Mobility and Bandwidth prediction in virtualized LTE systems: architecture and challenges

Georgios Karagiannis[1], Almerima Jamakovic[2], Keith Briggs[3], Morteza Karimzadeh[1], Carlos Parada[4], Marius Iulian Corici[5], Tarik Taleb[6], Andy Edmonds[7], Thomas Michael Bohnert[7]
[1]: University of Twente, the Netherlands, {g.karagiannis, m.karimzadeh}@utwente.nl
[2]: University of Bern, Switzerland, jamakovic@iam.unibe.ch
[3]: British Telecommunications plc, UK, keith.brigg@bt.com
[4]: Portugal Telecom Inovacao, Portugal, carlos-f-parada@ptinovacao.pt
[5]: Fraunhofer, Germany, marius-iulian.corici@fokus.fraunhofer.de
[6]: NEC Europe, Germany, Tarik.Taleb@neclab.eu
[7]: ZHAW/ICCLab, Switzerland, andrew.edmonds@zhaw.ch

***Abstract*— Long Term Evolution (LTE) represents the fourth generation (4G) technology which is capable of providing high data rates as well as support of high speed mobility. The EU FP7 Mobile Cloud Networking (MCN) project integrates the use of cloud computing concepts in LTE mobile networks in order to increase LTE's performance. In this way a shared distributed virtualized LTE mobile network is built that can optimize the utilization of virtualized computing, storage and network resources and minimize communication delays. Two important features that can be used in such a virtualized system to improve its performance are the user mobility and bandwidth prediction. This paper introduces the architecture and challenges that are associated with user mobility and bandwidth prediction approaches in virtualized LTE systems.**

Keywords—LTE; cloud computing; user mobility prediction; bandwdith prediction

I. INTRODUCTION

Mobile operators are currently focusing on providing technological solutions for the significant growth of mobile data traffic due to the continuous increase of mobile users, devices and new mobile applications. Long Term Evolution (LTE) can be considered as a promising cellular technology that could cope with this challenge. In particular, LTE represents the fourth generation (4G) technology that is standardized by the 3rd Generation Partnership Project (3GPP) and which is capable of providing high data rates, as well as support of high speed mobility. The LTE system consists of two main network parts, which are the Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) and the Evolved Packet Core (EPC). The E-UTRAN consists of base stations that are denoted as Evolved Node-Bs (eNodeBs or eNBs). The EPC is composed of several core network elements, where the main important ones are the Serving Gateway (S-GW), the Packet Data Network Gateway (P-GW) and the Mobility Management Entity (MME). The P-GW, which is the main mobility EPC anchor point, connects the EPC to external networks. Moreover, the P-GW also performs multiple functions such as, IP address/prefix allocation, policy control or charging. The S-GW supports the transport of the user data between the UE and the P-GW anchor point. The MME is the control node that processes the whole mobility management signalling, (e.g., handovers) between the User Equipment (UE) and the EPC.

The EU FP7 Mobile Cloud Networking (MCN) project [1], incorporates the use of cloud computing concepts in LTE mobile networks, in order to increase LTE's efficiency and performance. In particular, the integration of cloud computing benefits in an LTE system can be realized by: (1) extending the cloud computing concept beyond the typical (*macro*) datacenters towards new smaller (*micro*) datacenters that are distributed within the E-UTRAN and the EPC, and (2) deploying and running cloud-based (*virtualized*) E-UTRAN, denoted as RAN as a Service (RANaaS), and EPC, denoted as EPC as a Service (EPCaaS). The most important cloud computing principles integrated in this virtualized LTE system are the support of **on-demand provisioning** of LTE components and **on-demand elasticity**, allowing the virtualized LTE components to scale automatically, based on the data traffic load that they need to support. This scaling process can be controlled by real-time monitoring, or by predicting the mobility of LTE users as well as the bandwidth (data traffic) they are going to generate on certain locations and future moments in time. By monitoring the used resources on network components, the MCN management system can calculate the required computational and networking processing power. However, for abrupt changes due to a large movement of users (e.g. special events, like concerts or football games), the time elapsed to collect monitoring data could be quite high. Therefore, in such situations the prediction of (1) user mobility and (2) the generated bandwidth can be used, in combination with the monitoring process, to trigger the automatic scaling of resources. The capability to predict (1) how users are moving within a mobile network topology, and (2) the bandwidth that they are generating on certain locations, is a fundamental aspect to consider when virtualizing a LTE system. Both network parts, the access (RAN) and the core (EPC), can benefit from this type of prediction, getting prepared in anticipation for various types of traffic load scenarios. Consider, for example, a

virtualized EPC (EPCaaS) deployed in multiple datacenters along a serving country. Assume that, during the summer, people living in region A' are users that move on weekends to region B' (several tens of kms away) to spend some time on the beach. In this case, prediction systems are able to anticipate this social trend, such that some virtualized EPC components (S-GWs, MMEs or P-GWs) can be moved temporarily (during weekend) to datacenters in this region, increasing the resources in datacenter A' (in region A) and reducing them with the same amount in datacenter B' (in region B). In such situations, the prediction of (1) user mobility and (2) the generated bandwidth is an essential trigger that can be applied to start moving the networking processing power (components) in advance, to the regions where they are needed, and at the same time releasing resources when they are not required anymore.

Several mobility and bandwidth prediction algorithms have been presented in the literature, see e.g., [2].[3]. However, such prediction algorithms have not yet been applied to support on-demand elasticity in virtualized LTE cellular systems. This paper introduces the architecture and challenges associated with the Mobility and Bandwidth prediction approaches in virtualized LTE systems to support features, such as automatic scaling of LTE components.

The main research questions answered by this paper are:

(1). Which mobility prediction algorithms can be used in virtualized LTE systems?
(2). Which bandwidth prediction algorithms can be used in virtualized LTE systems?
(3). Which MObility and Bandwidth availability prediction as a Service (MOBaaS) architecture can be applied in virtualized LTE systems?
(4). What are the main challenges associated with MOBaaS in virtualized LTE systems?

This paper is organized as follows. Section II discusses the mobility and bandwidth prediction algorithms that can be applied in virtualized LTE systems; moreover, it answers questions (1) and (2). In Section III, the MOBaaS architecture and its service lifecycle are introduced; it also answers question (3). Section IV discusses the MOBaaS challenges and answers question (4). In section V the paper concludes and makes recommendations for future work.

II. MOBILITY AND BANDWIDTH PREDICTION ALGORITHMS

This section discusses the mobility and bandwidth prediction algorithms that can be applied in virtualized LTE systems.

A. Mobility Prediction Algorithms

A large number of different algorithms (strictly speaking, heuristics) have been proposed in the literature for forecasting the future position of mobile users, and for forecasting user mobility and bandwidth availability in neighbouring cells, all with the aim of making the handover process more efficient and a smoother experience for the user, see e.g., [2], [5]. It is also probably true to say that none of these heuristics is of proven efficacy or usefulness. There are also issues related to user privacy, and difficulties arising from the unreliability and/or unavailability of user positional data (for example, in areas with no GPS (Global Positioning System) signal). These heuristics generally fall into two classes, depending on whether the user is known to be constrained to a street network or not, and also whether a street map is available to the prediction algorithm. In the MCN project, we have compared and evaluated a large number of mobility prediction methods [4]. See Fig. 1 for a classification.

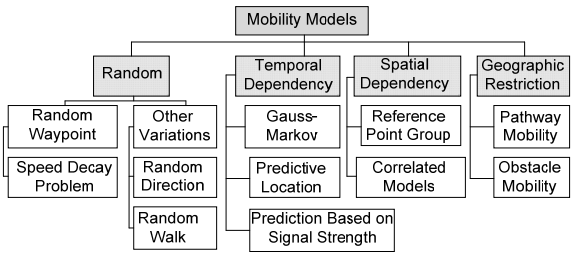

Fig. 1. Generic classification of mobility models, copied from [5]

Several mobility models with "Temporal Dependency", see Fig. 1, have been investigated. In the first (constrained) case, the "Predictive Location" model has been studied. In this model it is assumed that the mobility prediction algorithm has available a graph representing the street network topology. Then from the current user position x_i (which will be one of the graph nodes) a new position x_{i+1} can be predicted based on a known node-node probability transition matrix P. This matrix can never be known precisely, but estimates of its elements will have to be built up over time, by monitoring the positions of many users. By counting successful predictions, an element of machine learning can be introduced into the algorithm. Another approach (much less precise, but the street graph is not needed), is simply to use base-stations cells as the nodes of the graph.

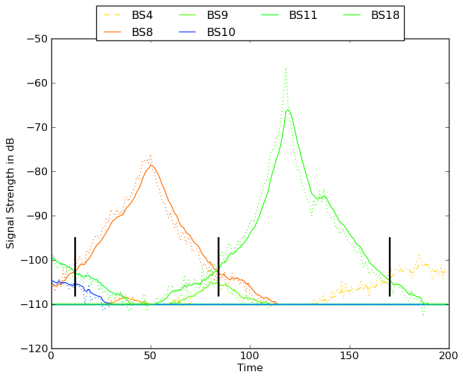

Fig. 2. Signal strength readings from a mobile user at several base stations

In the second (unconstrained) case, the "Gauss Markov" model has been studied. Such typically applied methods are just generic time series forecasting algorithms, which do not incorporate anything specific to the user mobility scenario. In other words, they are assuming no specific structure to the data used as inputs to the prediction. Methods of this type

include autoregressive (AR) time series models, and Kalman filters. Using such methods is purely heuristic, with no rigorous theoretical basis at all, and there are no guarantees about the performance. A final approach that has been studied is the "Prediction based on signal strength" model. From the performed studies we could deduce that in the long run this latter model will be more useful in virtualized LTE. The estimations of user positions can be based on signal strength readings, as reported by the user terminals to the base stations according to 3GPP protocols. Typical data from such readings is shown in Fig. 2. Not only can handover points by computed by comparison with a threshold, but the trend (up or down) of signal strength provides very valuable information to the algorithm about the direction of movement of the user.

B. Bandwidth Availability Prediction Algorithms

Over the last few years, several research activities have focused on bandwidth availability prediction models that are applied on older cellular systems. In a virtualized LTE system, bandwidth availability prediction models can be used to estimate the amount of bandwidth that will be used and is available on different parts of the LTE network topology in a future moment in time. A comprehensive bandwidth availability prediction model needs to monitor the available bandwidth at any time on different parts of the LTE network topology (e.g., the radio coverage cell, inter-radio coverage area, and inter-S-GW service area). For each C_i the available bandwidth at any T_j is denoted as $B_a(C_i, T_j)$. The mentioned parameter C_i could be a radio coverage cell, an inter-radio coverage area, an inter-S-GW service area, inter MME pool area or platforms where the virtualization of the LTE network entities can be realised, i.e., datacenters. The available bandwidth could be calculated considering the following:

- $B_t(C_i, T_j)$ the total bandwidth of C_i at time T_j,
- $B_u(C_i, T_j)$ the estimated bandwidth used by users at C_i at time T_j,
- $B_r(C_i, T_j)$ the bandwidth reserved at C_i for ongoing sessions at time T_j.

The available bandwidth at C_i at any time is calculated:

- $B_a(C_i, T_j) = B_t(C_i, T_j) - B_u(C_i, T_j) - B_r(C_i, T_j)$

In the MCN project, see [4] several bandwidth availability prediction methods have been compared. Typically, the existing bandwidth availability prediction models can be classified into two main categories: mobility non-predictive and mobility predictive schemes. In a **mobility predictive** scheme, the information of previous and/or current measured mobility factors, such as user location and direction of movement are used to estimate the available bandwidth. For instance, the solution proposed in [6] uses the location and direction of the user to estimate the next cells the user will be visiting. In particular, it calculates the probability of a UE of moving towards a new cell, depending on the UE's movement direction within the coverage area of the cell. Moreover, it takes also into account that the number of radio channels that can be reserved in a given cell has a close relationship to the number of users located in neighbouring cells. In another example, [7] describes a method used in cellular systems, based on the history of previously measured mobility factors, such as the spatial information (i.e., the location that the user has travelled) and temporal information (i.e., the time and day that data has been collected) maintained at the user's mobile host as local mobility profiles. Utilizing such data, this method is able to estimate the amount of bandwidth used in each cell and the amount of bandwidth that can to be served accordingly. On the opposite side, in a **mobility non-predictive** scheme, the mobility factors are not used during the bandwidth prediction process. In particular, the load generated by services is monitored and used in order to estimate the desired bandwidth for a user. This type of scheme can be, for example, applied to pre-reserve multimedia content delivery in Heterogeneous Wireless Networks, without requiring the knowledge of any mobility factors. An example of such a scheme is introduced in [8] and is built on the concept that the already existing Policy and Charging Rules Function (PCRF) in a LTE network can help estimating the resources a user is going to request. This is due to the fact that the PCRF stores information about how much load (bandwidth) a service will require in the future.

III. MOBAAS ARCHITECTURE AND ITS SERVICE LIFE CYCLE

This section introduces the MOBaaS (Mobility and Bandwidth Availability Prediction as a Service) architecture, and its service life cycle, see [4].

A. MCN Key Archtectural Entities and Service Lifecycle Stages

The MCN project has specified an architecture for Mobile Cloud Networking, see [9]. In this architecture all the functional elements are implemented as services. The MCN services are implemented by resources and these resources operate service related entities. These resources can be both physical or virtual (i.e., through some virtualisation technology). Furthermore, the entities that are used to represent certain aspects of service orientation are: (1) service, (2) service instance and (3) service instance components. In MCN the services can be classified into two main categories: (1) atomic services, i.e., virtualised computing, storage and networking, and (2) composed services, which can further be grouped in two main categories MCN (main) services and MCN support services. The service related entities are managed in a common consistent fashion regardless of their category by three key MCN architectural entities. The **Service Manager (SM)**, which offers to enterprise users multi-tenant capable services and an external interface, both programmatic and visual. Moreover, the SM is used in two dimensions; the business which encodes business agreements, and the technical that manages the different Service Orchestrators (SOs) of a particular tenant. The **Service Orchestrator (SO)** embodies how a service is actually implemented and is controlling the complete and end to end orchestration of a service instance (SI), including creation and scaling. The SO monitors metrics related to the service instance. The third key architecture entity is the **Cloud Controller (CC)** that is used to deploy, provision and dispose SOs. Moreover, the CC is responsible for cloud infrastructure resource provisioning and deployment, together with the management of virtual resources coordinating with the SO.

The technical lifecycle of a service can be separated in six stages. **Design**, where the service's technical design is carried out. **Implementation**, the service is implemented that includes the implementation of a SM and SO. **Deploy**, where the SO is deployed by the CC. **Provision**, the SO is instantiated and begins to create and provision the MCN services necessary to satisfy the SO's needs. **Runtime and Operation**, where the service instance is monitored and managed. During this stage the scaling in and out of components is carried out. Scaling in occurs when a component is releasing resources and scaling out occurs when new resources are allocated to a component. **Disposal**, is the lifecycle stage where the service instance's sub-components are disposed and deleted.

B. MOBaaS architecture

MOBaaS, see [4], is a MCN support service that assists a MCN service, such as RANaaS and EPCaaS, in predicting information regarding (1) the movement of individual end-users (estimated location of an individual end-user in a future moment in time), (2) the traffic that these individual end-users will be generating at a certain location in a future moment in time, (3) bandwidth available at a certain location in a future moment in time. The MOBaaS service architecture reference model can be seen in Fig. 3. In addition to the SM, SO and CC, this architecture encompasses the following service instance components. **Management**: operates in coordination with the SO and represents the component that is in charge of the management of the MOBaaS service. **Aggregation**: enables the fast collection of the datasets coming from Monitoring as a Service (MaaS), see [10] and from the Configuration Nodes. MaaS is a MCN support service that monitors the use of physical and virtual resources by other MCN services, see Fig. 4. **Frontend:** enables the configuration of MOBaaS and as well the way for MCN services and MCN support services to get the stored predicted data. **Configuration Node:** instantiated at each MCN service and MCN support service that requires the use of MOBaaS and it can be part of: (1) the MCN service management system, or (2) the SO to store, maintain and provide information required by the MOBaaS and it is not provided by MaaS, i.e., network graph and electronic map. **Prediction Data Storage**: collects and stores:

- monitored data (real time monitored and historical monitored data) via MaaS, including SLA (Service Level Agreement) related information;
- applied network graph (topology) and its parameters, obtained from other MCN services. In the MCN context, see [9], this graph is denoted as Service Template Graph and/or Infrastructure Template Graph);
- electronic road map, e.g., road map used by navigation systems.

Prediction Consumer: any other service instance component, which can be part of a MCN service or MCN support service, which can query and receive via the MOBaaS *Frontend*, the output (e.g., prediction of end-user location and prediction of bandwidth used and/or bandwidth available) generated by MOBaaS. **Mobility Prediction**: predicts the location of user in future moments in time, see Section 2. **Bandwidth Availability Prediction**: predicts the used bandwidth and/or available bandwidth in future moments in time, see Section 2.

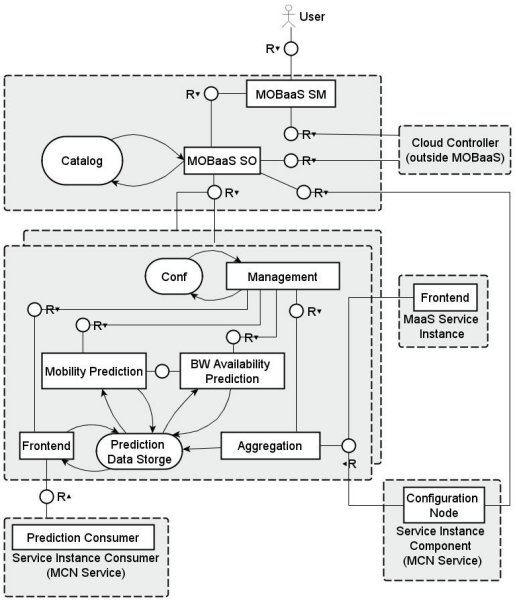

Fig. 3. MOBaaS Architecture Reference Model

As mentioned previously, MOBaaS is a support service for other MCN services. This implies that other MCN services, such as EPCaaS, can use MOBaaS in predicting information regarding (1) the mobility of users, (2) the bandwidth that they are generating and (3) the available bandwidth in the network.

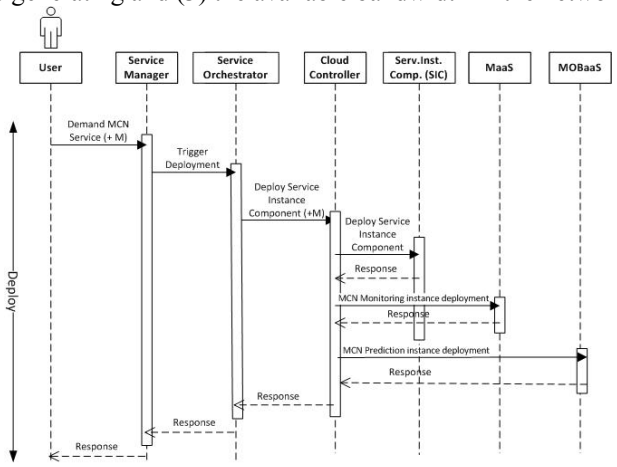

Fig. 4. MOBaaS deployment

Therefore, MOBaaS can be (1) considered as a specific service instance component of such a main MCN service and (2) follows the corresponding service lifecycle stages of that main MCN service, e.g., EPCaaS. As example, the deployment lifecycle stage of a MCN service that uses MOBaaS and MaaS as supporting services is shown in Fig. 4. After the MOBaaS is deployed and provisioned, it can predict and expose the requested MCN service system parameters,

e.g., the mobility of users, the bandwidth that they are generating and the available bandwidth in the network.

IV. MOBAAS CHALLENGES

In the virtualised LTE cellular system, the design of the mobile network core, i.e. the Evolved Packet Core (EPC), that displays the rapid and on-demand service elasticity, has been identified as one of the key design issues. During the Runtime and Operation phase, the SO of the EPCaaS runs decision algorithms to efficiently manage the deployed EPC service, e.g. to horizontally scale in/out one or more EPC Service Instance Components (SICs) to optimally place an EPC network function component in case of a massive users movement. Beside the internal logic, the SO's decision algorithms can use different data sources as input/trigger, where one of them is mobility and bandwidth availability prediction information, provided by MOBaaS service. The main challenges associated with the realization of the MOBaaS in cloud-based EPC, see also [4], are in a close relationship with the requirements to be fulfilled by the MOBaaS system:

- How to retrieve and process the traffic, location and movement of an end-user or group of end-users?
- How to estimate (with higher probability) the location of an end-user or group of end-users in a future moment in time?
- How to estimate (with higher probability) the traffic of an end-user or group of end-users at the predicted location in a future moment in time?
- How to estimate (with higher probability) bandwidth that will be available at the predicted location in a future moment in time?
- How to provide the accuracy of each estimated value in terms of confidence intervals?
- How to estimate the time until the predicted parameter value reaches the Target Parameter Value?
- Which kind of novel and accurate mechanisms should be used to make the MOBaaS service stable and other MCN services follow the prediction outcome seamlessly?
- How to design the MOBaaS system that is considered as being scalable with the increasing number of nodes in the system, reliable to the node and network failures, manageable from a single or several locations, adaptable to be used by different MCN (XaaS) services, configurable in order to increase the accuracy of the prediction, and simple for use by the consumer.

The MCN project identified and designed the MOBaaS architecture reference model as a first step to tackle the possible solutions for these challenges. Furthermore, several mobility prediction algorithms have been investigated, see Section II.A. During the next period of the MCN project, a first proof-of-concept of MOBaaS will be realized to demonstrate its feasibility.

V. CONCLUSIONS AND FUTURE WORK

The EU FP7 Mobile Cloud Networking (MCN) project [1], focuses on the integration of cloud computing concepts in LTE mobile networks, in order to increase LTE's efficiency and performance. One of the most important cloud computing principles integrated in this virtualized LTE system is the elasticity, which allows the virtualized LTE components to scale automatically, based on the data traffic load that they need to support. This automatic scaling process can be controlled by real-time monitoring, or by predicting the mobility of LTE users as well as the bandwidth (data traffic) they are going to generate on certain locations and future moments in time. This paper introduced the MOBaaS architecture, its service lifecycle and challenges. MOBaaS assists a MCN service, such as RANaaS and EPCaaS, in predicting user mobility and bandwidth usage and availability. The MOBaaS architecture and its lifecycle stages that have been defined in the first year of the MCN project hold a great promise. During the second year of the MCN project it is expected that a first version of MOBaaS will be implemented and verified.

ACKNOWLEDGEMENT

This work is accomplished in the context of the MCN project. We therefore, would like to thank the MCN project partners for their contributions and comments and acknowledge the European Commission, since the MCN project is an EC funded Integrated Project under the 7th RTD Framework Programme, FP7-ICT-2011-8-grant agreement number 318109.

REFERENCES

[1] EU FP7 Mobile Cloud Networking project, (visited in September 2013) http://www.mobile-cloud-networking.eu/site/.

[2] M. Abo-Zahhad, M. Ahmed Sabah, M. Mourad, "Future location prediction of mobile subscriber over mobile network using Intra Cell Movement pattern algorithm", Proc. of 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), pp. 1-6, 2013

[3] S. Rashad and J. Bradley, "SmartMobiMine: Smart mobile data mining techniques to support 4G mobile networks", Proc. of IEEE Consumer Communications and Networking Conference (CCNC), pp. 703-704, Jan. 2011.

[4] MCN D4.1, "Mobile Network Cloud Component Design", European Commission, EU FP7 Mobile Cloud Networking public deliverable (available at [1]), 2013.

[5] F. Bai and A. Helmy, "A survey of mobility models", Wireless Ad Hoc and Sensor Networks, Kluwer academic Publishers, 2004

[6] J. Hou and Y. Fang, "Mobility-based call admission control schemes for wireless mobile networks" Wireless Communications and Mobile Computing, Vol. 1, Iss. 3, pp. 269-282, July 2001.

[7] S. Rashad, M. Kantardzic, and A. Kumar, "A data mining approach for call admission control and resource reservation in wireless mobile networks", Advances in Data Mining, pp. 134 -143, 2005.

[8] A. Al-Hezmi and M. Corici, "Network Resources Pre-Reservation for Multimedia Content Delivery in Heterogeneous Wireless Networks", Proc. of 11th European Wireless Conference 2011 - Sustainable Wireless Technologies (European Wireless), pp. 515-521, 2011.

[9] MCN D2.2, "Overall Architecture Definition, Release 1", European Commission, EU FP7 Mobile Cloud Networking public deliverable (available at [1]), 2013.

[10] MCN D3.1, "Infrastructure Management Foundations - Specifications and Design for Mobile Cloud framework", European Commission, EU FP7 Mobile Cloud Networking public deliverable (available at [1]), 2013.