Toward Carrier Cloud: Potential, Challenges, and Solutions

Tarik Taleb

Abstract

Mobile operators are in need of means to cope with the ever increasing mobile data traffic, introducing minimal additional capital expenditures on existing infrastructures, principally due to the modest average revenue per user. Network virtualization and cloud computing techniques, along with the principles of the latter in terms of service elasticity, on-demand, and pay-per-use, could be important enablers for various mobile network enhancements and cost reduction. This article discusses the recent trends the mobile telecommunications market is experiencing, showcasing some of the emerging consumer products and services that are facilitating such trends. The article also discusses the challenges these trends present to mobile network operators. It also demonstrates the possibility of extending cloud computing beyond data centers toward the mobile end user, providing end-to-end mobile connectivity as a cloud service. The article introduces a set of technologies and methods for on-demand provision of a decentralized and elastic mobile network as a cloud service over a distributed network of cloud computing data centers. The concept of Follow-Me-Cloud, whereby not only data but also mobile services intelligently follow their respective users, is also introduced. The novel business opportunities behind the envisioned carrier cloud architecture and service are also discussed, considering various multi-stakeholder scenarios.

Introduction

The objective of this article is to marry the cloud domain and the telecom domain, highlighting why there is need for such a marriage and explaining how it can happen. For that purpose, it is important to understand the current landscape of both the mobile communications and cloud arenas. For the former, it is generally agreed that there is a global commitment from major operators to Long Term Evolution (LTE) in order to satisfy their continuous push for a new and fast radio technology. Indeed, the continuous need for higher data rates, shorter end-to-end communication delays, and short latencies for connection setup have all led to specifications of diverse access technologies. The diversity of these accesses was the driving force for what is perhaps the most significant transformation of mobile and wireless network systems since the emergence of digital cellular telephony. This has given birth to a new architecture, previously referred to as System Architecture Evolution and currently called Evolved Packet System (EPS) [1].

Simultaneously with the great changes in the mobile network architecture and its technologies, user equipment technologies have been progressing at a much faster speed. Smartphones are in fact even racing ahead of mobile networks, supporting diverse operating systems, and offering both users and developers a wide plethora of tools to generate thousands, if not millions, of mobile applications. The business of tablets is also growing at an ever accelerating pace, and forecasts indicate that there will be even much faster growth in the number of LTE devices, be they smartphones, tablets, or even LTE-modem-connected laptops.

From the mobile application and services side, mobile multimedia streaming services are gaining great momentum among the mobile user community. While it is incontestable that mobile multimedia services demand huge mobile resources, there are many "small" applications that do not involve the exchange of high amounts of data per individual mobile user, but cause big "headaches" to mobile operators [2]. An important feature of these mobile applications is that they are based on a one-to-many communication paradigm. Classical examples of such applications include news tickers, social network applications (e.g., Twitter), and location-based check-in services (e.g., Foursquare).

This wide variety of applications is changing the basic user equipment device (UE)-related assumptions, based on which mobile networks have been designed. Indeed, so far, a UE, in the context of the third generation (3G), is a mere phone with some intelligence to receive packet-switched (PS) services on top of a circuit-switched (CS) network. It has been assumed that it is in either active or idle state, but most of the time in idle mode. With the arrival of PS mobile networks such as LTE and the wide popularity of the above-mentioned applications, lots of background traffic is exchanged to keep sessions of these applications alive, changing the status of

The author is with NEC Europe Ltd.

1536-1284/14/$25.00 © 2014 IEEE

UEs toward the "always active" paradigm. In addition, UEs are no longer restricted to mobile phones, but may also include tablets, LTE-modem-equipped laptops, and so on. Furthermore, currently, users are assumed to connect to WiFi when they are indoors (e.g., at home, in an office, or in a mall) and use a mobile network only when they are outdoors. However, with the deployment of small cells, the coverage of mobile networks is expanding from outdoor only to also include indoors [3]. While this represents some great business opportunities for both telecom equipment vendors and mobile operators, they introduce important challenges. These challenges principally pertain to how to cope with the "IP tsunami" of mobile traffic, hitting both control and user data planes of EPS. Indeed, even if no data is exchanged, maintaining the session of a mobile application alive would require the periodic exchange of hundreds of signaling messages (i.e., background traffic) among multiple mobile network components, and that is as part of diverse procedures [1]. The deployment of millions, if not billions, of machine type communication (MTC) devices would make the problem more complicated, despite the potential business opportunities MTC exhibits [4].

All in all, every one of the above mentioned changes happening at the UE technology level as well as at the behavior level of an ever growing community of mobile users, expecting ubiquitous connectivity for all mobile application types, are resulting in an important "mobile IP tsunami." The question is how operators are tackling it. One principal approach is to throttle specific types of traffic (e.g., YouTube) during specific periods of time (e.g., peak hours). Other solutions consider transcoding and caching over the top (OTT), using tools such as video quality measurement (VQM) and setting the amount of reduction in a video file size. While this kind of solutions may go unnoticed for some time, users may ultimately realize the tricks as their quality of experience (QoE) may be impacted, and operators may risk losing some of their subscribers. Other solutions proposed in the recent literature, intended to cope with the increase in mobile data traffic and the limited resources at the physical layer, tend toward the direction of using multiple wireless interfaces (e.g., LTE and WiFi) and steering traffic among them based on service and application type (e.g., voice over LTE and YouTube traffic over WiFi) to ensure acceptable QoE/quality of service (QoS) to a large number of users sharing the wireless medium [3].

In addition to protocol-level solutions, operators have also been looking into architectural solutions that enable them to decentralize/distribute, to a certain extent, the user plane of their network architecture by moving data anchor gateways to the network edge [5] and selectively offloading IP traffic as near to the edge of operators' networks as possible through offload points at nearby radio access networks (RANs) (Figs. 1a and 1b) [6, 7]. Others considered the edification of mobile content delivery networks (CDNs), placing content close to the network edge and enabling some level of self-organization among caches for the sake of load balancing, energy saving, and so on (Fig. 1c). While this intuitively yields important benefits to end users as they access popular content from

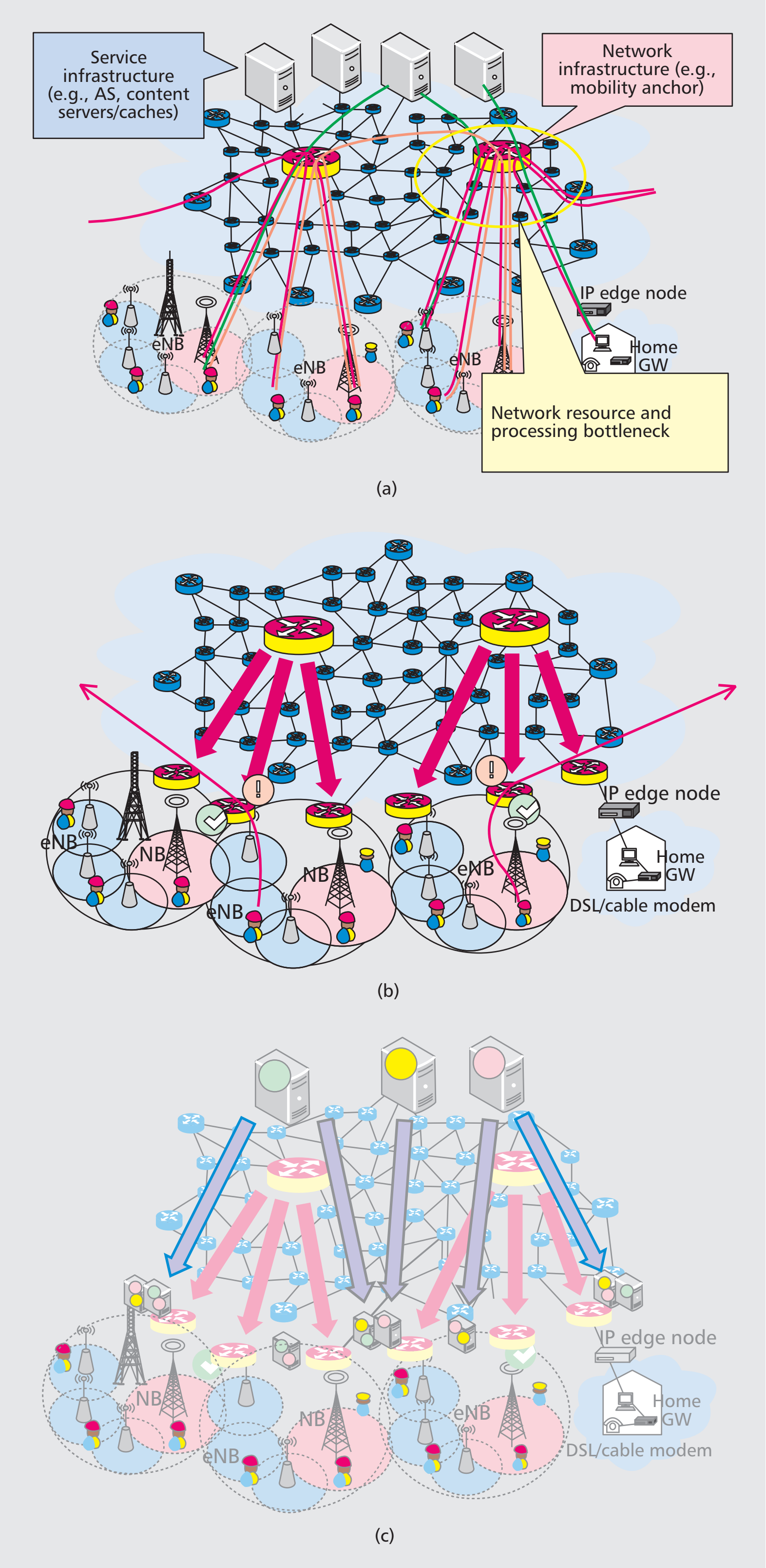
Service infrastructure (e.g., AS, content servers/caches)
Network infrastructure (e.g., mobility anchor)
IP edge node
Home GW
eNB
eNB
eNB
Network resource and processing bottleneck
(a)
IP edge node
Home GW
DSL/cable modem
eNB
NB
eNB
NB
eNB
(b)
IP edge node
Home GW
DSL/cable modem
NB
NB
eNB
eNB
(c)

Figure 1. Architectural solutions for mobile networks to cope with mobile IP traffic growth: a) current mobile core networks, highly centralized; b) decentralizing the mobile core network and selectively offloading IP traffic; c) building and decentralizing the mobile content delivery network.

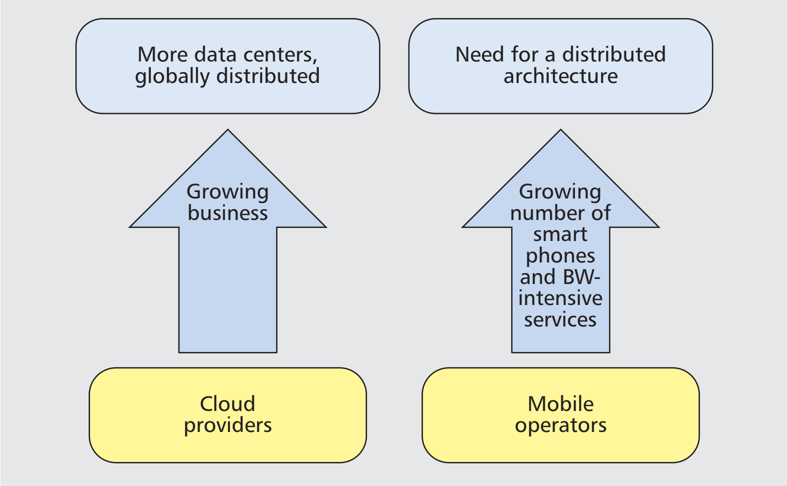
More data centers, globally distributed
Need for a distributed architecture
Growing business
Growing number of smart phones and BW-intensive services
Cloud providers
Mobile operators

Figure 2. Trends in cloud computing and mobile telecommunications.

nearby caches via nearby data anchor gateways, these solutions require significant investment into the infrastructure as operators need to massively invest in the backhaul speed, upgrade and purchase more network nodes, and design a scalable decentralized network to accommodate peak hours. In addition to the important requirement of energy consumption, this will require massive investment, which unfortunately cannot bring an immediate return on investment, particularly due to the relatively low increase in the average revenue per user (ARPU).

On the other hand, cloud computing is gaining great momentum. Indeed, it is a fast growing business with many emerging players (Amazon, Verizon Terremark, Salesforce, Rackspace, etc.). The size of its market is on the order of hundreds of billions of U.S. dollars. Statistics also indicate that an important portion of companies are using or considering the use of cloud services, moving their provided services and applications to the cloud. This is effectively thanks to the nice features cloud computing offers, summarized in on-demand self-service support, elasticity support, multi-tenancy support, and pay-as-you-go fair billing. Cloud computing offers three main business models: infrastructure, platform, and software as a service (I/P/SaaS). The flexibility these three business models come with offers users huge economic potential, enabling them to significantly reduce their capital expenditure (CAPEX) and operational expenditure (OPEX) costs. The promising business of cloud computing and the need for it have even stimulated multiple open sources (e.g., KVM, Xen, and OpenVZ), different open sources for IaaS (e.g., OpenStack, CLoudStack, and Eucalyptus Systems), and different open sources for PaaS (e.g., Cloud Foundry, SmartOS). Different standardization activities relevant to cloud computing are also taking place, with the Open Cloud Computing Interface (OCCI) being the leader. The importance of the cloud computing business, the tremendous demand for its services, and the rapid growth of its market are even stimulating the deployment of multiple distributed regional data centers [8, 9].

Putting the above-mentioned observations together, as depicted in Fig. 2, on one hand, cloud providers are distributing their "cloud/network," globally deploying more regional data centers to meet the cloud business demands. On the other hand, mobile operators need to decentralize their architecture to cope with the growing number of smartphones and bandwidth-intensive mobile services, all while ensuring minimal investment in infrastructure due to the relatively low ARPU. The billion dollar question, defining the core of this article, is the following. Currently, clouds are limited to offer computing and storage as a service. Why not use clouds to build mobile networks and help decentralize mobile networks on demand, elastically, and in the most cost-efficient way? This shall help operators to minimally invest in building virtualized mobile networks (i.e., carrier cloud) and grow on demand. It shall be noted that many telcos and carriers already own clouds. Some telco cloud deployments are only for internal use. Others are selling it as a service, either owning the cloud infrastructure or through partnership with cloud providers (e.g., Verizon offering Terremark, Deutsche Telekom partnering with EdgeCast, Bell with Limelight, Telecom Italia, BT, Telefonica).

The remainder of this article is structured as follows. The following section highlights some enabling technologies and portrays a high-level diagram of the carrier cloud architecture along with its main components. It also highlights some of the key issues along with a high-level description of solutions to tackle them. The third section describes some of the potential use cases of carrier cloud. Finally, the article concludes in the last section.

Carrier Cloud

Enabling Technologies: Network Function Virtualization

Carrier network nodes traditionally refer to dedicated hardware boxes that are single-service and single-tenant boxes with well defined functional behavior, and well defined and standardized external interfaces. They are built as a network function (i.e., code) running on a particular operating system on top of a dedicated hardware platform. The concept of network function virtualization (NFV) aims to decouple the software part from the hardware part of a carrier network node, using virtual hardware abstraction techniques. The objective is to run network functions as software in standard virtual machines (VMs) on top of a virtualization platform in a general-purpose multi-service multi-tenant node (e.g., a carrier grade blade server). A suitable software defined networking (SDN) technology can be used to interwork between the different virtualized network functions on the different VMs within the same data center or across multiple data centers (Fig. 3).

NFV is deemed to give a high degree of flexibility to mobile operators in the deployment of their mobile networks on the cloud, as the same general-purpose physical node can dynamically run different network functions (and services) on multiple virtual instances as per the needs and requirements of the mobile operator. Cost-efficient scalability is achieved with no major effort as elasticity is an intrinsic feature of cloud computing. Rapid deployment of mobile services can be guaranteed as network functions can be run on demand and in a dynamic way on VMs. As an attempt to advance work on NFV, many operators and vendors have started standards activities, for example, creating an Industry Specifications Group (ISG) to understand technical challenges for NFV, and to recommend and define strategies to impact relevant standard development organizations [10]. Among many, a highly important issue pertains to the adaptability and resiliency of virtualized/soft network functions originally designed for a (relatively) static mobile network to the dynamicity, frequent redistribution, and scaling of the carrier cloud.

Substantive research work has been carried out in the recent literature on adopting this concept of NFV to virtualize mobile networks. The principal focus of most of these research works has been on the virtualization of IP routing functions, particularly to decouple the control and user data planes of mobile networks, principally using OpenFlow [11, 12]. Further research work has analyzed the impact of EPC nodes on both planes to decide on the routing function virtualization [13], and others have proposed possible architectural modifications to EPC to render its key components "instantiable" in data centers [14]. Some other research work further describes testbeds and experimental environments to validate the concept of OpenFlow-based decoupling of control and user data planes along with the support of mobility [12, 15]. While the work presented herein can certainly benefit from the findings of these research works, this article advances the state of the art by describing a complete architecture that supports the on-demand creation of cloud-based elastic mobile networks, detailing interactions occurring among the resource controllers of the different stakeholders that shall enable the vision of the carrier cloud.

Carrier Cloud Architecture

As explained earlier, the overall objective of the carrier cloud is to enable the on-demand edification of a carrier-grade mobile network on the cloud in an elastic way. In other words, we intend to enlarge the service domain of cloud computing from the traditional way, whereby storage and computing are provided as a service at data centers, to the provision of mobile connectivity (as well) as a service.

Figure 4 shows the most important stakeholders that shall enable the vision of the carrier cloud. The physical infrastructure may be owned by the same or different providers. It may consist of a public network connecting different data centers across a particular geographical area. These data centers can be large-scale or regional [8, 9]. The server racks within a data center may be connected through a private network. Using the underlying physical infrastructure, a virtual infrastructure may be created with the help of one or multiple cloud infrastructure providers (CIPs), that is, a classical CIP providing IaaS, such as Amazon EC2). A platform may be offered as a service on top of the virtual infrastructure by one or multiple cloud service providers. It shall be noted that the physical infrastructure provider, CIP, and CSP could be owned by the same or different parties. Using 3GPP terminology, a packet data network (PDN) may be running on a platform or an infrastructure provided by a CSP or CIP, respectively. In that case, we hereafter refer to it as the PDN cloud. It may also be independent of the cloud and built in a traditional way. At the carrier cloud service platform (CC-SP) layer, two new roles are introduced: mobile RAN cloud and mobile core network cloud service platform providers (MRAN-CSP and MCN-CSP). The former provisions mobile RAN functions, runs adequate intelligence to decide where to place them on the infrastructure or platform provided by the CIP or CSP, respectively, and ensures adequate resources (i.e., compute and storage) from either the CIP or CSP to run the mobile RAN functions. This service can be provided with or without mobile RAN system management. The MCN-CSP provisions mobile core

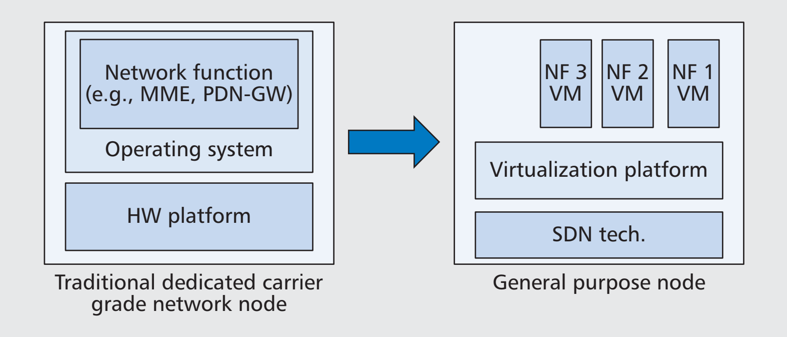

Figure 3. The concept behind network function virtualization.

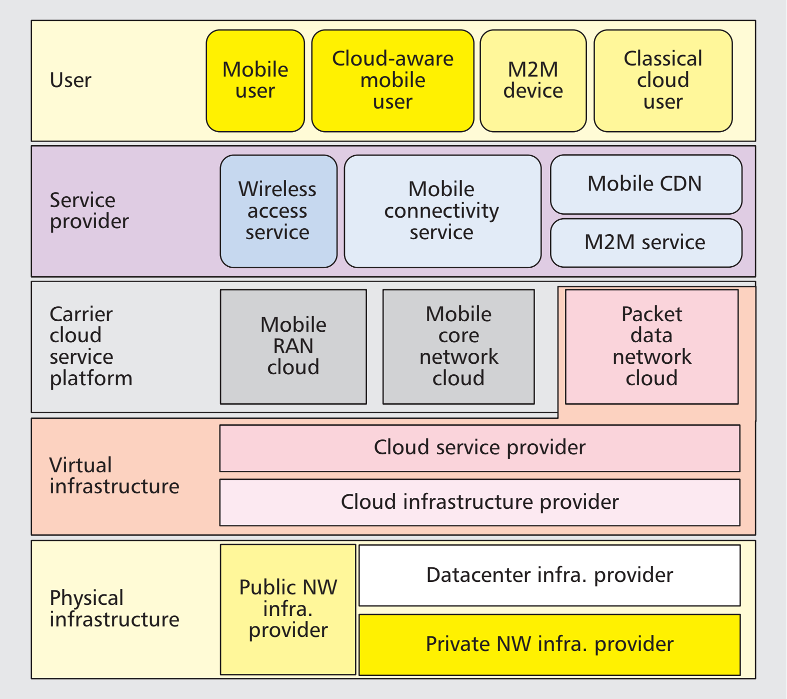
User
Mobile user
Cloud-aware mobile user
M2M device
Classical cloud user
Service provider
Wireless access service
Mobile connectivity service
Mobile CDN
M2M service
Carrier cloud service platform
Mobile RAN cloud
Mobile core network cloud
Packet data network cloud
Virtual infrastructure
Cloud service provider
Cloud infrastructure provider
Physical infrastructure
Public NW infra. provider
Datacenter infra. provider
Private NW infra. provider

Figure 4. Key stakeholders of the carrier cloud architecture.

network functions, runs adequate intelligence to decide where to place them on the infrastructure or platform provided by CIP or CSP, respectively [16], and ensures adequate resources (i.e., compute and storage) from either the CIP or CSP for running the mobile core network functions. This service can be also accompanied with or without mobile core network system management. The carrier cloud service platform (CC-SP) provider ensures that in case a mobile core cloud platform and a mobile RAN platform are provided (by MCN-CSP and MRAN-CSP, respectively) to the same mobile service provider, these two platforms are adequately integrated and interworked following existing standards. The network system management of both platforms during runtime may also be provided as a service. Different service providers may benefit

A mobile virtual network operator can be also built using a mobile core network platform provided by MCN-CSP and hosting its mobile services at the premises of CSP and/or CIP. Different end-users can be envisioned for the service providers using the service platforms provided by CC-SP.

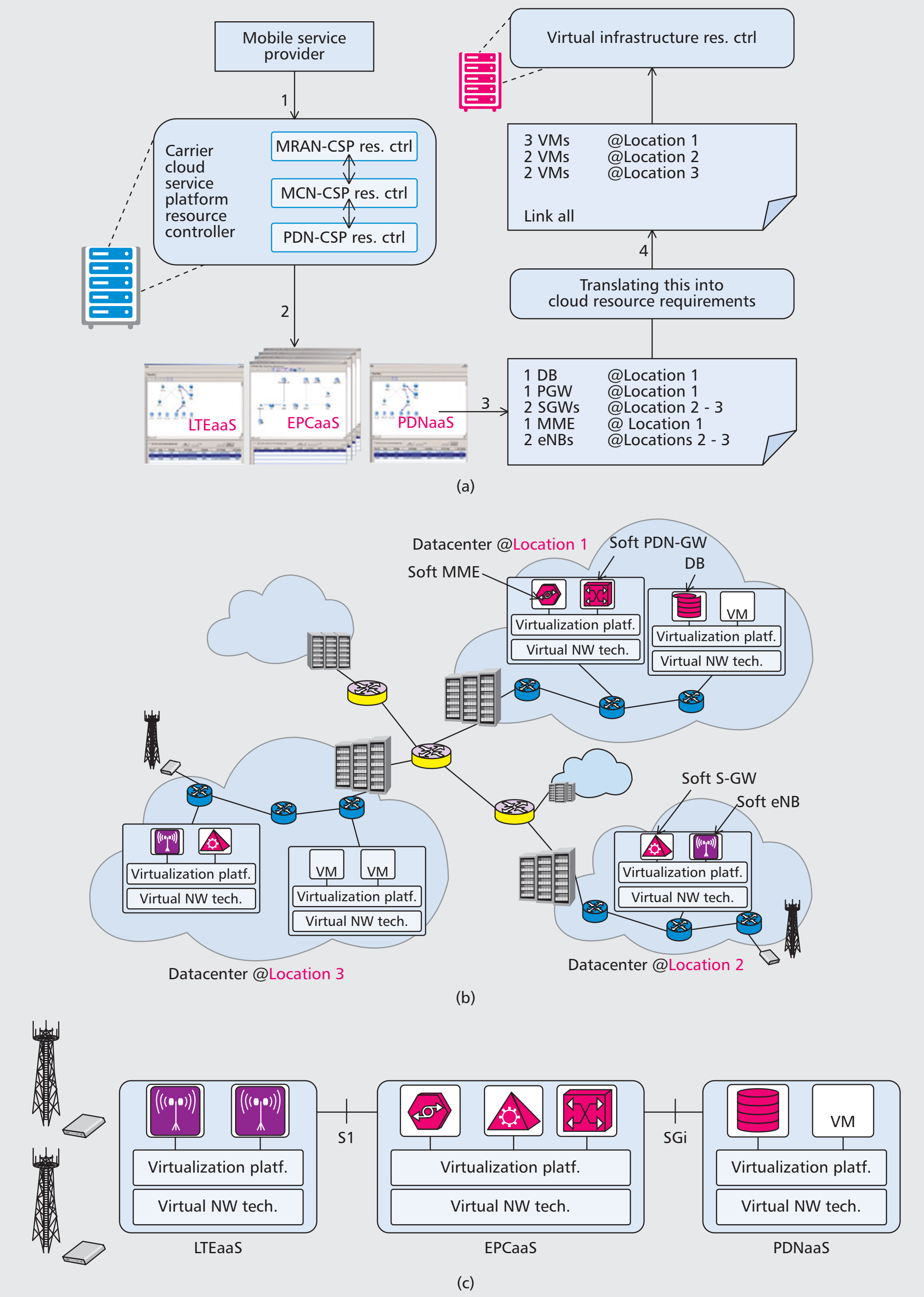

Figure 5. Carrier cloud how-to: (a) major steps to deploy a mobile service and a mobile network on the cloud; (b) placing and running soft network functions on obtained VMs at the different locations; (c) outcome network: a full end-to-end mobile network provided on the cloud.

from the service platforms provided by CC-SP. A wireless access service provider may use the platform provided by the MRAN-CSP to provide local wireless connectivity service to enable radio communications among, for example, organizers of a local event or large-scale concerts (e.g., enabling device-to-device, D2D, communications). A mobile network operator (MNO), with both its mobile core network and its RAN running on the cloud, can benefit from the service platforms provided by both the MRAN-CSP and MCN-CSP. The MNO's network management can be either carried out by the MNO or offered as a service by the CC-SP. A mobile virtual network operator (MVNO) can also be built using a mobile core network platform provided by an MCN-CSP and hosting its mobile services at the premises of the CSP and/or CIP. Different end users can be envisioned:

- A conventional cloud service user using the cloud service or PaaS provided by a CSP/CIP
- A conventional mobile service user using the mobile connectivity service provided by a mobile service provider using platforms provided by a CC-SP
- Semi-advanced CC users using elastic mobile connectivity service provided by a mobile service provider using platforms provided by a CC-SP
- An advanced CC user simultaneously using both elastic mobile connectivity as a service provided by a CC-SP and compute/storage as a service provided by a CSP/CIP

Carrier Cloud: How To?

The remainder of this section describes major steps behind the deployment of a mobile network on the cloud. As an example, we use the case in which both mobile service and mobile connectivity are provided in the cloud; that is, the PDN, mobile core network, and mobile RAN are all deployed in the cloud.

As a first step, the mobile service provider, a customer of the carrier cloud service platform provider, expresses its requirements to the latter. These requirements can be explicitly expressed, for example, specifying that adequate resources are required to serve a given number N_1 of subscribers in location 1, N_2 subscribers in location 2, and N_3 subscribers in location 3. During the runtime of the virtual mobile network platform for the mobile service provider, its requirements may be dynamically and automatically assessed based on an autonomic prediction of the needs of the subscribers or users of the mobile service provider. They can also be assessed by the mobile service provider based on expectations after the launch of a new service, the subscriptions of specific user types (e.g., VIP users) with specific QoS requirements, and so on. These requirements are then used by the carrier cloud service platform resource controller to decide the optimal network configuration to satisfy the needs of the mobile service provider. This operation is carried out while consulting the MRAN-cloud service platform resource controller, MCN-cloud service platform resource controller, and PDN-cloud service platform resource controller, ensuring that the outcome network configuration is interoperable and standards-compliant. At this stage, the service platform resource controllers translate the requirements from the mobile service provider into a network configuration, deciding where mobile RAN network functions, mobile core network functions, and PDN "caches or servers" need to be placed. The placement of these functions is of utmost importance and shall be based on different metrics (e.g., application type, data center location, data center load, and end-to-end QoS/QoE) that shall render the overall end-to-end communications optimal [3, 16, 17]. In steps 2 and 3, the decision of the carrier cloud service platform resource controller is a mere network configuration that specifies "what shall be placed where." In the example of Fig. 5, the outcome network configuration is one database (forming a PDNaaS) and one PDN gateway (P-GW) in location 1, two serving gateways (S-GWs) in locations 2 and 3, and one mobility management entity (MME) in location 1 (forming an EPCaaS: Evolved Packet Core as a service), and two soft eNBs in locations 2 and 3 (forming an LTEaaS). This network configuration is then translated into requirements that can be understood by the underlying cloud service provider. In the example of Fig. 5a, the requirements are translated into three VMs in location 1, two VMs in location 2, two more VMs in location 3, and a virtual network connecting the seven VMs.[1]

Once these resources are secured from the cloud service provider (e.g., using a suitable open source such as OpenStack), the carrier cloud service platform resource controller places and runs the virtualized network functions (i.e., images thereof) in the created VMs as per the determined network configuration, as illustrated in Fig. 5b. Assuming that the virtualized network functions are standards compliant, the outcome network is an end-to-end mobile network as shown in Fig. 5c, through which LTEaaS is connected to EPCaaS via the well standardized S1 interface, and EPCaaS is, in turn, connected to PDNaaS using the standardized SGi interface. It shall be noted that whenever there is a need for more resources at any of the three network blocks (i.e., LTEaaS, EPCaaS, PDNaaS), the steps of Fig. 5a are followed to assess the user and network elasticity, and to determine the amount and type of resources required, the kind of network functions to be run, and the locations where to be placed. This operation ensures the dynamic and on demand creation of a fully elastic end-to-end mobile network on the cloud.

Follow Me Cloud: Toward Better Interworking of EPCaaS and PDNaaS

As explained earlier, with efficient algorithms for placement and scaling of mobile core network and mobile service functions [16], the vision of carrier cloud shall enable optimal distribution of the mobile core network, allowing user equipment to be always connected to optimal data anchor gateways (i.e., the placement of a data anchor gateway will take into account the geographical distribution of users and their behavior in the usage of mobile services) [7, 18]. This shall ensure optimal mobile connectivity service (i.e., optimality of the path between the

With efficient algorithms for placement and scaling of mobile core network and mobile service functions, the vision of carrier cloud shall enable optimal distribution of the mobile core network, allowing user equipment to be always connected to optimal data anchor gateways.

[1] *In this example, for the sake of simplicity, we assume that all requested VMs have the same characteristics. However, they may have different characteristics, particularly those running different virtualized network functions.*

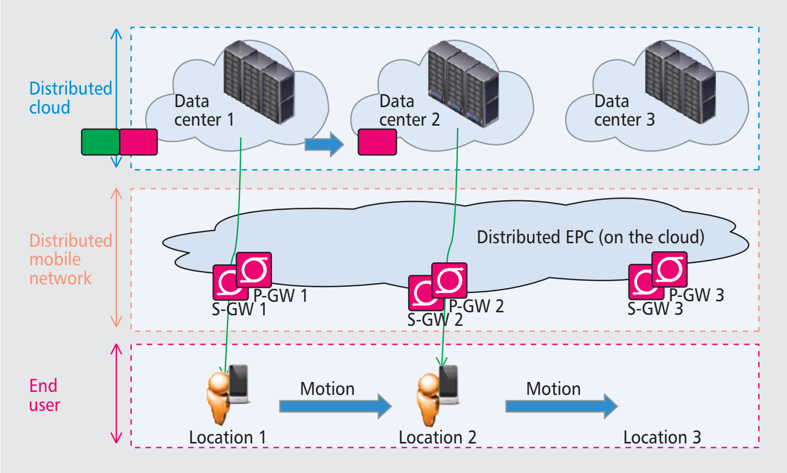

Figure 6. Follow Me Cloud: achieving optimal end-to-end mobile connectivity.

UE and the data anchor gateway path). However, this may be inefficient in the absence of an optimal end-to-end connectivity. Indeed, it is likely that at the beginning of a service session, a user gets connected through an optimal data anchor gateway (e.g., P-GW1 in Fig. 6) and starts receiving the service from a nearby data center (e.g., data center 1 in Fig. 6). However, upon its movement to a different location, the user connects to an optimal data anchor gateway (e.g., P-GW2 in Fig. 6) but keeps receiving the service from data center 1. While the mobile connectivity from the user to the mobile data anchor gateway is always optimal [7, 18], the end-to-end mobile service is not optimal as the user keeps receiving the service from a distant data center after its movement to a different location. The concept of Follow Me Cloud is to enable not only the content, but also the service to follow the user during her movement. Thus, the service will be always accessed from the optimal data center and via the optimal data anchor gateway, ensuring an optimal end-to-end connection [19].

Service migration, through VM migration and during runtime (i.e., a highly costly operation), has recently become technologically possible using different technologies and supporting VM relocation at the network level at either layer 2 or layer 3. The former intuitively takes place in case there is Ethernet continuity between DCs; Transparent Interconnection of Lots of Links (TRILL) and Shortest Path Bridging (SPB) can be used. The latter can be supported by mobile IP or Locator/Identifier Separation Protocol (LISP). For mixed L2/L3 VM relocation, virtual extensible LAN (VXLAN), Network Virtualization Using Generic Routing Encapsulation (NVGRE), or Stateless Transport Tunneling (STT) can be used, for example, integrated in VMware NSX only for intra-DC VM migration. Inter-DC VM migration can be enabled using products integrating LISP (e.g., Cisco and VMware joint product [20]), under some operating conditions (e.g., max distance between DCs, min required bandwidth). Inter-DC VM migrations can also be achieved; with negligible down times, using a pure control plane solution triggered by the hypervisor (e.g., KVM) as in [21]. Despite the ongoing progress in the area of VM migration, service migration control and its orchestration logic (i.e., when to migrate what) remain the most missing elements. Resource management (i.e., resource slicing, sliding, and cleanup), fast and secure VM migration, and service migration transparency to the user also define interesting and challenging research problems that are worth investigating. One important aspect that shall render service migration control easier pertains to algorithms for efficient user service and content distribution/replication, deciding where and when to place which, adopting numerous smart caching techniques. The service migration control can also benefit from efficient interworking functions between the "service layer" (i.e., PDNaaS) and the mobile core network layer (i.e., EPCaaS) of the overall network, implementing data-center-aware data anchor gateway selection mechanisms [19]. Using the architecture illustrated in Fig. 5, should a Follow Me Cloud logic be implemented at the PDN-cloud service platform controller with the necessary interworking functions with the MCN-cloud service platform controller, these issues can easily be solved [19]. While the work presented in [19] achieves the objectives of the Follow Me Cloud concept without the use of any SDN technologies, an OpenFlow-based implementation of Follow Me Cloud is described in [22]. Furthermore, a Markov-chain-based analytical model of the concept is available in [23].

Putting It All Together: High-Level Architecture

Figure 7 depicts a high-level diagram of the CC architecture, illustrating the main units of the MRAN-CSP resource controller, the MCN-CSP resource controller, and the PDN-CSP resource controller, along with interfaces between the different units, and toward the cloud controller and the cloud. For the sake of simplicity, we assume that all data centers belong to the same cloud provider, avoiding the complexity that might otherwise come with multi-cloud providers or cloud federations. However, it is stressed that the envisioned architecture does not exclude the case of multi-cloud providers. Indeed, it can be built on any cloud platform that is, in turn, built on multiple clouds (i.e., federated cloud) using adequate cloud brokering and technologies. Devising these technologies is outside the scope of this article.

***MCN-CSP Resource Controller* —** The MRAN-CSP resource controller consists of seven units, referred to as UR1-7 (i.e., UR: Unit for RAN); the role of each is described below.

UR1 — Service Provider Requirement Assessor: This is a module for assessing requirements from the different service providers using an MRAN-CSP. This assessment can be based on either explicit feedback from the different service providers (using interface R1) or autonomic prediction of their requirements. Different protocols (e.g., Diameter) and different implementation options can be envisioned for interface R1, and how the requirements from the different

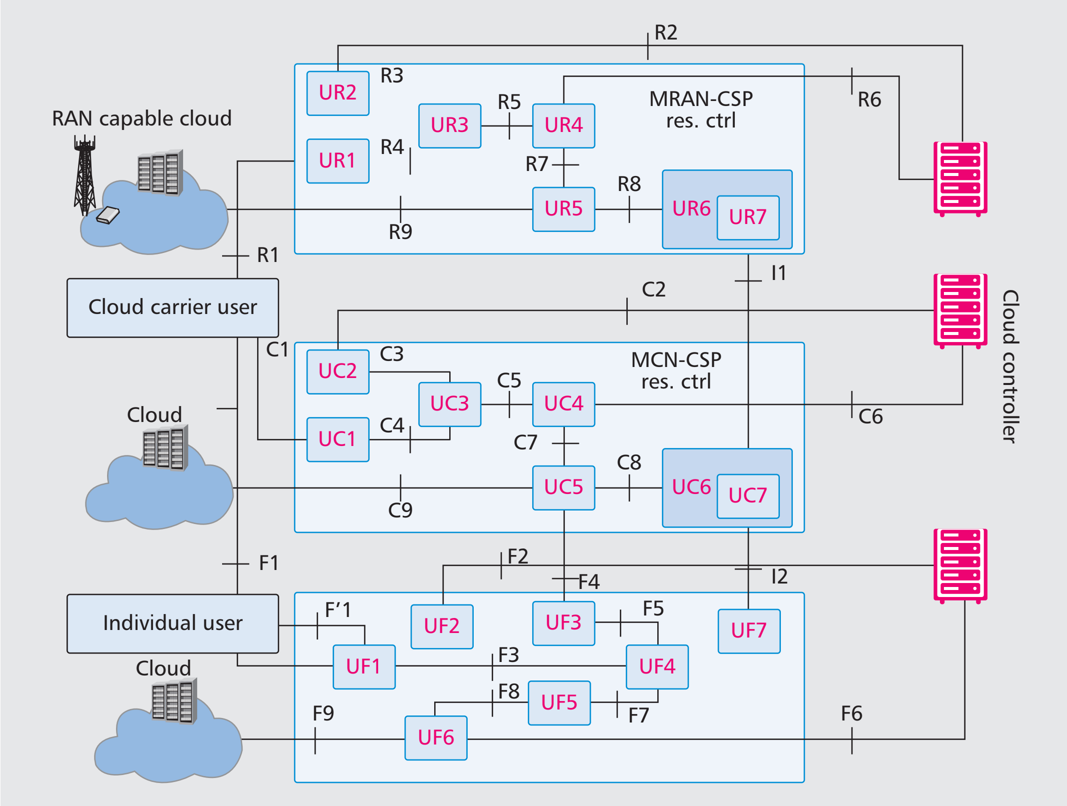

Figure 7. High-level architecture.

The concept of Follow Me Cloud is to enable, not only the content, but also the service to follow the user during her movement. In this way, the service will be always accessed from the optimal data center and via the optimal data anchor gateway, ensuring an optimal end-to-end connection.

CC users are communicated and negotiated. These implementation options are outside the scope of this article.

UR2 — VI Resource Assessor: This is a module for receiving and processing information (monitored by the VI cloud controller) about the resource availability in the underlying RAN-capable cloud. The monitoring information is exchanged using interface R2.

UR3 — RAN Network Configurator: This develops knowledge on the resources of the RAN-capable cloud from UR2 using interface R3, understands requirements from service providers from UR1 using interface R4, and accordingly computes a RAN network configuration that shall adequately meet the requirements of a service provider given the underlying cloud resources. UR3 employs different algorithms that decide where to place each RAN network function (e.g., eNB and RNC: radio network controller). It may be called on for initial deployment of a RAN network on the cloud for a service provider, migration of network functions during runtime of the RAN network, extension/shrinkage of the RAN network during its runtime, and dynamic placement of network functions.

UR4 — MRAN-CSP Res. Ctrl./VI Res. Ctrl. Moderator: Once an adequate RAN network configuration is defined by UR3, it is communicated via interface R5 to UR4, which translates it into specific cloud requirements and communicates the request to the VI resource controller via interface R6. If the request cannot be accommodated (e.g., due to sudden unavailability of resources), a round of negotiations takes place between the VI resource controller and UR4 on one side, and UR4 and UR5 on another side, until an agreement is made on the right RAN network configuration.

UR5 — RAN Network Configuration Enforcer: Once cloud resources are allocated for the RAN network configuration, UR5 is triggered via interface R7 to enforce the RAN network configuration on the allocated resources, that is, running RAN network functions (e.g., soft eNBs) on provided VMs and cloud resources as per the RAN network configuration determined by UR3. Instantiation of RAN network functions on VMs and their management are done via direct communication with the RAN-capable cloud using interface R9.

UR6 — Virtual RAN Network Function Manager: This is in charge of managing the RAN network functions to ensure that the virtual RAN network built on the cloud is appropriately working during runtime.

UR7 — Cross-Domain Interworking Agent: As part of UR6, UR7 is in charge of providing and enforcing interworking functions (e.g., necessary functions and procedures relevant to the Third Generation Partnership Project's, 3GGP's, S1 interface) between MRAN and MCN. This interworking is carried out through interface I1.

***MCN-CSP Resource Controller* —** The high-level structure of the MCN-CSP resource controller is similar to that of the MRAN-CSP resource controller. The MCN-CSP resource controller also consists of seven main units, referred to as UC1–7 (i.e., UC: unit for core); the role of each is described below.

Operating and managing the core network requires staff with strong technical expertise, which may be missing in these emerging markets. For these reasons and more, providing EPC as a service to "ease" the deployment of LTE and beyond in these regions would be of vital importance, for both societal and economic reasons.

UC1 — Service Provider Requirement Assessor: This is a module for assessing requirements from the different service providers using the MCN-CSP. Similar to UR1, this assessment can be based on either explicit feedback from the different service providers (using interface C1) or autonomic prediction of their requirements. Similar to interface R1, different protocols and implementation options can be envisioned for interface C1.

UC2 — VI Resource Assessor: This is a module for receiving and processing information (monitored by the VI cloud controller) about the resource availability in the underlying cloud. The monitoring information is exchanged using interface C2. It shall be noted that when the RAN-capable cloud and the cloud used by the MRAN-CSP and MCN-CSP, respectively, belong to the same cloud provider, UC2 and UR2 may be the same unit.

UC3 — Core Network Configurator: This develops knowledge of the resources of the cloud from UC2 using interface C3, understands requirements from service providers from UC1 using interface C4, and accordingly computes a mobile core network configuration that shall adequately meet the requirements of a service provider given the underlying cloud resources. UC3 employs different algorithms that decide where to place which network function (e.g., MME, PDN-GW, S-GW, and policy and charging rules function, PCRF) [16]. It may be called in the initial deployment of a mobile core network on the cloud for a service provider, for migration of network functions during runtime of the core network, extension/shrinkage of the core network during its runtime, and dynamic placement of network functions.

UC4 — MCN-CSP Res. Ctrl./VI Res. Ctrl. Moderator: Once an adequate mobile core network configuration is defined by UC3, it is communicated via interface C5 to UC4, which translates it into specific cloud requirements and communicates the request to the VI resource controller via interface C6. If the request cannot be accommodated (e.g., due to sudden unavailability of resources), a round of negotiations take place between the VI resource controller and UC4 on one side, and UC4 and UC5 on the other side, until an agreement is made on the right mobile core network configuration.

UC5 — Mobile Core Network Configuration Enforcer: Once cloud resources are allocated for the mobile core network configuration, UC5 is triggered via interface C7 to enforce the mobile core network configuration on the allocated resources, that is, running network functions (e.g., MME, PDN-GW, S-GW, etc) on provided VMs and cloud resources as per the mobile core network configuration determined by UC3. Instantiation of mobile core network functions on VMs and their management are done via direct communication with the cloud using interface C9.

UC6 — Virtual Mobile Core Network Function Manager: This is in charge of managing the mobile core network functions to ensure that the virtual mobile core network built on the cloud is working appropriately during runtime.

UC7 — Cross-Domain Interworking Agent: As part of UC6, UC7 is in charge of providing and enforcing interworking functions (e.g., necessary functions and procedures relevant to both S1 and SGi, two well standardized 3GPP interfaces) between MRAN and MCN on one hand, and MCN and PDN on the other hand. This interworking is carried out through interfaces I1 and I2.

PDN-CSE Resource Controller — The PDN-CSP resource controller also consists of seven units, referred to as UF1-7 (i.e., unit for supporting FMC: UF) with a structure different than that of the MRAN-CSP and MCN-CSP resource controllers. The role of each of these units is described below.

UF1 — Individual End User/Service Provider Requirement Assessor: Similar to UR1 and UC1, UF1 is a module for assessing requirements from the different service providers using the PDN-CSP, based on either explicit feedback from the different service providers (using interface F1) or autonomic prediction of their requirements. UF1 also communicates with individual end users (i.e., using the service providers) via interface F1 to assess when a user would be interested in the mobility of a service she receives (e.g., triggering FMC [19, 22]). This assessment can, for example, be based on change in the IP address of the user's UE.

UF2 — VI Resource Assessor: Similar to UR2 and UC2, UF2 is a module for receiving and processing information (monitored by the VI cloud controller) about the resource availability in the underlying cloud. The monitoring information is exchanged using interface F2. It shall be noted that when MRAN-CSP, MCN-CSP, and PDN-CSP use the same cloud, UR2, UC2, and UF2 may be the same unit.

UF3 — Data Anchor GW/PDN-Cloud Data Center Mapper: This may map onto the data center/GW mapping entity in [19]. It interfaces with the MCN-SCP resource controller to UC5 via interface F4. UF3 receives constant updates from UC5 about the locations of the different data anchor gateways of the different service providers using the MCN-SCP resource controller and maps them to adequate data centers. This mapping can, for example, be based on geographical proximity between data centers and the data anchor gateways.

UF4 — Follow Me Cloud Controller: This is triggered by UF1 through interface F3 to initiate service mobility for a particular individual end user or a group of users (e.g., users of a particular service provider), uses information on the mapping of data centers to data anchor gateways communicated from UF3 through interface F5, and decides whether a service needs to be moved for the user(s) and to where.

UF5 — Service Placement/Duplication Decision Maker: This runs numerous smart caching techniques and algorithms to decide where to place/replicate services based on triggers from UF4 through interface F7 and also based on available cloud resources as communicated by UF2 through F9.

UF6 — Service Placement Enforcer & PDN-CSP Res. Ctrl./VI Res. Ctrl. Moderator: This is in charge of enforcing decisions of UF5 by requesting resources from the VI resource controller through interface F6 and placing/distributing the services directly on the determined data centers using interface F10.

UF7 — MCN/PDN-Cloud Cross Domain Interworking Agent: It is a cross domain interworking agent providing and enforcing interworking functions between MCN and PDN (e.g., enforcing SGi interface).

Use Scenarios

Different use scenarios can be envisioned for the above described carrier cloud vision. One particular use case could be visited public land mobile network (PLMN) as a service. For example, a mobile operator, operating in country A, may want to make its services more attractive to its mobile subscribers who frequently travel overseas by lowering roaming cost. When a potential number of mobile subscribers roam to country B, the mobile operator may use the above described carrier cloud concept to create, on demand, a visited PLMN (vPLMN) for its roamers using resources of data centers available in country B. This shall enable the mobile operator to ensure low roaming cost to its mobile subscribers, rather than having them pay huge fees to roaming partners in country B. Two variants of this use case scenario should be supported. If the mobile operator's home PLMN (hPLMN) is a legacy network, it would create a complete on-demand vPLMN in the visited country, and enable roaming between its hPLMN and this vPLMN. If the mobile operator's hPLMN is also built on the cloud, an alternative would be to expand the hPLMN itself on demand into the visited country. This could be easier if the cloud provider of the hPLMN also operates in country B.

Another potential use case of the carrier cloud vision described herein could be EPC as a service (EASE), that is, providing core network functions and their management as a service. Indeed, a vendor or global mobile operator may provide the core network as a cloud service to RAN partners owning a RAN spectrum license and operating in specific regions. This use case could be of particular interest in emerging markets or developing countries to "ease" the deployment of LTE and beyond networks there. Table 1 provides a comparison of some telecom-relevant economics in developed countries and emerging markets. In the latter, RAN spectrum license is relatively cheap, and the real estate required for the deployment of RAN nodes is available. However, the core network nodes are quite expensive, and their purchase may not bring immediate return on investment given the low mobile service fares in emerging markets. Additionally, operating and managing the core network requires staff with strong technical expertise, which may be missing in these emerging markets. For these reasons and more, providing EASE to "ease" the deployment of LTE and beyond in these regions would be of vital importance, for both societal and economic reasons.

	Developed countries	Emerging markets
RAN	Expensive due to • High RAN license fees • Hgh/scarce real estate	Cheap due to • Lower RAN license fees • Cheap and available real estate
Core network	Relatively affordable	Not affordable
Technical expertise	Available	Shortage

Table 1. A comparison between some telecom-relevant economics in developed countries and emerging markets.

A mobile operator may desire to create its own content distribution network (i.e., mobile CDN) to cache popular videos for its users. Using the above described carrier cloud vision, the mobile operator may request the deployment of relevant network functions — video-on-demand (VoD) cache functions and their placement at strategic data centers — taking into account the geographical proximity of users to these data centers and also the amount of video traffic (i.e., hit ratio) associated with them. This defines a new use case of the above described carrier cloud concept, mobile CDN as a service. Another use case could be "offload network as a service." Effectively, a mobile operator may desire to selectively offload some IP services (e.g., YouTube and Facebook traffic) from its "legacy" core network [6, 7, 18]. For this purpose, the mobile operator may create on demand a "cheap" cloud-based mobile core network in charge of handling those offloaded IP services.

In the area of MTC [4], operators currently charge an MTC provider on the order of €10/mo and per MTC device. Depending on the MTC application/service, MTC providers (e.g., a gas/electricity/water utility provider) may need to deploy a potentially high number of MTC devices (e.g., millions of devices). With the current pricing model, such an MTC provider may have to pay on the order of €10 million/mo. This price can be considered significantly high given the fact that such MTC devices connect to the mobile network only once a month and for a short time (i.e., only for a few seconds at the end of each month for sending measurements of the gas/electricity/water consumption). An important use case of the above described carrier cloud concept could then be the on-demand creation of a mobile network on the cloud to be functional from the specific time when the measurements have to be communicated to the MTC server(s), and only for the duration in which all measurements need to be reported. Once the measurements are reported, the mobile network created on the cloud may be deleted, and the allocated resources may be released. Allocating resources from the cloud and running mobile network functions there for a very short time shall certainly be less costly than paying multiple traditional MTC device subscription fees (for the multitude of MTC devices) to a traditional mobile operator. In addition to this significant cost saving, such an MTC EPC as a service concept shall allow important network configuration flexibility and ensure short time to market of new MTC applications and services.

All in all, the above mentioned use cases are just a few examples. Using the above described carrier cloud concept, users, enterprises, as well as mere individual users, will be given the possibility and flexibility to deploy, run, and manage their own mobile networks, and use its mobile connectivity service to launch any IP service requiring mobility support. Mobile networks will then be built to meet the services' requirements, elastically accommodating their changing demands, and ultimately enabling the concept of service-driven mobile networking, whereby mobile networks are designed and built from the viewpoint of target services and independent from the underlying infrastructure.

Resource management over a federated cloud, supported by cross-data-center monitoring, and VM migration, particularly its orchestration logic, are also important research topics that deserve further studies and investigations from the wide community of interested researchers and practitioners.

CONCLUSION

In spite of the fact that most existing mobile services are still circuit-switch-based, the penetration of smartphones into the mobile market has brought a tremendous increase in mobile Internet traffic. Along with vast deployment of LTE and their packet-switched services, mobile Internet traffic is foreseen to increase further by a large factor in the next few years. To accommodate this tsunami of mobile IP traffic, mobile operators need to invest in their infrastructure by purchasing more core network nodes. No matter how powerful these nodes are, they are limited in the maximum number of packets or messages they can handle per second. For instance, depending on the vendor, an MME node can typically process/handle a maximum of 1000 mobility-relevant signaling messages per second from user equipment residing in its pool area. In a pool area in which 1001 mobility-relevant signaling messages are transmitted on average, the operator needs to deploy two MME nodes. This means that the operator needs to deploy an additional MME node just to accommodate one additional mobility signaling message exceeding the capacity of a single MME. This could be a costly investment for the operator given the fact that core network nodes tend to be expensive. The cost becomes even more significant given the trend toward flat data rate models and the modest ARPU.

As a remedy to these economic challenges faced by mobile operators and to achieve mobile core network decentralization, this article introduces the concept of the carrier cloud, its high-level architecture, and the mechanisms to achieve it. To achieve the carrier cloud, there is also a need for network function virtualization whereby the software components of mobile core network nodes are decoupled from the hardware. For efficient end-to-end mobile connectivity, the Follow Me Cloud concept is also introduced for better interworking between the mobile core network and the packet data network, provided on the cloud.

With the carrier cloud concept, a mobile operator does not necessarily need to purchase completely new nodes to accommodate changes in its mobile traffic. Instead, the mobile operator needs to request resources only in terms of memory storage and CPU to run the adequate carrier network node functionality as per the needs of its customers, on demand, and in an elastic way. Furthermore, the mobile operator pays the cloud provider only per usage of its resources, which are ultimately covered by the users. This enables the mobile operator to have a certain level of control on its investment cost and also ensure a certain level of profit/return on investment. This is not to mention the rapidity of deploying the mobile network on the cloud in an elastic way, in comparison to the traditional deployment of mobile network nodes that usually takes weeks, if not months (e.g., time for network dimensioning and planning, organization, delegating engineers, delivering and setting up the nodes at the right spots). Furthermore, the presented carrier cloud concept is also expected to generate new revenue streams to mobile operators, which are today often limited to connectivity services.

Admittedly, the carrier cloud cannot be realized without tackling important challenges. Building mobile systems, traditionally known for their high reliability, availability, and security, on less reliable and less secure systems, such as the cloud, is an important challenge. The adaptability and resiliency of virtualized network functions in a carrier cloud to its dynamicity, frequent redistribution, and scaling is another important hurdle. Resource management over a federated cloud, supported by cross-data-center monitoring, and VM migration, particularly its orchestration logic, are also important research topics that deserve further studies and investigations from the wide community of interested researchers and practitioners.

REFERENCES

[1] 3GPP TS 23.401, "General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access."

[2] T. Taleb and A. Ksentini, "Impact of Emerging Social Media Applications on Mobile Networks," *Proc. IEEE ICC 2013*, Budapest, Hungary, June 2013.

[3] T. Taleb and A. Ksentini, "QoS/QoE Predictions-Based Admission Control for Femto Communications," *Proc. IEEE ICC 2012*, Ottawa, Canada, June 2012.

[4] T. Taleb and A. Kunz, "Machine Type Communications in 3GPP Networks: Potential, Challenges, and Solutions," *IEEE Commun. Mag.*, vol. 50, no. 3, Mar. 2012.

[5] T. Taleb, K. Samdanis, and F. Filali, "Towards Supporting Highly Mobile Nodes in Decentralized Mobile Operator Networks," *Proc. IEEE ICC 2012*, Ottawa, Canada, June 2012.

[6] K. Samdanis, T. Taleb, and S. Schmid, "Traffic Offload Enhancements for eUTRAN," *IEEE Commun. Surveys & Tutorials*, vol. 11, no. 3, Aug. 2012, pp. 884–96.

[7] T. Taleb, Y. Hadjadj-Aoul, and K. Samdanis, "Efficient Solutions for Data Traffic Management in 3GPP Networks," to appear, *IEEE Sys. J.*

[8] R. Miller, "Solar-Powered Micro Data Center at Rutgers," *Data Center Knowledge*, May 2012; http://www.datacenterknowledge.com/archives/2012/05/31/solar-powered-micro-data-center-at-rutgers/.

[9] R. Miller, "AOL Gets Small with Outdoor Micro Data Center," *Data Center Knowledge*, July 2012; http://www.datacenterknowledge.com/archives/2012/07/06/aol-micro-data-centers/.

[10] Authored by network operators, "Network Functions Virtualization: An Introduction, Benefits, Enablers, Challenges, & Call for Action," Oct. 2012

[11] J. Batalle *et al.*, "On the Implementation of NFV over an OpenFlow Infrastructure: Routing Function Virtualization," *Proc. 2013 IEEE SDN for Future Networks and Services*, Trento, Italy, Nov. 2013.

[12] K. Pentikousis, Y. Wang, and W. Hu, "MobileFlow: Toward Software-Defined Mobile Networks," *IEEE Commun. Mag.*, vol. 51, no. 7, July 2013, pp. 44–53

[13] A. Basta *et al.*, "A Virtual SDN-Enabled LTE EPC Architecture: A Case Study for S-/P-Gateways Functions," *Proc. 2013 IEEE SDN for Future Networks and Services*, Trento, Italy, Nov. 2013.

[14] J. Kempf *et al.*, "Moving the Mobile Evolved Packet Core to the Cloud," *Proc. WiMob*, Barcelona, Spain, 2012.

[15] R. Riggio, T. Rasheed, and F. Granelli, "EmPOWER: A Testbed for Network Function Virtualization Research and Experimentation," *Proc. 2013 IEEE SDN for Future Networks and Services*, Trento, Italy, Nov. 2013.

[16] T. Taleb and A. Ksentini, "Gateway Relocation Avoidance-Aware Network Function Placement in Carrier Cloud," *Proc. ACM MSWIM 2013*, Barcelona, Spain, Nov. 2013.

[17] T. Taleb and A. Ksentini, "On Efficient Data Anchor Point Selection in Distributed Mobile Networks," *Proc. IEEE ICC 2013*, Budapest, Hungary, June 2013.

[18] T. Taleb, Y. Hadjadj-Aoul, and S. Schmid, "Geographical Location and Load Based Gateway Selection for Optimal Traffic Offload in Mobile Networks," *Proc. IFIP Networking*, Valencia, Spain, May 2011.

[19] T. Taleb and A. Ksentini, "Follow Me Cloud: Interworking Federated Clouds and Distributed Mobile Networks," *IEEE Network*, vol. 27, no. 5, Sept. 2013.

[20] "Cisco Locator/ID Separation Protocol and Overlay Transport Virtualization Data Center Infrastructure Solutions for Distributed Data Centers," Cisco White Paper, 2011.

[21] P. Raad *et al*., "Achieving Sub-Second Downtimes in Internet Virtual Machine Live Migrations with Lisp," *Proc. IEEE/IFIP Int'l. Symposium on Integrated Network Management*, Ghent, Belgium, May 2013.

[22] T. Taleb, P. Hasselmeyer, and F. Mir, "Follow-Me Cloud: An OpenFlow-Based Implementation," *Proc. IEEE GreenCom '13*, Beijing, China, Aug. 2013.

[23] T. Taleb and A. Ksentini, "An Analytical Model for Follow Me Cloud," *Proc. IEEE GLOBECOM 2013*, Atlanta, GA, Dec. 2013.

Biography

TARIK TALEB [SM] (talebtarik@ieee.org) is an IEEE Communications Society (ComSoc) Distinguished Lecturer. He is currently working as a senior researcher and 3GPP standardization expert at NEC Europe Ltd, Heidelberg, Germany. Prior to his current position, he worked as a faculty member at Tohoku University, Japan. He received his B.E. degree in information engineering with distinction, and his M.Sc. and Ph.D. degrees in information sciences from Tohoku University in 2001, 2003, and 2005, respectively. His research interests lie in the field of architectural enhancements to mobile core networks, mobile cloud networking, mobile multimedia streaming, intervehicular communications, and social media networking. He has also been directly engaged in the development and standardization of the Evolved Packet System as a member of 3GPP's System Architecture working group. He serves as Vice-Chair of the Wireless Communications Technical Committee. He is the recipient of many awards, including the 2009 IEEE ComSoc Asia-Pacific Best Young Researcher award (June 2009). Some of his research work has also received best paper awards at prestigious conferences.