Benefits and challenges of cloud technologies for 5G architecture

Dario Sabella[1], Peter Rost[2], Albert Banchs[3], Valentin Savin[4], Marco Consonni[5], Marco Di Girolamo[5], Massinissa Lalam[6], Andreas Maeder[2], Ignacio Berberana[7]

[1] Telecom Italia, [2] NEC Laboratories Europe, [3] IMDEA Networks, [4] CEA-LETI, [5] HP Italy Innovation Center, [6] Sagemcom Broadband, [7] Telefonica I+D

Abstract— **This paper focuses on the practical implementation of a Cloud-RAN architecture in the context of future 5G systems, with particular emphasis on different aspects of the functional split between the cloud platform and the radio access points. First, we provide a comprehensive overview of implementation aspects and how different hardware options impact the implementation of RAN functionality. We further discuss a virtualized infrastructure which may have a significant impact on how algorithms are implemented, how they interact with each other, and how they can be scaled within the RAN. We also analyze implementation constraints to be considered to provide backwards compatibility with 3GPP LTE; such constraints on the computing platforms result from the RAN requirements in terms of latency and throughput. Finally, the level of flexibility achievable by the proposed architecture is described from a practical point of view.**

Keywords-component; Cloud-RAN, RAN-as-a-Service, energy efficiency, virtualization intrastructure, flexible funcitonal spliy

I. INTRODUCTION

Recently the C-RAN architecture [1] received a lot of attention from mobile operators, due to the advantages resulting from the usage of an energy efficient and green infrastructure, cost-saving on CAPEX & OPEX, as well as capacity improvement and adaptability to non-uniform traffic.

The final step of C-RAN architecture is the virtual RAN architecture, where highly reconfigurable general purpose Hardware (HW) interacts with a significant number of controlled cells, and Baseband (BB) resources can in principle be located in the same or different physical locations (in the cloud). In this phase, virtual RAN gets all the C-RAN advantages and adds the following ones (due to usage of General Purpose Hardware - GP HW):

- dynamical reallocation of processing resources within a centralized baseband pool to different virtualized base stations, in presence of different air interface standards;
- HW and Software (SW) totally decoupled for both, cost and management efficiency;
- simpler inter-vendor interoperability;
- cost reduction to manage, maintain, expand and upgrade the base station.

Nevertheless, in the view of practical implementation of this kind of architecture for 5G systems, some challenges still need to be addressed, in particular the extremely demanding backhauling requirements resulting from a full centralization of RAN functionalities.

In recent publications [2][3], we introduced the concept of RAN-as-a-Service (RANaaS), indicating a set of computing and storage infrastructure resources (typically a Cloud computing Infrastructure-as-a-Service (IaaS) platform, but potentially any industry standard, general purpose computing infrastructure) where part of the processing functions of the lowest OSI layer(s) (L1/L2), normally distributed across a number of base stations (eNB), are moved to the RANaaS and centralized. The remaining part of RAN functionalities is still decentralized in Radio Access Points (RAPs), as indicated in the Fig. 1. With this approach, partial centralized solutions are envisaged, in order to cope with backhauling bandwidth and latency constraints.

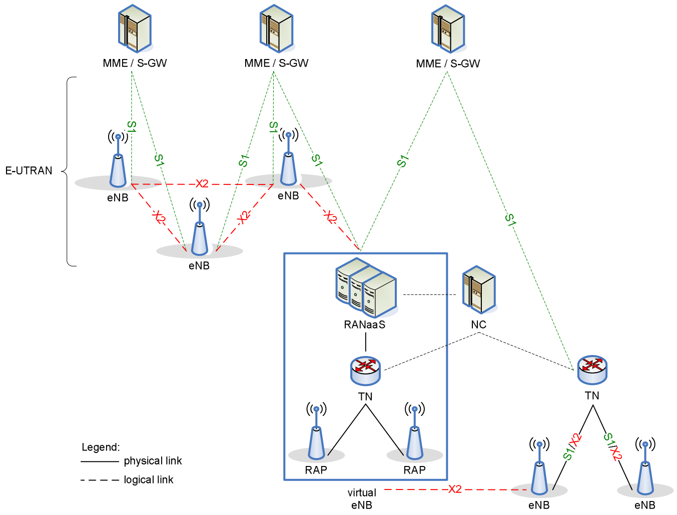

Fig. 1. RAN-as-a-Service concept in Cloud architecture toward 5G

978-1-4799-8088-8/15/$31.00 ©2015 IEEE

This paper goes further in examining the implementation aspects related to the functional split, as a reference for the introduction of a computational architecture for 5G Cloud-RAN systems. Moreover, flexible functional splits are considered for future systems, especially in order to adapt to changing network conditions or different deployment choices of the operator in different regions.

Section II describes the HW/SW implementation of the proposed Cloud Architecture, while Section III provides a deeper level of detail, analysing the constraints to be considered if the 3GPP LTE compliancy is targeted. Finally, in Section IV we discuss the flexibility of functional split assignment, and Section V concludes the paper.

II. Implementation of Cloud Architecture toward 5G systems

A. HW/SW partitioning

In general, there are three main options to implement the RANaaS concept (depicted in Fig. 1) on RAPs and RANaaS platform. Different functionalities could be implemented either on dedicated hardware – such as Application Specific Integrated Circuit (ASIC), Field Programmable Gate Arrays (FPGAs), or Digital Signal Processors (DSPs) – or on General Purpose Processors (GPPs). Furthermore, hybrid approaches are possible where a software implementation on GPPs is complemented with hardware accelerators. Fig. 2 illustrates the main options and how the proposed Cloud Architecture implementation can be categorized.

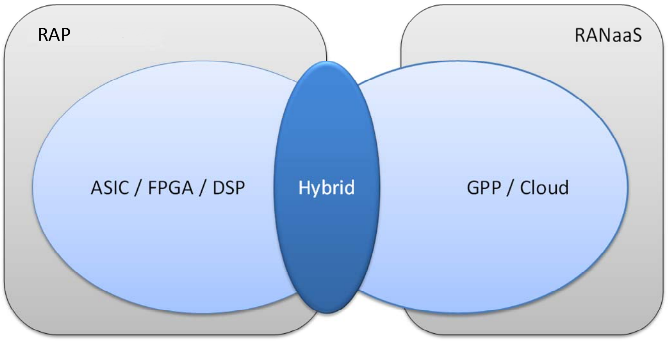

Fig. 2. HW/SW implementation of cloud architecture

Currently, state-of-the-art virtualized baseband processing in C-RAN is based on hybrid solutions consisting of GPPs with hardware accelerators or co-processors that implement specialized digital signal processing functionalities [5]. In the first case, HW accelerators can be implemented by means of DSPs, FPGAs, ASICs or a combination of them. GPPs can be based on ARM, MIPS or x86 ISAs (Instruction Set Architectures). Co-processors communicate with the CPU using a standard interface such as PCI Express. The approach is illustrated in Fig. 3 with an example of splitting of digital signal processing across different hardware (GPP, DSP, and FPGA). Algorithmic bottlenecks that prevent a pure-software implementation running on GPP can be eliminated by the use of custom hardware accelerators that offload data processing from the CPU. The same approach has been followed to support in an integrated way graphic processing capabilities (combination of CPU and GPU) or packet processing capabilities. The next logical step is to define a programming model for the co-processor that is also GPU-reminiscent, akin to DirectX or OpenGL's abstraction of a computer's graphics subsystem.

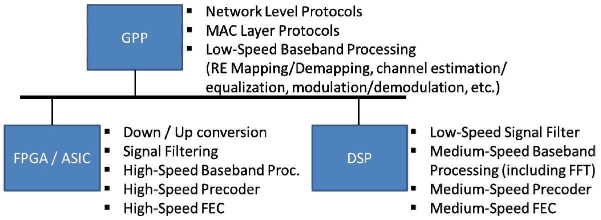

Fig. 3. Example of splitting of digital signal processing across GPP, DSP, and FPGA

It can be noticed that functionalities identified as “software,” e.g., channel estimation or MIMO processing, are not supported by the co-processor. It can be argued that the same kind of approach is already followed by the solutions provided for baseband processing in commercial base stations. However, two main differences between base station solutions and virtual RAN solutions should be stressed:

- In most baseband processing units for base stations, GPP responsibilities are limited to scheduling and coordination of DSP functionalities carried by specialized hardware, while in the virtualization solution, a significant part of the processing can be carried out by GPPs.
- Solutions for the virtualized architectures should support resource virtualization as understood in the Information Technologies (IT) realm, while base station solutions are dedicated to single access points.

The main reason for using a hybrid solution with co-processors is the fact that the full implementation of radio interface baseband processing by means of GPPs may be suboptimal in terms of required investment, energy consumption and other performance parameters. On the other hand, centralization of conventional baseband processing units does not allow for an easy virtualization of the resources and the reuse of IT solutions.

The support of the RANaaS concept and flexible functional split introduce a new level of complexity as it may require the solution to support different levels of processing without penalizing the network Total Cost of Ownership (TCO). It should be noted that the solutions previously described are expected to be deployed within a RANaaS instance while the distributed elements are RAPs, which only support a (limited) set of baseband processing functionalities depending on the functional split.

In the context of the proposed architecture, the RAP may implement different levels of baseband processing based on the functional split. This may range from only RRH functionalities, as in C-RAN, to a full support of the whole radio interface protocol stack, as in a conventional distributed implementation. The same operating scenarios are applicable to the RANaaS infrastructure. The requirements for an ideal solution would be the following:

- It should be possible to reuse the same solution for both RAPs and RANaaS infrastructure in such a way that processing elements may be moved from the RAP to the RANaaS and vice versa.

- It should be possible to switch off those processing units (CPU cores, DSP co-processors) that are not required for the selected functional split.
- It should be possible to virtualize the capabilities of the processing elements in such a way that the functionalities they implement may be decoupled from their locations. For instance, the processing elements of a RAP may be used for processing connections of other RAPs if the backhaul infrastructure provides the necessary connectivity.
- It should be possible to reuse the same solution for the virtualization of other network elements, not necessarily of the mobile network, e.g., virtualization of CPEs or implementation of virtual switches.
- It should be possible to scale the usage and provision of the data processing capacity with the actual data traffic load that is processed within the RANaaS platform.

The main argument to deploy hybrid solutions within datacenters is the required computational complexity of RAN processing and the energy and cost-efficiency of GPP hardware. While custom hardware solutions offer a performance advantage over GPP based solutions, they do not offer the same scalability and flexibility as GPP based solutions.

In particular, custom hardware solutions are not able so far to scale the consumed data processing complexity with the actual data traffic demand. This implies a potential overprovisioning by solutions which are not based on GPPs even in the case of centralized processing. The assignment of functionality to custom hardware or GPP allows for an additional degree of freedom.

Nevertheless, in order to reduce the complexity, in our approach, we consider the virtualization of RAN processing where any functionality executed at the RANaaS platform is executed on GPPs and any functionality executed at RAP is either implemented in dedicated hardware or GPPs for providing a fall-back solution at the RAP (as in Fig. 2).

B. *Virtualization infrastructure*

As described in the previous section, common solution designs for supporting virtualized baseband processing use a combination of general purpose processors with hardware accelerators or co-processors for implementing specialized digital signal processing functionalities. Mapping this kind of architecture into an IaaS platform raises the problem of how to distribute the related workload on a virtualized platform. Specific attention must be paid for parallel computation which is traditionally obtained using specialized hardware devices, e.g., DSPs or FPGAs.

IaaS platforms implement server virtualization where more than one virtual server runs on top of a single physical computer. This is implemented by using a hypervisor which runs on the physical hardware and takes care of running several virtual servers.

Through virtualization, IaaS platforms provide the low-level building blocks for implementing systems where parallel programming can run. In fact, each Virtual Machine (VM) can utilize multiple cores of the hosting physical machine for elaborating the assigned workload. In addition, several VMs can be activated and can collaborate for expanding the computing power dedicated to the workload at hand. In this manner, parallelization can be implemented either within a single VM or across collaborating VMs.

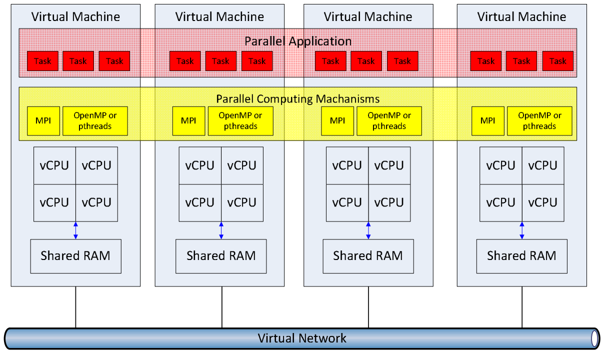

Fig. 4. Hybrid Programming Model on an IaaS VM Cluster

Fig. 4 shows a parallel application distributed on a cluster of VMs, where each VM runs a certain number of tasks. A software technology such as Open Multi-Processing (OpenMP) or Pthreads is used for managing the tasks running within a single VM. It provides mechanisms for creating, coordinating and synchronizing tasks which share the memory of the VM (Parallel Computing Mechanisms layer). Another software technology, such as a Message Passing Interface (MPI) implementation, is used for managing and interconnecting tasks running on different VMs (Parallel Application layer).

III. IMPLEMENTATION CONSTRAINTS

The previous section has addressed the capability of a cloud platform using GPP to perform RAN functionalities (with the PHY layer demanding the most extensive computational task) and related virtualization infrastructure considered in RANaaS platform. Many functional splits can be considered (see Fig. 5) to implement the cloud infrastructure described in Fig. 1. Nevertheless, if nothing prevents one functional split to be applied in theory with any kind of backhaul, there are still strong timing constraints to consider if the 3GPP LTE compliancy is targeted. The lower we perform the functional split in the protocol stack, the more bandwidth is required to support the forwarding of the data between the RAP and the RANaaS instance. In addition, the split point within the PHY processing chain itself (see Fig. 5) can also significantly impact the required bandwidth as shown in [6] and [9].

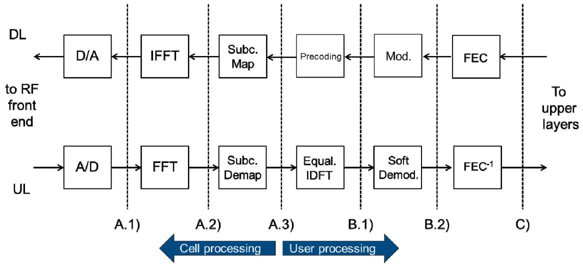

Fig. 5. Functional split options for the PHY layer [6]

More important than the bandwidth requirements, the timing requirements must be carefully considered in the functional split decision. Indeed, 3GPP has defined many timers for each of the layer (from MAC to RRC) which dictate the behaviour of the complete LTE system. They may impose some serious constraints on the feasibility of one specific functional split within a legacy 3GPP LTE ecosystem. In [7] and [8], those timers have been all gathered and analysed. Table I presents the identified ones that may possibly impact a functional split decision. Many of the other higher layer timers not presented here are indeed configurable within a specified range definition which is large enough to allow for a setting adapted to the required backhaul and processing time.

TABLE I. 3GPP TIMING REQUIREMENTS [7]

	Timer	Short description	Max Value
PHY	Subframe	Physical subframe length	1 ms (fix)
	Frame	Physical frame length	10 ms (fix)
MAC	HARQ RTT Timer	When an HARQ process is available	8 ms (fix)
RLC	t-PollRetransmit	For AM RLC, poll for retransmission @TX side	500 ms
	t-Reordering	For UM/AM RLC, RLC PDU loss detection @RX side	200 ms
	t-StatusProhibit	Prohibit generation of a status report @RX side	500 ms
PDCP	discardTimer	Discard PDCP SDU / PDU if expiration or successful transmission	Infinite
RRC	TimeToTrigger	Time to trigger of a measurement report	5.12 s
	T300	RRCConnectionRequest	2 s
	T304	RRCConnectionReestablishment Request	2 s
	T310	RRCConnectionReconfiguration	2 s or 8 s
	T311	Detection of physical problem	2 s
	T304	(successive out-of-sync from lower layers)	30 s

The heaviest timing constraint will appear as soon as the uplink HARQ process will be performed in the RANaaS instance and not at the RAP. In the case a transmission attempt failed, the HARQ procedure repeats the same message (Chase combining) or a different encoding of the same message (incremental redundancy). To avoid a communication to stall waiting for the acknowledgment (ACK) of a message transmission, several HARQ processes are used in parallel, allowing one message transmission to be done while waiting for a previous message transmission to be acknowledged. In FDD, eight HARQ processes are defined which fits well the way data are acknowledged in particular in the uplink. When one UE is scheduled for transmission in subframe *n*, it received this scheduling information in subframe *n*-4. At subframe *n*+4, the UE is expecting an ACK from the RAP which can either implicitly schedule a retransmission or explicitly schedule a (re)transmission at subframe *n*+8. The HARQ process used at subframe *n* can therefore not be used before subframe *n*+8. Thus, the round-time trip (RTT) timer of one HARQ process is defined by 8 ms as listed in Table I. In uplink LTE, a synchronous behavior is implemented. This implies that one HARQ process can only be used in modulus eight subframes, i.e., eight HARQ processes in parallel. This has a strong impact on the functional split implementation in the case that decoding processing is performed at the RANaaS instance. Indeed, a UE expects a positive or negative ACK of each transmission within 4 subframes, i.e., 4 ms. This implies that the two-way transmission from the RAP to the RANaaS physical Point of Presence (PoP) and the decoding itself must finish in less than ACK in the next subframe (*n*+4). This timing requirement is the heaviest constraint with such functional split.

If a legacy UE fails to receive this ACK in subframe *n*+4 for an HARQ process used at subframe *n*, it will not be able to use this HARQ process at subframe *n*+8, and a next scheduling occasion which will only be possible at subframe *n*+12 for subframe *n*+16. An LTE-compliant solution can be for the RAP to send a positive ACK on the PHICH, even if the result of the decoding is not yet known. By not sending scheduling information on the PDCCH, the UE will not flush its HARQ process unless receiving an explicit order to do so, i.e. with the NDI indicator set to 1. This compatibility comes at the cost that the possible maximum throughput of the UE will be reduced.

IV. FLEXIBLE FUNCTIONAL SPLIT ASSIGNMENT

The objective of this section is to analyze, from a practical viewpoint, the actual flexibility that can be achieved in terms of functional split. Ideally, a fully virtualized environment where each function could be moved at any place would be feasible. In practice, there will be different small-cell implementations which may or may not support different functional splits, or different co-processors which can be turned on/off. In addition, the assignment of the RANaaS PoP cannot be assumed to be changed on the fly.

In practice this means that the set of supported functional splits will be a reduced set. If only one functional split would be supported, it will be the one that provides the best ratio of potential benefits associated to the centralization degree with respect to potential cost variations. The benefits that can be obtained from centralization are mainly associated to increased spectral and energy efficiencies [4]. This ratio does not only depend on technical factors but also on other factors such as the traffic demand, the user distribution, and the possibility of reusing deployed infrastructure. For example, the centralization of the baseband processing may allow for an implementation of cooperation mechanisms that may help to improve the spectral efficiency, reducing the need for new deployments in high traffic demand areas and making it a sensible option from a techno-economic viewpoint. But if demand is relatively low or other solutions such as carrier-aggregation are available, then it may be that the opposite is reached. On top of this, different technical criteria should also be taken into account:

- Cell based vs. user based processing (see Fig. 5): cell based processing should be distributed as far as possible, because it reduces the transport requirements and does not exhibit potential processing multiplexing gains. On top of this, this processing is better implemented using hardware solutions;
- Software based processing vs. hardware based processing: As has been indicated in previous sections,

some functionality is more efficiently implemented by means of hardware based solutions, while others benefit from a software based implementation;

- Latency requirements: Some processing functionalities are very sensible to latency. Obviously, this factor determines whether a potential centralization in the RANaaS infrastructure is feasible or not.

Based on this description and the previous section, it may be feasible that the implementation at the RAP splits into two parts: a hardware implementation and a software implementation. Then, based on the actual functional split, individual modules would be turned on and off. Each module is of course composed by a functionality of the eNB radio protocol stack (and some of these modules may be implemented in hardware and some in software). In a practical setup, a RAP may only support two functional splits:

- A preferred functional split where functionality at the RAP is executed on hardware and all remaining functionality is executed in software at the RANaaS entity. The individual modules at the RAP may be implemented on different co-processors in order to allow for flexible functional split configurations based on a single platform.
- A fall-back solution which is applied in the case that the RANaaS platform is not used and the RAP connects directly to the core network. In this case, all upper layers are executed in software at the RAP in addition to the functionality that is executed in the case of the preferred functional split.

In addition, a third intermediate functional split option between preferred functional split and full decentralization could be supported by the RAP. This could be useful in order to adapt to the backhaul network as well as the current computational load of the RANaaS platform.

V. Conclusions

This paper analyzed the aspects related to the practical implementation of Cloud-RAN architecture in the view of future 5G systems, with particular emphasis on different aspects of the functional split between RANaaS platform and radio access points. First, we provided a comprehensive overview of implementation aspects and how different hardware options impact the implementation of RAN functionality. We further discussed a virtualized infrastructure which may have a significant impact on how algorithms are implemented, how they interact with each other, and how they can be scaled with the RAN. Then we described implementation constraints to be considered if the 3GPP LTE compliancy is targeted. Finally, we analysed, from a practical viewpoint, the actual flexibility that can be achieved in terms of functional split.

Future work may include the assessment of computational performance in the RANaaS testbed currently hosted by Telecom Italia.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 317941. The authors would like to acknowledge the contributions of their colleagues in iJOIN, although the views expressed are those of the authors and do not necessarily represent the project.

References

[1] C-RAN: The Road Towards Green RAN, China Mobile Research Institute, http://labs.chinamobile.com/cran/wp-content/uploads/ CRAN_ white_paper_v2_5_EN.pdf

[2] D. Sabella, P. Rost, Y. Sheng, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamo, M. Di Girolamo, and G. Guiliani, “RAN as a Service: Challenges of designing a flexible RAN architecture in a cloud-based heterogeneous mobile network,” 2013 Future Networks Summit, July 2013, Lisbon, Portugal.

[3] P. Rost, C.J. Bernardos, A. De Domenico, M. Di Girolamo, M. Lalam, A. Maeder, Dario Sabella, D. Wübben “Cloud technologies for flexible 5G radio access networks,” IEEE Communications Magazine, May 2014.

[4] D. Sabella, A. De Domenico, E. Katranaras, M. Imran, M. Di Girolamo, U. Salim, M. Lalam, K. Samdanis, A. Maeder, “Energy Efficiency benefits of RAN-as-a-Service concept for a cloud-based 5G mobile network infrastructure”, IEEE Open Access 2014.

[5] G. Li, S. Zhang, X. Yang, F. Liao, T. Ngai, S. Zhang, and K. Chen, “Architecture of GPP based, scalable, large-scale C-RAN BBU pool,” in IEEE GLOBECOM 2012 Workshops, International Workshop on Cloud Base-Station and Large-Scale Cooperative Communications, Anaheim, CA, USA, Dec.2012.

[6] D. Wübben, P. Rost, J. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, “Benefits and impactImpact of cloud computingCloud Computing on 5g signal processing”. IEEE Signal Processing Magazine, Special Issue 5G Signal Processing, November 2014.

[7] A. Maeder, M. Lalam, A.D. Domenico, E. Pateromichelakis, D. Wübben, J. Bartelt, R. Fritzsche, P. Rost, “Towards a Flexible Functional Split for Cloud-RAN Networks”, European Conference on Networks and Communications, July 2014, Bologna, Italy.

[8] iJOIN, Deliverable D3.2 “Definition of MAC and RRM approaches for RANaaS and a joint backhaul/access design,” report, November 2014

[9] iJOIN, Deliverable D2.2 “Definition of PHY layer approaches that are applicable to RANaaS and a holistic design of backhaul and access network,” report, November 2014.