Benefits and Challenges of Virtualization in 5G Radio Access Networks

The authors focus on the benefits, challenges, and limitations that accompany virtualization in 5G radio access networks (RANs). Within the context of virtualized RAN, they consider its implementation requirements and analyze its cost. They also outline the impact on standardization, which will continue to involve 3GPP but will engage new players whose inclusion in the discussion encourages novel implementation concepts.

Peter Rost, Ignacio Berberana, Andreas Maeder, Henning Paul, Vinay Suryaprakash, Matthew Valenti, Dirk Wübben, Armin Dekorsy, and Gerhard Fettweis

Communications Standards

Abstract

Future 5G deployments will embrace a multitude of novel technologies that will significantly change the air interface, system architecture, and service delivery platforms. However, compared to previous migrations to next-generation technologies, this time the implementation of mobile networks will receive particular attention. The virtualization of network functionality, the application of open, standardized, and inter-operable software, as well as the use of commodity hardware, will transform mobile-network technology. In this article we focus on the benefits, challenges, and limitations that accompany virtualization in 5G radio access networks (RANs). Within the context of virtualized RAN, we consider its implementation requirements and analyze its cost. We also outline the impact on standardization, which will continue to involve 3GPP but will engage new players whose inclusion in the discussion encourages novel implementation concepts.

Introduction

Cloud computing has dramatically transformed the information technology (IT) sector by introducing new ways to store and process data, create and offer services, and operate complex systems. Recognizing this power, mobile network operators are beginning to leverage cloud-computing technologies by migrating mobile network functionality to the cloud. At first, operator services and functions in the core network were the focus of the research and standards communities [1], e.g. in the European Telecommunication Standards Institute (ETSI) Network Functions Virtualization (NFV) Industry Specification Group (ISG) [2]. Recent attention has shifted to meeting the baseband-processing requirements of the radio access network (RAN) on high-volume IT hardware [3, 4].

Concurrently, an increasing number of mobile terminals and an increased demand for data motivate massive network densification through the use of small cells. In a macro-cell network, each cell serves a large number of users, which enables modeling aggregated traffic as being homogeneous even if the users have different traffic and mobility profiles. In contrast, each cell in a small-cell network serves fewer users, and hence the traffic profile observed is less homogeneous; i.e. there are areas with significant peak traffic (e.g. metro stations) and areas with no (or low) traffic (e.g. a business district during weekends). Additionally, finer spatial sampling of traffic by small cells implies stronger traffic variation per cell. For instance, [5] shows that macro-cell utilization is typically around 20 percent to 40 percent. However, since each base station (BS) must be equipped with sufficient computing resources to handle its peak load, resources are over-provisioned by a factor of 5 to 10, which is both expensive and wasteful. Centralized RAN and resource virtualization avoids over-provisioning by assigning resources intelligently and elastically based on the actual need.

Virtualization in the Context of RAN

Forms of RAN Virtualization

Virtualization can be applied to different aspects of the RAN, through spectrum virtualization, hardware sharing, virtualization of multiple radio access technologies (RATs), and virtualization of computing resources. Spectrum virtualization allows the available spectrum to be utilized more efficiently by permitting multiple network operators to share the same spectrum. Hardware and network sharing is of particular relevance for small cells in order to avoid massive over-provisioning. Virtualization of multiple RATs allows simplified management of different RATs, each dedicated to different services and offering a different quality of service (QoS). Virtualization of computing resources is a new option that builds upon the idea of co-locating the processing resources of multiple BSs at a central processing center. While early implementations provided each physical BS with its own dedicated computing resources, which resulted in an over-provisioning of computing resources, more advanced implementations permit a dynamic reassignment of processing resources to BSs. This article focuses on the potentials and challenges of moving the processing required for a mobile network to a centralized computing cloud that houses a virtualized computing infrastructure based on commodity hardware. In the following, we refer to this system as *cloud-RAN*.

Cloud-RAN as an Enabler of RAN Virtualization

A fully commoditized implementation permits complete programmability and flexibility, and it facilitates realizing the gains of a cloud-RAN implementation to the fullest extent. However, the question of whether (or not) the computational power of commoditized hardware is sufficient remains open. In order to fulfill real-time guarantees, computationally intensive parts may be executed on dedicated support hardware, e.g. co-processors similar to a graphical processing unit (GPU). To avoid a hardware lock-in, these

Peter Rost and Andreas Maeder are with Nokia Networks.

Ignacio Berberana is with Telefonica I+D.

Henning Paul, Dirk Wübben, and Armin Dekorsy are with the University of Bremen.

Vinay Suryaprakash and Gerhard Fettweis are with Technische Universität Dresden.

Matthew Valenti is with West Virginia University.

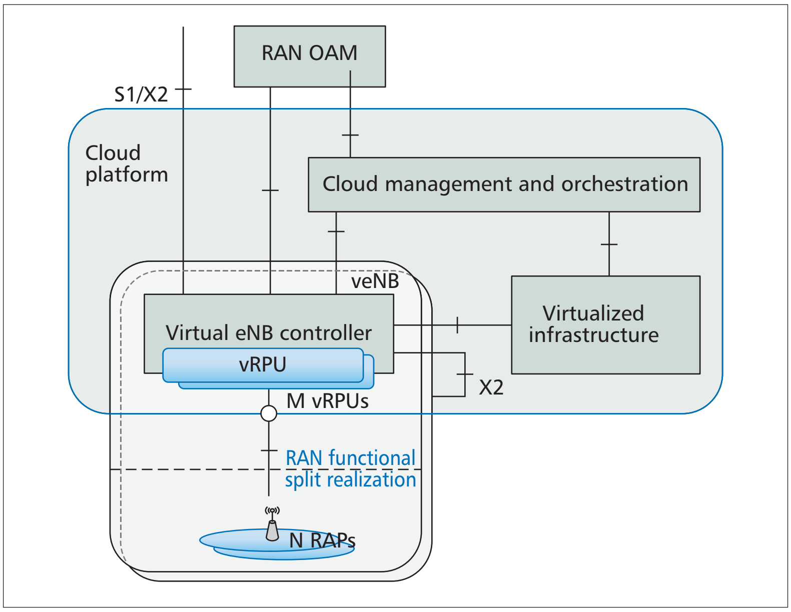
RAN OAM
S1/X2
Cloud platform
Cloud management and orchestration
veNB
Virtual eNB controller
vRPU
Virtualized infrastructure
X2
M vRPUs
RAN functional split realization
N RAPs

Figure 1. Cloud-RAN architecture implementation example.

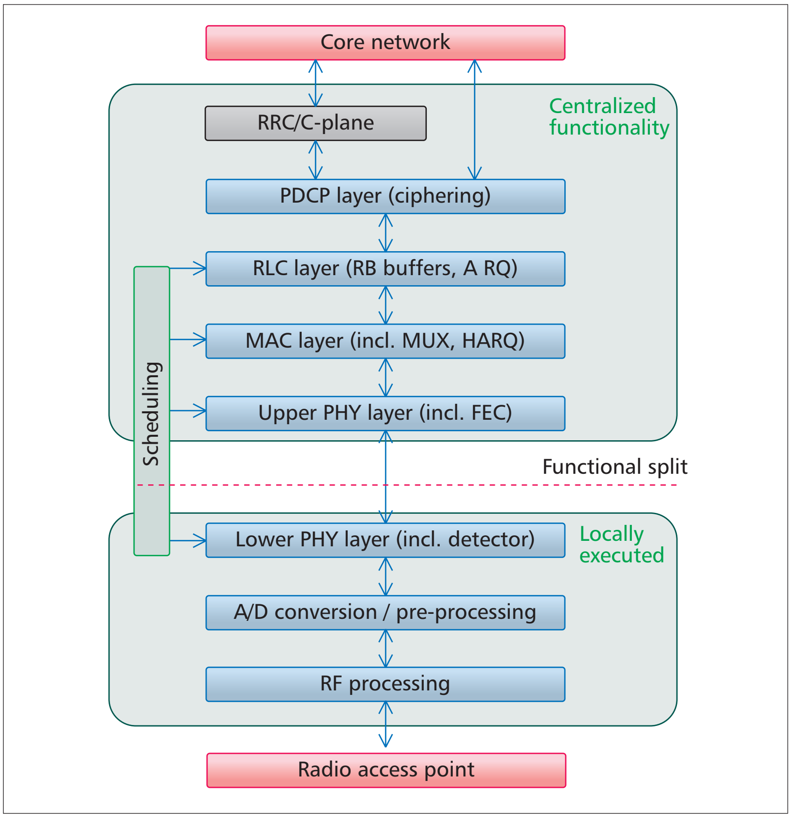
Core network
RRC/C-plane
Centralized functionality
PDCP layer (ciphering)
RLC layer (RB buffers, A RQ)
MAC layer (incl. MUX, HARQ)
Upper PHY layer (incl. FEC)
Scheduling
Functional split
Lower PHY layer (incl. detector)
Locally executed
A/D conversion / pre-processing
RF processing
Radio access point

Figure 2. RAN functional split.

co-processors could be made accessible through open interfaces (similar to OpenGL). Furthermore, the size and location of the computing centers are important design choices for a cloud-RAN system. Larger computing pools improve efficiency by reducing the likelihood of insufficient computing resources, and moving the data and its processing to remote centers may leverage potentially reduced operating costs (for instance, cheaper electricity, land, or labor). However, moving the data to a remote location will add to the communication latencies and costs. Please note that this article uses 3GPP LTE RAN functionality to explain RAN virtualization, although RAN virtualization itself will be an essential part of future 5G mobile networks.

The virtualization of computing resources and its application to RANs will require an interaction among standardization bodies concerned with RAN such as 3GPP (for which the implementation platform has so far been out of scope) and those focusing on virtualization aspects such as the ETSI NFV and Mobile Edge Computing (MEC) ISGs. Ideally, the standardization of RAN protocols takes into account characteristics stemming from virtualization and partial centralization, allowing “cloud-native” RAN implementations. Alternatively, the computing platform and its interaction with BSs may also be adapted to the RAN protocol constraints, as in the case of legacy system implementations.

Figure 1 shows an example cloud-RAN architecture that applies the ETSI NFV architectural principle to an E-UTRAN based system. The enhanced NodeB (eNodeB) as a logical network entity defined in 3GPP is implemented both in the cloud platform and at physical radio access points (RAP). The “cloudified” part of the eNodeB consists of a number of virtual RAN processing units (vRPUs), responsible for executing the RAN protocol stack, and a virtual eNodeB controller that terminates 3GPP interfaces toward the core network and other eNodeBs (virtual or non-virtual). In the context of the ETSI NFV framework, the virtualized eNodeB can be implemented as a virtual network function (VNF), which is instantiated on a virtualized infrastructure. Scaling, lifecycle management, and performance monitoring are handled by corresponding management and orchestration functions that take information from the RAN OAM (operations, administration and management/maintenance) system into account.

RAN Functional Decomposition

One of the central issues in a cloud-RAN environment is determining which functionality is executed centrally at the data center and which remains local to the RAP. Based on the results in [2] and [3], this article considers a cloud-RAN that applies the RAN functional split that is illustrated in Fig. 2. We assume that all RAN functionality starting with forward error correction (FEC) and including Layer 2 and Layer 3 processing is centralized and processed on high-volume commodity IT hardware. All functionality below FEC is executed locally at the RAP. Note that other degrees of centralization are possible but are not elaborated upon here (see [2, 3] for further details). The functional split highlighted by this article centralizes a majority of the computations, thereby exploiting most of the centralization gains, yet it relaxes the latency and throughput requirements on the backhaul as compared to a fully centralized solution that requires substantial “fronthaul” connectivity. Hence, the solution can be implemented using today’s backhaul and switching technologies.

Implementation Aspects of Cloud-RAN Systems

Cloud-RAN requires novel technologies that are provisioned along three dimensions:

- Radio access equipment: power-efficient and cost-efficient multi-RAT BSs.
- Backhaul: flexible and economical connectivity of BSs and centralized data-processing resources.
- Data processing: economical, elastic, and easily programmable centralized processing resources.

The virtualization and centralization of the RAN requires a platform that lies at the intersection of real-time architectures for processing communication signals and large-scale information processing systems. This creates dependencies between the data-processing capabilities of the computing infrastructure and the achievable communication rates of the RAN. Hence, the design, optimization, and analysis of such a system require a new conceptual framework that links the theories of data processing and communications. In this section we elaborate upon the implementation aspects of cloud-RAN and discuss challenges (and opportunities) that arise.

Real-Time Requirements of Cloud-RAN

Cloud-RAN implementations must take into account the stringent real-time requirements of the RAN. For instance, hybrid automatic repeat-request (HARQ) in LTE requires that a positive or negative acknowledgement be sent 3 ms after receiving a transport block. Failure to do so induces an unnecessary HARQ transmission, thereby lowering the throughput. In the downlink, link adaptation and radio frame generation are the main challenges. Link adaptation becomes suboptimal as the channel information becomes outdated, and radio frame generation must be synchronized between the cloud platform and the BS. Furthermore, the QoS profile enforces end-to-end latency guarantees, which require processing to be completed within a stipulated amount of time.

The actual real-time constraints that need to be fulfilled depend strongly on the functionality that is performed centrally, the dependencies between individual functional components of the RAN, and the ability to predict the processing requirements. Depending on the degree of centralization, some of the real-time constraints may be relaxed, e.g. performing decoding locally at the BS gives more time to meet the HARQ timing constraint. However, the dependencies between individual RAN functions play an important role, e.g. link adaptation determines the actual data rates on the air interface but is susceptible to changes in channel quality. If link adaptation is performed locally, then it is difficult to perform scheduling and packet segmentation at the central processor. Finally, the predictability of processing requirements is important because software jitter may violate real-time constraints. For instance, turbo-decoding requires about 80 percent of the uplink processing [11]. However, depending on the number of iterations and the actual number of information bits processed, the required complexity and decoding time may vary significantly.

The actual real-time constraints that need to be fulfilled depend strongly on the functionality that is performed centrally, the dependencies between individual functional components of the RAN, and the ability to predict the processing requirements.

Implementation Constraints of Cloud-RAN

Software implementation of RAN functionality requires a new way to design and operate the RAN. Until now, RAN functionality has been executed on dedicated hardware such as digital signal processors (DSPs) or application specific integrated circuits (ASICs). Dedicated hardware is precisely dimensioned and provides the required resources to cope with peak-traffic demands; it is highly reliable and has high performance, but does not permit sharing or virtualization of resources. In contrast, software implementation on commodity hardware may be more flexible and allow for resource sharing and virtualization. However, it is usually less reliable and has lower performance. Therefore, such implementations need to be “cloud-native” and must be designed for resilience. This cannot be achieved by merely porting existing implementations, but rather requires more advanced concepts.

Commodity hardware may be implemented by general purpose processors (GPPs) or a mix of GPPs for upper-layer processing and complementary network processors for lower-layer processing, similar to GPUs in current computer architectures. The network processors may be addressed through open interfaces (similar to OpenGL) to allow flexibility. Additionally, the processing may be performed in virtual machines or in more lightweight environments such as containers [6].

Cloud-RAN will pose new challenges to data-center architectures since it may require dedicated platforms rather than the existing platforms that have been optimized for Internet services. However, they will still be considered “commodity” due to the pervasiveness of mobile network technology. In particular, the distribution and execution of RAN processing jobs in data centers requires high-performance software defined networking (SDN) architectures that route RAN data and address processing elements within data centers efficiently. Similarly, the real-time requirements in a RAN may not allow simple migration of virtual machines (or containers) but require new mechanisms that facilitate fast transfer of processing states or RAN protocol states. The efficiency with which processing elements (containers or virtual machines) are assigned to data packets has a major impact on the elasticity of the system.

The requirements on the data center will also depend on the manner in which the processing is implemented. For instance, processing may be performed on a per-user-terminal basis, a per-BS basis, or a per-cluster basis. The first option provides higher scheduling granularity; the second option may simplify the process of merging data originating from different user terminals (e.g. for scheduling); the third option may simplify the joint processing of data across multiple BSs. Furthermore, the parallelization of processing could be done with a very low granularity (on a per packet basis) or with higher granularity (on a per-BS basis). The need for synchronization objects (semaphores) also increases with an increase in granularity, and this limits the processing performance significantly. In contrast, processing each packet on a separate processor (or core) allows for decoupling processes, and therefore avoids the need for synchronization objects.

In a cloud-RAN system, the radio and data-processing resources should be managed jointly; i.e., radio resource allocation must adapt not only to the prevailing channel conditions and required quality of service, but must take into account the demand for computational resources imposed by the radio allocation.

Joint RAN/Cloud Resource Management

In a cloud-RAN system, the radio and data-processing resources should be managed jointly, i.e. radio resource allocation must adapt not only to the prevailing channel conditions and required quality of service, but must take into account the demand for computational resources imposed by the radio allocation. This is a predictive task, i.e. the system has to estimate the required computational complexity, estimate the available computational resources, and then adapt the RAN resource allocation accordingly.

One possibility for carrying out this joint optimization is to account for the data-processing load during link adaptation, i.e. the resource scheduler could incorporate a weighted metric that penalizes choices that lead to high computational demands. Furthermore, as the number of users served by a BS increases, the expected traffic and processing requirements will also increase. This may require a re-assignment of processing resources to virtual machines within the data-processing center and should be anticipated by the scheduler. Additionally, the scheduler must be able to operate at the computational capacity, i.e. the maximum system throughput using a given amount of computational resources. This requirement is particularly important during peak-traffic hours when many users connect to the mobile network and the computational load approaches the system limit.

The previous examples described operational tasks. However, there are also dimensioning and positioning challenges. For instance, [7] provides a framework for estimating the amount of computational resources required for an expected number of users. The concept of computational outage, which is the likelihood that the available computational resources are insufficient to meet the instantaneous computational load, is introduced. Using this framework, the required computational resources can be predicted, and computationally aware schedulers that maximize the system utilization and prevent computational outage can be designed.

Data-Processing Complexity of RAN Protocols

In a cloud-RAN system, the data-processing requirements depend on many different factors. For example, if the transmission rate increases, more information bits need to be processed, which in turn linearly increases the computational load. Additionally, if a communication link operates close to its Shannon capacity, even more receiver processing is required, which can be attributed to the need for additional turbo decoder iterations. As a result, the processing load increases super-linearly as the system operates increasingly close to capacity, and the load depends on both the instantaneous channel conditions and the scheduling policy, which determines how close to capacity the system operates.

Therefore, in a manner similar to exploiting channel diversity in mobile networks (e.g. multi-user diversity in scheduling or spatial diversity in multiple antenna systems), *computational* diversity can also be exploited. Computational diversity exploits the large fluctuations in the data-processing load imposed by multiple users with diverse channel conditions. Hence, if multiple users are served by a cloud-RAN instance, their diverse computational requirements may be used to improve the resource utilization since the computational assets need to be provisioned according to the expected cumulative load of the users rather than the peak load of any given user. Furthermore, by dynamically adapting the modulation and coding scheme through the use of appropriate *computationally aware* link-adaptation algorithms, the data-processing load can be controlled.

From a user's perspective, there is no difference between a channel outage and a computational outage: in either case, the communications fail and another attempt must be made to transmit the packets. The model introduced in [7] accurately predicts the data-processing resources required to perform uplink decoding in a multi-cell scenario for a given threshold on computational outage probability. Using the empirically determined computational load discussed in [11] and assuming a 10 MHz LTE channel and that each turbo-decoder iteration requires 1000 FLOPS per data bit, we can estimate the overall required data-processing capabilities for a reference setup involving server blades equipped with four Intel Xeon 4870 (10-core processor) and 128GB RAM.

Based on these assumptions and the framework introduced in [7], Fig. 3 shows the computational resources required for LTE in a data center. We compare two cases of centralized computing: in the cloud-RAN case with virtualization, processing resources may be flexibly re-assigned to BSs, while in the second case without virtualization, each BS is serviced by its own dedicated computational resources (as is the case in fully centralized RAN). For both cases, cells are assumed to be fully loaded and the computational outage probability is set to 10 percent. We quantify the computational requirement by the number of servers using the aforementioned architecture. When resources are shared, we see a reduction of approximately 50 percent in the data-processing resources required.

Cost Analysis

A major issue in cloud-RAN implementation is its impact on the cost of mobile networks (and the capital expenditure (CAPEX) in particular). RAN virtualization and centralization over a non-ideal backhaul may allow cost-efficient implementations. The backhaul deployment and the ability to use existing infrastructure as well as non-ideal backhaul technologies play a critical role in the economic analysis of cloud-RAN. This section presents a cost-analysis for the RAN functional split illustrated in Fig. 2.

Cost-Components in Cloud-RAN

An evaluation of the CAPEX for a cloud-RAN system must consider costs stemming from different network components. In [8] four different network layers[1] as illustrated in Fig. 4 are distinguished: users, BSs, backhaul nodes, and data centers. The lowest layer represents the users and assumes a particular average traffic demand per user. Both micro and macro BSs are overlaid on the same layer. Backhaul nodes consist of aggregation points that are then connected to data centers wherein centralized processing is performed. This model captures the most important cost components and facilitates analysis of the salient trends in a cloud-RAN deployment.

BSs in a cloud-RAN environment process may perform only part of the RAN protocol stack at the local site. Hence, their size and costs might depend upon the amount of processing performed locally. Furthermore, the difference in cost between macro and micro BSs can vary significantly because the latter use lower transmit power and fewer antennas, thereby reducing the cost per access point significantly. In contrast, increased centralization also requires adequate backhaul technologies that cater to the throughput and latency requirements. In fully centralized systems, high-performance optical fiber is required. Since it is very expensive, its cost may even outweigh the RAN cost reduction described above. The functional split considered in this article (Fig. 2) does not require specific backhaul technologies and can also be applied to non-ideal backhaul (with latencies above 1 msec). Therefore, the backhaul costs may be significantly lower compared to a fully centralized RAN. Finally, the deployment of additional data centers will be necessary. However, since each data center hosts only a few servers (as seen in the number of servers required in Fig. 3), the additional data-processing hardware required may be deployed at preexisting points in order to simplify site acquisition and reduce costs.

Example Cost Analysis

In [8] a generic framework and an expression for the CAPEX of the entire network have been derived, which can be applied to three different scenarios by substituting the appropriate component cost values. The first scenario is cloud-RAN with virtualization and using a mix of optical and wireless backhaul technologies (each 50 percent). The second scenario is a fully centralized RAN, which refers to a functional split above A/D conversion in Fig. 2, without virtualization, and requiring optical fiber connectivity (as prevalent today). The third scenario is distributed RAN (DRAN), which refers to a conventional implementation where all RAN processing is performed locally at the BSs. Table 1 shows an example budget for a cloud-RAN network that uses the functional split illustrated in Fig. 2. It compares the costs for DRAN and cloud-RAN. In our example, we assume 170 active users per km^2, an average traffic demand per user of 10 Mb/s, and a mix of 50 percent microwave and 50 percent optical fiber technology for the backhaul.

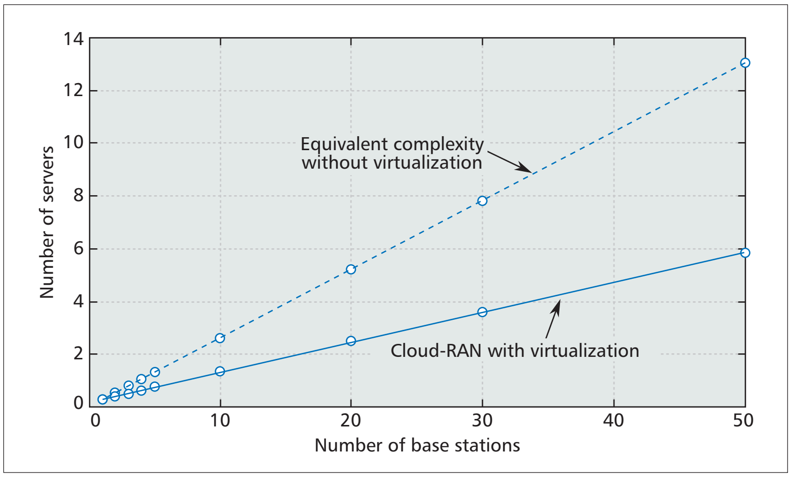

Figure 3. The number of IT server blades required depending upon the number of centralized (fully loaded) BSs.

Type of cost	DRAN	Cloud-RAN
Macro base station	$50k	$25k
Micro base station	$20k	$10k
Microwave BH	$50k per link plus $5k per kilometer	
Optical fiber BH	$5k per link plus $100k per kilometer	
Data center		$40k
Server blades		$20k each (Fig. 3)

Table 1. Exemplary budget for Cloud-RAN analysis (more details are given in [7]).

Figure 5 shows the resulting CAPEX for the three cases of DRAN, cloud-RAN, and fully centralized RAN. In all three cases, the expected area throughput is used as a basis for normalization and results are plotted over different data center densities. It is important to note that one data center may consist of only a few server racks at an existing point of presence within the mobile network. This reduces the operational expenditure (OPEX) because no additional site rental is necessary. Furthermore, small data centers promote greater failure resilience and they reduce the traffic within the metropolitan transport network. Therefore, considering Fig. 5, a density of one or two data centers per square kilometer appears realistic in a very dense urban small-cell deployment. If we further increase the density of small-cells, e.g. due to higher data rate demands and user density, then the cost effectiveness of cloud-RAN would increase even further as the exploited centralization gains in cloud-RAN also increase (similar to the over-provisioning of distributed RAN) and the cost-reduction per BS becomes more dominant.

The results show that cloud-RAN based on the applied RAN functional split can be more cost effective than a DRAN implementation. However, the actual benefit may depend on the scenario, parameterization, and actual traffic

[1] Note that the term "layer" here does not refer to the layers in the Open Systems Interconnection (OSI) model, which standardizes the internal functions of a communication system, but instead refers to each network component.

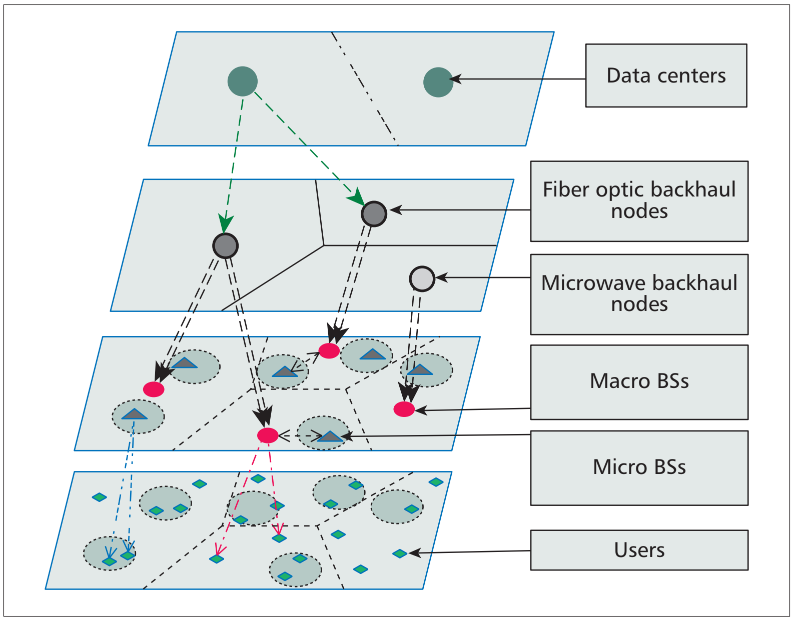

Figure 4. Model of a heterogeneous network with multiple network components.

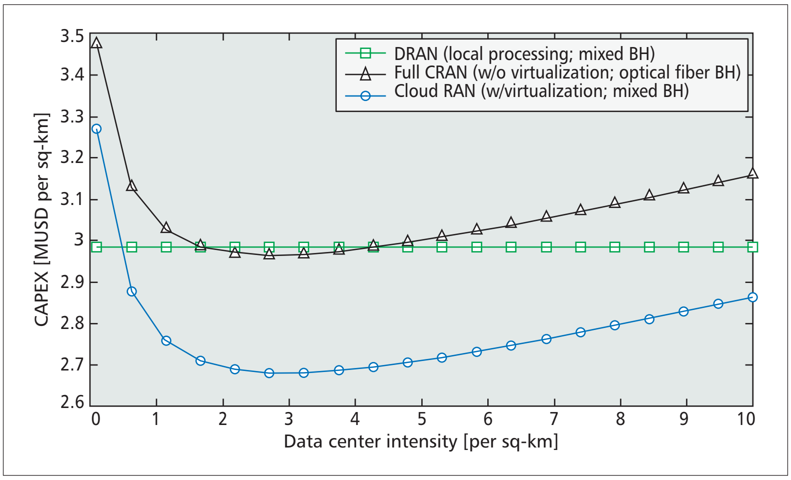

Figure 5. Fost efficiency analysis.

demand. Additionally, the architecture presented here holds the potential for reduced OPEX due to lower maintenance costs on site as well as easier management through standard IT management mechanisms.

Standardization Impact

Several standards development organizations (SDOs) have recognized virtualized cloud-RAN as one of the key technologies to meet the requirements of 5G networks. The mobile communication industry (including operators, vendors, and chipmakers) collaborates in various industry fora and projects on drafting 5G requirements. The Next Generation Mobile Networks (NGMN) Alliance created its 5G Initiative, focusing on an operator's view of 5G requirements. In Europe, the ETSI and several 5G-related projects funded by the European Commission, such as iJOIN[2] and METIS[3], work toward a common view on 5G. For the North American market, the 4G Americas industry forum published recommendations on 5G requirements and solutions [12]. Similar publications are available from Asian 5G-related activities, e.g. from the 5G Forum in Korea, ARIB in Japan, and the IMT-2020 Promotion Group as well as the Chinese Ministry of Science and Technology project 863-5G. Finally, the International Telecommunication Unit promotes its "IMT-2020 and beyond" program in the ITU-R working group 5D.

While all these activities are pre-standardization efforts, they frequently mention coordinated transmission and flexible networks/services as key enabling technologies, and cost and resource efficiency as key requirements for future 5G solutions, all of which strongly point toward virtualized cloud-RAN as a solution.

Standardization of virtualized cloud-RAN requires activities in both networking and computing frameworks, such as SDN and NFV, as well as enhancements to the mobile network functionalities and architecture in order to take full advantage of the virtualization approach.

Network and computing frameworks are addressed by several SDOs such as the Open Networking Foundation (ONF, [15]) for SDN based on the OpenFlow protocol, ETSI NFV[4] and ETSI MEC.[5] Virtualized cloud-RAN as a use case implies new and stringent requirements (e.g. on latency) for these frameworks. The ONF Wireless & Mobile project, which aims to collect use cases and determine architectural as well as protocol requirements for extending ONF-based technologies to wireless and mobile domains, is a first step toward the identification of such requirements in the transport-network area.

For computing frameworks, ETSI NFV aims to evolve quasi-standard IT virtualization technology to consolidate many network equipment types into industry-standard high-volume servers, switches, and storage. It enables the implementation of network functions in software that can run on a range of industry-standard server hardware and can be moved to, or loaded in, various locations in the network as required, without the need to install new equipment. RAN virtualization use cases are described, but not yet addressed, in the current ETSI NFV recommendations. Meanwhile, the newly created ETSI MEC aims to offer application developers and content providers cloud-computing capabilities and an IT service environment at the edge of the mobile network.

The mobile network aspect of 5G will be led by the 3rd Generation Partnership Project (3GPP). Partners in 3GPP are still awaiting a consensus on 5G requirements before concrete actions are taken. Some aspects of virtualized network functions have already been addressed, such as in the SA2 system architecture working group with the new work item on flexible mobile service steering (FMSS) in the operator's core network. However, substantial impact on specifications in 3GPP RAN working groups cannot be expected before future LTE Releases 14, 15, and beyond. Thus it can be expected that virtualization and software control may help simplify the network architecture and support the flexible allocation of radio processing functionalities.

[2] http://www.ict-ijoin.eu

[3] www.metis2020.com

[4] http://www.etsi.org/technologies-clusters/technologies/nfv

[5] http://www.etsi.org/technologies-clusters/technologies/mobile-edge-computing

A first glimpse of RAN-related efforts is already visible in the network virtualization work stream of the Small Cell Forum [14], which analyzes requirements of RAN virtualization. In particular, different functional splits and their associated performance benefits and constraints are discussed.

Conclusions and Further Challenges

This article discussed the challenges, benefits, and opportunities of virtualizing RAN functions. We paid particular attention to the data-processing requirements, which are directly influenced by the operation and design of the RAN itself. Using results from this complexity analysis, we discussed the main contributors to the cost of a virtualized RAN system. We further discussed prominent implementation challenges such as joint resource optimization for RAN and the cloud computing platform.

Based on the discussion in this article, we can conclude that virtualized RAN provides greater flexibility to the mobile network operator and potentially reduces network costs. It is our opinion that RAN virtualization will be an integral part of 5G and that commodity IT platforms have the potential to host cloud RAN networks. However, there are many challenges beyond those tackled in this article, such as communication interfaces within data centers, parallelization of RAN functions, state maintenance, and the impact of the RAN protocol stack. Many of these aspects are detailed in [9].

There are many challenges beyond those tackled in this article, such as communication interfaces within data centers, parallelization of RAN functions, state maintenance, and the impact of the RAN protocol stack

Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Program FP7/2007-2013 under grant agreement no. 317941 – project iJOIN, http://www.ict-ijoin.eu. The European Union and its agencies are not liable or otherwise responsible for the contents of this document; its content reflects the view of its authors only.

References

[1] T. Taleb, "Toward Carrier Cloud: Potential, Challenges, and Solutions," *IEEE Commun. Mag.*, June 2014.

[2] P. Rost ***et al.***, "Cloud Technologies for Flexible 5G Radio Access Networks," *IEEE Commun. Mag.*, May 2014.

[3] D. Wübben ***et al.***, "Benefits and Impact of Cloud Computing on 5G Signal Processing," *IEEE Signal Proc. Mag.*, Nov. 2014.

[4] J. Kerttula ***et al.***, "Implementing TD-LTE as Software Defined Radio in General Purpose Processor," *ACM SIGCOMM Software Radio Implementation Forum 2014*, Chicago (IL), USA, Aug. 2014.

[5] H. Guan, T. Kolding, and P. Merz, "Discovery of Cloud-RAN", *Cloud-RAN Wksp.*, Apr. 2010.

[6] G. Banga, P. Druschel, and J. C. Mogul, "Resource Containers: A New Facility for Resource Management in Server Systems," *Proc. Symp. Operating Systems Design and Implementation*, New Orleans, LA, USA, vol. 99, Feb. 1999, pp. 45–58.

[7] P. Rost, S. Talarico, and M.C. Valenti, "The Complexity-Rate Tradeoff of Centralized Radio Access Networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6164-6176, Nov. 2015.

[8] V. Suryaprakash, P. Rost, and G. Fettweis, "Are Heterogeneous Cloud-Based Radio Access Networks Cost-Effective?" *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, October 2015

[9] EU FP7 Project iJOIN, "iJOIN Deliverable D5.2: Final Definition of iJOIN Requirements and Scenarios," Nov. 2014.

[10] Open Networking Foundation, "Software-Defined Networking: The New Norm for Networks" (online document]), white paper, Apr. 2012.

[11] S. Bhaumik ***et al.***, "CloudIQ: A Framework for Processing Base Stations in a Data Center," *IEEE MobiCom*, Istanbul, Turkey, Aug. 2012.

[12] 4G Americas, "Recommendations on 5G Requirements and Solutions," white paper, Oct. 2014.

[13] 3GPP, "TR 36.839 V11.1.0, Mobility Enhancements in Heterogeneous Networks," technical report, Sept. 2012.

[14] Small Cell Forum, "Virtualization for Small Cells: Overview," technical report, June 2015.

[15] Open Networking Foundation, "OpenFlow-Enabled Mobile and Wireless Networks," technical report, Sept. 2013.

Biographies

Peter Rost [SM] (peter.rost@ieee.org) received his Ph.D. degree from Technische Universität Dresden, Dresden, Germany, in 2009 (supervised by Prof. G. Fettweis), and his M.Sc. degree from the University of Stuttgart, Stuttgart, Germany, in 2005. Since May 2015 Peter has been a member of the Radio Systems research group at Nokia Network, focusing on 5G RAN architecture. From 2010 to 2015 he was a member of the Wireless and Backhaul Networks group at NEC Laboratories Europe. He has been active in 3GPP RAN2 and several EU projects (e.g. EU FP7 iJOIN as technical manager (www.ict-ijoin.eu)). He is a member of IEEE ComSoc GITC, IEEE Online GreenComm Steering Committee, VDE ITG Expert Committee "Information and System Theory." He is an executive editor of *IEEE Transactions on Wireless Communications*.

Ignacio Berberana (ignacio.berberana@telefonica.com) received the M.S. degree in mining engineering from Madrid Polytechnic University in 1987. In 1987 he enjoyed a National Research Grant for studying adaptive control systems. In 1988 he joined Telefonica I+D (Telefonica research labs), where he has worked in areas covering satellite and wireless communications, including several European projects (CODIT, MONET, Rainbow, Artist4G, iJOIN). Currently he is responsible for the Innovation unit in the Radio Access Networks direction of the Telefónica Global CTO office, which deals with long term evolution of mobile access, including 5G systems.

Andreas Maeder [M] (andreas.maeder@nokia.com) is a senior researcher at Nokia Networks, Munich, Germany. He has been with NEC Laboratories Europe in Heidelberg, Germany from 2008 to 2015. Andreas received his Ph.D. from the Unversity of Wuerzburg, Germany. Currently, his main area of research is the convergence of IT and telecommunication technologies. Andreas has contributed to the standardization of broadband wireless access technologies in IEEE and 3GPP since 2008. He was rapparteur of the work item on user plane congestion management in 3GPP SA2. He was chair of the IEEE BWA workshop 2013, and has authored numerous scientific articles, conference papers, and patents.

Henning Paul [M] (paul@ant.uni-bremen.de) received his Dr.-Ing. (Ph.D.) and Dipl.-Ing. (equivalent to M.Sc.) degrees from the University of Bremen, Germany in 2012 and 2007, respectively. He has been with the Department of Communications Engineering, University of Bremen, Germany since 2007, where he currently is a senior researcher and lecturer. His research is focused in the field of wireless sensor networks, in-network processing, and distributed signal processing algorithms for mobile communications. He is member of VDE ITG.

Vinay Suryaprakash [M] (vinay.suryaprakash@alcatel-lucent.com) received his doctorate from the Technische Universität Dresden, Germany under the supervision of Prof. Gerhard Fettweis in 2014. He received his master of science in electrical engineering from the University of Southern California, Los Angeles in 2007, after which, as an employee of Cisco Systems Inc., San Jose, CA from 2008 to 2010, he was involved in analysis and testing of load balancers that help regulate traffic in large networks. His current research focuses on using stochastic geometry for the system level analysis of wireless networks. In 2013 he was nominated as one of the six finalists of the Qualcomm Innovation Fellowship 2013 from contestants all across Europe.

Matthew C. Valenti [SM] (Matthew.Valenti@mail.wvu.edu) received his Ph.D. from Virginia Tech, USA, and has been on the faculty of West Virginia University (WVU) since 1999, where he is currently a professor and the Director of the Center for Identification Technology Research (CITeR). His research interests are in wireless communications, cloud computing, and biometric identification. He serves on the Executive Editorial Committee of *IEEE Transactions on Wireless Communications*, as an editor for *IEEE Transactions on Communications*, and as chair of the Communication Theory Technical Committee of the IEEE Communications Society (ComSoc). He is active in the organization of several ComSoc sponsored conferences, including MILCOM, ICC, and Globecom.

Dirk Wübben [SM] (wuebben@ant.uni-bremen.de) received the Dipl.-ing. (FH) degree in electrical engineering from the University of Applied Science Münster, Germany, in 1998, and the Dipl.-ing. (Uni) degree and the Ph.D. degree in electrical engineering from the University of Bremen, Germany in 2000 and 2005, respectively. In 2001 he joined the Department of Communications Engineering, University of Bremen, Germany, where he is currently a senior researcher and lecturer. His research interests include wireless communications, signal processing, cooperative communication systems, and channel coding.

Armin Dekorsy (Dekorsy@ant.uni-bremen.de) is the head of the Department of Communications Engineering, University of Bremen. He received his Dipl.-Ing. (FH) (B.Sc.) degree from Fachhochschule Konstanz, Germany; the Dipl.-Ing. (M.Sc.) degree from the University of Paderborn, Germany; and the

Ph.D. degree from the University of Bremen, Germany, all in communications engineering. From 2000 to 2007 he worked as a research engineer at Deutsche Telekom AG, and as a distinguished member of technical staff (DMTS) at Bell Labs Europe, Lucent Technologies. In 2007 he joined Qualcomm GmbH as European research coordinator conducting Qualcomms' internal and external European research activities. He has long-term expertise in the research of wireless communication systems, baseband algorithms, and signal processing. Prof. Dekorsy has published more than 160 journal and conference publications and holds more than 17 patents in the area of wireless communications. Prof. Dekorsy is a member of the IEEE Communications Society and IEEE Signal Processing Society, the VDE/ITG expert committee on "Information and System Theory", and represents the University at ETSI, NETWORLD2020, and at the OneM2M forum.

GERHARD P. FETTWEIS [F] (gerhard.fettweis@vodafone-chair.com) earned his Ph.D. under H. Meyr's supervision from RWTH Aachen in 1990. After one year at IBM Research in San Jose, CA, he moved to TCSI Inc., Berkeley, CA. Since 1994 he has been Vodafone Chair Professor at TU Dresden, Germany, with 20 companies from Asia/Europe/US sponsoring his research on wireless transmission and chip design. He coordinates two DFG centers at TU Dresden, namely cfaed and HAEC. He is an IEEE Fellow and a member of the German academy Acatech. His most recent award is the Stuart Meyer Memorial Award from IEEE VTS. In Dresden his team has spun-out 13 start-ups, and set up funded projects in volume of close to EUR 1/2 billion. He has helped organize IEEE conferences, most notably as TPC Chair of ICC 2009 and TTM 2012, and as General Chair of VTC Spring 2013 and DATE 2014.