areas, and the increasing number of policy-, charging- and mobility-related dialogues.

Since the actual FMC traffic scenarios are not live yet, at this stage it is not possible to provide models based on analysis results. Nevertheless, it is possible to provide the building blocks of such models and to define the methodology on building such models. During WP6 tasks at COMBO it is worth considering building the actual traffic models for live (or laboratory) scenarios, based on the results of the current document.

Although analysing real data of “FMC traffic” was not possible at this stage, this deliverable describes how traffic patterns and models are expected to be affected by some of the FMC scenarios. There are two models detailed in this document: one for offloading scenarios with WLAN femtocell, and another for evolution of traffic composition within aggregation networks.

Regarding the traffic of control messages, the important finding is that without making deliberate changes in the control message flow, some elements of the architecture (especially S-GW, MME, PCRF) will become overwhelmed with control-dialogues. This is mainly due to the fact that the policy-, mobility- and charging-related messages scale up with the number of subscribers, and that the subscribers do prefer to stay mobile with their equipment. The overload of the aforementioned core elements is a high risk, which yields for changes – actually: decoupling – of how control and user plane traffic is handled.

For this deliverable, a traffic model for aggregation networks has been developed based on a network topology optimized for Germany. During this work, a method for aggregated FMC traffic planning has been developed. This method works as follows. To obtain a holistic view on fixed and mobile traffic, each fixed network access area (of the country) serves as a source of both types of traffic. For the fixed network, the forecasted fixed network traffic has to be regionally distributed based on the number of copper pairs available for xDSL connections per access area. For the mobile network, both mobile traffic volume measurements as well as estimates for cell coverage areas can be used to interpolate a country-wide mobile traffic distribution. The next step is to scale the traffic in both domains by a market share assumption to reflect the network load of an isolated ISP. Subsequently, traffic forecasts for fixed and mobile traffic at regional and core PoP locations can be developed for years ahead for conservative, average and progressive traffic growth assumptions.

During the analysis of aggregated traffic at PoPs it was found that mobile traffic, though growing exponentially, is still contributing only a small fraction to the overall traffic (2012: 1.6%, 2017: 7.3%) in the average traffic growth scenario. However, under a progressive traffic growth assumption, mobile traffic at the busiest core PoP will grow to 104 Gbit/s, which corresponds to 16.6% of the traffic being processed at that particular PoP. The average regional PoP traffic in 2017 amounts to 4 Gbit/s and the average traffic at a core

PoP to 200 Gbit/s, again assuming average traffic growth over the next years. This clearly indicates that traffic growth due to mobile traffic alone is not a potential driver to change the aggregation network structure or even re-direct traffic from mobile users closer to the edge of the network affecting the network architecture.

In summary, the advances of FMC and traffic analysis and modelling introduced in this document should be used within COMBO in the following way:

- current traffic demands, trends, drivers are identified; Task 3.3 (Fixed/mobile convergence at equipment and infrastructure level) can build on these during equipment and infrastructure planning, whereas WP5 can build these results in during the techno-economic planning (both Task 5.2 - Cost analysis, and Task 5.3 - Impact of convergence on business ecosystems);
- similarities on the traffic analysis results reinforced that generalized models can be built, differences reveal the important variables to be built into future models - this can be used within Task 4.2 (Performance Management) on performance considerations, as well as within WP6 (Task 6.2 - Lab-based practical work and development): test output gathered by using D2.3 traffic models will be applicable more widely than just within the laboratory;
- measurements showed that real traffic is inevitably complex and self-similarity should be considered for capacity planning: we do have to over-provision the network elements, because unforeseen bursts will occur – this can be used within WP3 (Task 3.2 - Fixed/mobile convergence at protocol level, Task 3.3 - Fixed/mobile convergence at equipment and infrastructure level), Task 4.2 as well as Tasks 5.2 and 5.3 during traffic-planning.

7 REFERENCES

[1] M. Scott, “Study on Impact of traffic off-loading and related technological trends on the demand for wireless broadband spectrum”, 2013.

[2] M. Feknous, T.Houdoin, B. Le Guyader, J. De Biasio, A. Gravey, J.Torrijos, « Internet Traffic Analysis : A Case Study From Two Major European Operators », ISCC , June 2014.

[3] S.Gosselin, J. De Biasio, M. Feknous, T.Mamouni, J. Torrijos and al, « Fixed and Mobile Convergence :Needs and Solutions », EW, May 2014

[4] M.Feknous, B. Le Guyader, A.Gravey, “Revisiting Access and Aggregation Network Architecture, CCSCI, January 2014.

[5] Sandvine, “Global Internet Phenomena Report”, 2013. http://www.sandvine.com

[6] UMTS Forum Report 44 “Mobile traffic forecasts 2010-2020© UMTS Forum January 2011”, http://www.umts-forum.org/component/option,com_docman/task,doc_download/gid,2537/Itemid,213/

[7] Cisco. Cisco visual networking index: Forecast and methodology, 2011-2016. White paper, 2012. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf

[8] Solon. Broadband on demand: Cable’s 2020 vision. White paper, 2011.

[9] Minnesota Internet Traffic Studies (MINTS), Online: http://www.dtc.umn.edu/mints/

[10] Kilper, D.; Atkinson, G.; Korotky, S.; Goyal, S.; Vetter, P.; Suvakovic, D. & Blume, O.; Power Trends in Communication Networks; IEEE Journal of Selected Topics in Quantum Electronics, 2011, 17, 275-284

[11] Cisco. The zettabyte era. White paper, 2013.

[12] Korotky, S. K.; Traffic Trends: Drivers and Measures of Cost-Effective and Energy-Efficient Technologies and Architectures for Backbone Optical Networks; Optical Society of America, 2012

[13] AT&T Labs. Traffic types and growth in backbone networks. Technical report, 2011.

[14] ITU-R M.2243. Assessment of the global mobile broadband deployments and forecasts for international mobile telecommunications. Technical report, 2012.

[15] Ericsson. Traffic and market report. White paper, 2013.

[16] A.T. Kearney, GSMA, “The Mobile Economy 2013". http://www.atkearney.com

[17] “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast 2012–2017”, Cisco Systems.

[18] “25 labs”, Technology blog, 8 September 2011. 25labs.com

[19] Analysys Mason “Global mobile network traffic” ,29 June 2011.

[20] "Facebook's Mobile Users, Q4 2010 - ", May 2013, available at http://trends.e-strategyblog.com/2013/05/08/facebooks-mobile-monthly-active-users/10979

[21] D. Wonak, "A storm is brewing – an LTE signalling storm", Diametriq, white paper, September, 2012.

[22] Tekelec LTE Diameter signalling index: forecast report and analysis 2011-2016, 2012

[23] 3GPP TR 23.843, "Study on Core Network Overload Solutions", Rel'12, V0.8.0 (2013-04)

[24] M. Olsson, et al., "SAE and the Evolved Packet Core: Driving the mobile broadband revolution", Elsevier, 2009.

[25] C. Knight "Measuring the explosion of LTE signalling traffic – a Diameter traffic model", Diametriq, white paper, September 2012.

[26] Diametriq Diameter Traffic Calculator, http://www.diametriq.com/lte-signalling-traffic-model/

[27] K. Iniewski, et al. "Convergence of mobile and stationary next-generation networks", Wiley, 2011.

[28] Balakrishnan Chandrasekaran, "Survey of Traffic Models", available at http://www.cse.wustl.edu/~jain/cse567-06/ftp/traffic_models3/index.html.

[29] Jacques Lévy Véhel, Rudolf H. Riedi, "Fractional Brownian motion and data traffic modelling: The other end of the spectrum", INRIA Rapport de Recherche August 1997.

[30] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, D. Veitch, "The Multiscale Nature of Network Traffic", IEEE Signal Processing Mag., vol. 19, no. 3, pp. 28-46, May 2002.

[31] K. Park, W. Willinger, "Self-Similar Network Traffic: An Overview", in Self-Similar Network Traffic and Performance Evaluation, K. Park, W. Willinger, Eds. Chichester, UK: John Wiley & Sons, 2000, pp. 1-38.

[32] P. Abry, P. Flandrin, M. S. Taqqu, D. Veitch, "Wavelets for the Analysis, Estimation, and Synthesis of Scaling Data", in Self-Similar Network Traffic and Performance Evaluation, K. Park, W. Willinger, Eds. Chichester, UK: John Wiley & Sons, 2000, pp. 39-88.

[33] V. Paxson, S. Floyd, "Wide-Area Traffic: the Failure of Poisson Modelling", IEEE/ACM Trans. Networking, vol. 3, no. 6, pp. 226-244, June 1995.

[34] G. W. Wornell, A. V. Oppenheim, "Estimation of Fractal Signals from Noisy Measurements Using Wavelets", IEEE Trans. Signal Processing, vol. 40, no. 3, pp. 611-623, March 1992.

[35] M. S. Taqqu, V. Teverovsky, W. Willinger, "Estimators for Long-Range Dependence: an Empirical Study", Fractals, vol. 3, no.4, pp. 785-798, 1995.

[36] P. Abry, P. Gonçalvès, P. Flandrin, "Wavelets, Spectrum Analysis and 1/f Processes", in Lecture Notes in Statistics: Wavelets and Statistics, vol 103, A. Antoniadis, G. Oppenheim, Eds. Springer, 1995, pp. 15-29.

[37] P. Abry, D. Veitch, “Wavelet Analysis of Long-Range Dependent Traffic”, IEEE Trans. Inform. Theory, vol. 44, no.1, pp. 2-15, Jan. 1998.

[38] D. Veitch, P. Abry, “A Wavelet-Based Joint Estimator of the Parameters of Long-Range Dependence”, IEEE Trans. Inform. Theory, vol. 45, no.3, pp. 878-897, Apr. 1999.

[39] ITU T Rec. G.810 "Definitions and Terminology for Synchronisation Networks", Rec. G.811 "Timing Characteristics of Primary Reference Clocks", Rec. G.812 "Timing Requirements of Slave Clocks Suitable for Use as Node Clocks in Synchronization Networks", Rec. G.813 "Timing Characteristics of SDH Equipment Slave Clocks (SEC)", Geneva, 1996-2003.

[40] S. Bregni, "Chapter 5 - Characterization and Modelling of Clocks", in Synchronization of Digital Telecommunications Networks. Chichester, UK: John Wiley & Sons, 2002, pp. 203-281.

[41] S. Bregni, L. Primerano, “The Modified Allan Variance as Time-Domain Analysis Tool for Estimating the Hurst Parameter of Long-Range Dependent Traffic“, Proc. IEEE GLOBECOM 2004, Dallas, TX, USA, 2004.

[42] S. Bregni, L. Jmoda, "Accurate Estimation of the Hurst Parameter of Long-Range Dependent Traffic Using Modified Allan and Hadamard Variances", to appear in IEEE Trans. Communications, 56.11 (2008): 1900-1906.. Extended version available: http://home.dei.polimi.it/bregni/public.htm.

[43] S. Bregni, R. Cioffi, M. Decina, "An Empirical Study on Time-Correlation of GSM Telephone Traffic", IEEE Transactions on Wireless Communications, Vol. 7, No. 9, Sept. 2008, pp. 3428-3435.

[44] A. Pattavina, A. Parini, "Modelling Voice Call Interarrival and Holding Time Distributions in Mobile Networks", Proc. 19th International Teletraffic Congress (ITC), Bejing, Aug. 2005.

[45] S. Hur, et al. "Performance limits of pico-cell systems using radio-over-fiber techniques with an electro absorption modulator." Optical and Quantum Electronics 39.7 (2007): 561-569.

[46] H. Claussen. "Co-channel operation of macro-and femtocells in a hierarchical cell structure." International Journal of Wireless Information Networks 15.3-4 (2008): 137-147.

[47] S. Anand, A. Sridharan, and K. N. Sivarajan. "Performance analysis of channelized cellular systems with dynamic channel allocation." Vehicular Technology, IEEE Transactions on 52.4 (2003): 847-859.

[48] D. Hong, and S. S. Rappaport. "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures." Vehicular Technology, IEEE Transactions on 35.3 (1986): 77-92.

[49] T.S. Kim, et al. "Mobility modelling and traffic analysis in three-dimensional indoor environments." Vehicular Technology, IEEE Transactions on 47.2 (1998): 546-557.

[50] Y. C. Kim, et al. "Dynamic channel reservation based on mobility in wireless ATM networks." Communications Magazine, IEEE 37.11 (1999): 47-51.

[51] P. A. Dintchev, B. Perez-Quiles, and E. Bonek. "An improved mobility model for 2G and 3G cellular systems." 3G Mobile Communication Technologies, 2004. 3G 2004. Fifth IEE International Conference on. IET, 2004.

[52] S. Bhattacharya, H. M. Gupta, and S. Kar. "Traffic model and performance analysis of cellular mobile systems for general distributed handoff traffic and dynamic channel allocation." Vehicular Technology, IEEE Transactions on 57.6 (2008): 3629-3640.

[53] W.A. Massey, and W. Whitt, “Networks of finite server queues with nonstationary Poisson input”, Queueing Syst. Vol. 13, pp. 183-250.

[54] G. Foschini, B. Gopinath, and Z. Miljanic. "Channel cost of mobility." Vehicular Technology, IEEE Transactions on 42.4 (1993): 414-424.

[55] M. Rajaratnam, and F. Takawira. "Hand-off traffic modelling in cellular networks." Global Telecommunications Conference, 1997. GLOBECOM'97., IEEE. Vol. 1. IEEE, 1997.

[56] S. Bhattacharya, H. M. Gupta, and S. Kar. "Performance Modelling of Cellular Mobile Systems: A Review of Recent Advances." IETE Technical Review 27.1 (2010): 15.

[57] M. Z. Chowdhury, and Y. M. Jang. "Handover management in high-dense femtocellular networks." EURASIP Journal on Wireless Communications and Networking 2013.1 (2013): 1-21.

[58] E. Chlebus, and W. Ludwin, “Is Handoff Traffic Really Poisson?”, IEEE International Conference on Universal Personal Communications (ICUPC), 1995.

[59] S. Choi, and K. Sohraby, “Analysis of a mobile Cellular System with Hand-off Priority and Hysteresis Control”, IEEE International Conference on Computer Communications (INFOCOM), 2000.

[60] M. Sidi, and D. Starobinski, “New Call Blocking Versus Handoff in Cellular Networks”, IEEE International Conference on Computer Communications (INFOCOM), 1996.

[61] Cisco, Traffic Analysis for Voice over IP, White paper, 2001., http://www.cisco.com/en/US/docs/ios/solutions_docs/voip_solutions/TA_ISD.pdf

[62] S. Aalto and P. Lassila, “Impact of size-based scheduling on flow level performance in wireless downlink data channels”, Proceedings of the 20th International Teletraffic Conference, 2007.

[63] The CAIDA UCSD Anonymized Internet Traces 2013. <2013/05/29 and 2013/06/20>. http://www.caida.org/data/passive/passive_2013_dataset.xml

[64] The CAIDA UCSD OC48 Internet Traces Dataset. <2002/08/14> http://www.caida.org/data/passive/passive_oc48_dataset.xml

[65] Cisco Visual Network Index Global Mobile Data Traffic Forecast Update, 2012-2017,February 2013

[66] http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf

[67] J.Ankey, “Google: Mobile devices generate 10% of all YouTube video downloads”, September 2011\ www.fiercemobileit.com

[68] 3GPP TS 22.278 “Service requirements for the Evolved Packet System (EPS)”, V12.2.0 (2013-03).

[69] “Integrated Femto-Wi-Fi (IFM) Networks”, Small Cell Forum, February, 2012.

[70] M. Taqqu, W. Willinger, and R. Sherman, ”Proof of a fundamental result in self-similar traffic modelling”, Computer Communications Review 26:5-23, 1997.

[71] S. Stoev, G. Michailidis, and J. Vaughan, ”On Global Modelling of Network Traffic”, INFOCOM 2010, The 29th Conference on Computer Communications, San Diego, California, March 2010.

[72] White paper ”Analysis of Traffic Offload: Wi-Fi to Rescue”, WirelessE2E LLC, September, 2010.

[73] 3GPP TR 23.829, “Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO)”, V10.0.1 (2011-10).

[74] 3GPP TR 23.861 “Network based IP flow mobility ”, V1.7.0 (2012-11).

[75] K. Lee, Y. Yi, J. Lee, I. Rhee, S. Chong, ”Mobile Data Offloading: How Much Can Wi-Fi Deliver?”, in proc. of ACM CoNEXT 2010, USA, December, 2010.

[76] Bundesnetzagentur: „Wettbewerbsintensität im Mobilfunk nimmt weiter zu“, press report, 2012. [online] http://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Allgemeines/Presse/Pressemitteilungen/2012/120824_WettbewerbMobilfunkpdf.pdf

8 Glossary

Acronym / Abbreviations	Brief description
2G	2nd Generation (mobile service)
3G	3rd Generation (mobile service)
3GPP	3rd Generation Partnership Project
ADSL	Asymmetric Digital Subscriber Line
ALM	Application-Level Multicast
AVAR	Allan Variance
AVSP	Adaptive Video Streaming Protocol
BHCA	Busy-Hour Call Attempts
BMGF	Binomial Moment Generating Function
BRAS	Broadband Remote Access Server
CAGR	Compound Annual Growth Rate
CDN	Content Delivery Network
COMBO	COnvergence of fixed and Mobile BrOadband
DAG	Data Acquisition and Generation
DHCP	Dynamic Host Configuration Protocol
DSLAM	Digital Subscriber Line Access Multiplexer
EDGE	Enhanced Data Rates for GSM Evolution
EIR	Equipment Identity Register
eNodeB	Evolved Node B (base station)
EPC	Evolved Packet Core
E-UTRAN	Evolved UMTS Terrestrial Radio Access network
FBM	Fractional Brownian Motion
FMC	Fixed Mobile Convergence (Converged)
FTTH	Fiber to the Home
GGSN	Gateway GPRS Support Node
HLR	Home Location Register
HSPA	High Speed Packet Access
HSS	Home Subscriber Server
HTTP	Hypertext Transfer Protocol
i.i.d.	Independent and Identically Distributed
IP	Internet Protocol
ISP	Internet Service Provider
ITU-T	International Telecommunications Union- Telecommunication Standardisation Sector
LTE	Long Term Evolution (3GPP standard)
LRD	Long-Range Dependence
M2M	Machine-to-Machine
MAVAR	Modified Allan Variance
MSC	Mobile Switching Center
MME	Mobile Management Entity
MNO	Mobile Network Operator
MPS	Messages Per Second
MSS	Mobile Switching Centre Server
NG-PoP	Next Generation Point of Presence
NUT	Network Under Test
OCS	Online Charging System

Acronym / Abbreviations	Brief description
OLT	Optical Line Termination
P2P	Peer-to-Peer
PALM	Poisson-arrival-location model
PAMR	Public Access Mobile Radio
PCRF	Policy and Charging Rules Function
PDN	Packet Data Network
PDN-GW	Packet Data Network Gateway
PDP	Packet Data Protocol
PON	Passive Optical Line
PoP	Point of Presence
PGF	Probability Generating Function
P-GW	Packet Data Network Gateway
PSD	Power Spectral Density
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RAN	Radio Access Network
RNC	Radio Network Controller
SCTP	Stream Control Transmission Protocol
SIM	Subscriber Identity Module
SGSN	Serving GPRS Support Node
S-GW	Serving Gateway
SS	Self-Similar
SSSI	Self-Similar Processes with Stationary Increments
TEID	Tunnel Endpoint Identifier
TFT	Traffic Flow Template
TCP	Transmission Control Protocol
UE	User Equipment
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
WLAN	Wireless Local Area Network
VoD	Video on Demand
VoIP	Voice over Internet Protocol
VoLTE	Voice over LTE
VLR	Visitor Location Register
WAN	Wide Area Network
WCDMA	Wideband Code Division Multiple Access
Wi-Fi	Wireless Local Area Network – Commercial name
WiMAX	Worldwide Interoperability for Microwave Access

9 List of Tables

10 List of Figures

11 About this document

11.1 List of authors

Full Name – E-mail	Company – Country Code
C. Behrens (BehrensC@telekom.de) E. Bogenfeld (eckard.bogenfeld@telekom.de)	DTAG – DE
J. De Biasio (joseph.debiasio@orange.com) M. Feknous (moufida.feknous@orange.com)	FT – FR
S. Höst (Stefan.Host@eit.lth.se)	ULUND – SE
A. Krendzel (andrey.krendzel@cttc.es) P. Dini (paolo.dini@cttc.es)	CTTC – ES
S. Bregni (bregni@elet.polimi.it) A. Pattavina (pattavina@elet.polimi.it)	POLIMI – IT
P. Varga (pvarga@tmit.bme.hu) P. Olaszi (polaszi@aitia.ai)	AITIA – HU
O. Eker (onur.eker@turktelekom.com.tr)	TT – TR

12 Further information

Grant Agreement number: 317762

Project acronym: COMBO

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

Funding Scheme: Collaborative Project – Integrated Project

Date of latest version of the Deliverable: 18-06-2014

Delivery Date: **Month 11**

Leader of the Deliverable: Dr. Pal Varga, AITIA

File Name: COMBO_D2.3_WP2_20June2014_AITIA_V2.0.docx

Version: V2.0

Authorisation code: PU = Public

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

Tel: + 33 2 23 27 12 46

E-mail: pointjc@jcp-consult.com

Project website address: www.ict-combo.eu

- - - End of Document - - -