Gkatzios et al. EURASIP Journal on Wireless Communications and Networking (2019) 2019:183
https://doi.org/10.1186/s13638-019-1488-z

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Efficiency gains in 5G softwarised radio access networks

Nikolaos Gkatzios[1], Markos Anastasopoulos[2*], Anna Tzanakaki[1,2] and Dimitra Simeonidou[2]

Abstract

This paper proposes the concept of compute resource disaggregation in centralized softwarised radio access networks. This approach allows individual allocation of processing functions to different servers depending on the nature and volume of their processing requirements. The benefits of the proposed approach are evaluated through experiments analyzing the baseband unit (BBU) processing requirements of LTE PHY using an open source suite for wireless systems and a purposely developed multistage linear programming modeling framework. To address the high computational complexity of integer linear programming (ILP), limiting real-time services, a heuristic algorithm was also developed. Our results show that when using the proposed approach, the overall system power consumption is reduced by 40% under high loading scenarios compared to the traditional solution where all BBU functions are hosted in the same physical servers. The heuristic results achieve relatively good agreement with the ILP results (the gap is less than 5%), particularly for low and high loading conditions. At the same time, the heuristic requires less than 0.3 s to identify the optimal allocation policy.

Keywords: RAN, BBU, Softwarized radio, Compute disaggregation, Multistage linear programming

1 Introduction

The increase of mobile traffic predicted in 5G networks as well as the wireless access technology densification and advancements proposed to address this introduces very stringent requirements in the radio access networks (RANs). Traditionally distributed RAN solutions, where baseband units (BBUs) and radio units (RUs) are co-located, suffer several limitations. To overcome these limitations, cloud radio access networks (C-RANs) have been proposed. In C-RAN, distributed remote radio heads (RRHs) are connected to a BBU pool, the central unit (CU), through high bandwidth transport links known as fronthaul (FH). The CU can be hosted in data centers (DCs) comprising general purpose processors (GPPs) that can be accessed through FH services supported by the transport network of the 5G infrastructure [1]. The interface between RUs and CU is standardized through the Common Public Radio Interface (CPRI).

In this environment, it is very important to identify the optimal allocation of BBU functions to the appropriate servers hosted by the CU, as it is expected to give significant efficiency gains (such as power consumption). To the best of the authors' knowledge, up to date, this is performed without taking into consideration the details and specificities of the individual processing functions that BBUs entail. To take advantage of appropriate mapping of processing functions to suitable available compute resources within the CU, we have proposed the concept of compute resource disaggregation [2]. This approach allows the individual allocation of processing functions, associated with a specific FH service, to different servers depending on the nature and volume of their processing requirements. To date, several studies have focused on optimal BBU placement through 5G network topology design [3] and the optimal optical network design serving 5G transport network requirements [4]. For a survey on C-RAN architectures and technologies used, the reader is referred to [5]. In addition, work on identifying optimal BBU functional split options over integrated wireless/optical 5G infrastructures has been reported in [1, 6–12]. Specifically, in [12], the authors investigate various functional split architectures and connectivity options between the RRHs and the BBU with the objective to reduce the associated FH bandwidth requirements. However, most of these studies are focusing on the inter-DC network design

* Correspondence: m.anastasopoulos@bristol.ac.uk
[2]HPN Group, University of Bristol, Bristol, UK
Full list of author information is available at the end of the article

problem supporting C-RAN services while little attention has been given on the intra-DC domain. Given that the operation of future C-RAN networks will be supported by virtualized BBU that will operate in a combination of general and specific purpose servers [1, 4], it is necessary to analyze the specificities and characteristics of the individual processing functions forming the BBU service chain (SC). One of the first studies on the topic has been reported in [13] where a closed-form approximation of the energy consumed at the base stations (BSs) has been developed, where based on the type of BSs (macro, pico, femto), the consumed power at the ASICs supporting BBU processing is estimated. However, the performance of virtualized BBUs running of general purpose processors (i.e, × 86) has not been considered.

In order to respond to this observation, in the present study, we rely on an implementation of the LTE protocol stack, namely WiBench [14]. WiBench is an open source kernel suite for benchmarking wireless systems available at [15]. Using this platform, the BBU processing requirement of individual LTE PHY are analyzed for various wireless access requirements and traffic load scenarios. Once the construction elements of the BBU SC have been analyzed, a multistage integer linear programming (ILP) modeling framework was developed able to assign the construction elements of the BBU chain to the suitable servers hosted by the CU. In the majority of the existing LTE PHY implementation, the whole BBU SC runs on a single physical machine. Typical example includes WiBench, [16], the "All-in-one OpenAirInterface" [17], and srsLTE [18]. A cloud-based architecture supporting the C-RAN paradigm is reported in [19] where the multiside/standard baseband unit (MSS-BBU) is introduced that provides *radio control*, *user processing* (UP), and *cell processing* functionalities. In the present study, these models are referred to as softwarized BBUs (SW-BBU). In [19], the only function that is virtualized (and possibly relocated) is the UP. UP comprises layer 3, layer 2, and part of PHY layer functions. In addition, this paper presents an architecture where multiple RRHs are connected to a baseband processing pool (MSS-BBU). Different MSS-BBUs, in different locations, are interconnected with each other. A Decentralized Cloud Controller (DCC) is connected to every MSS-BBU, and it manages the load balancing inside the MSS-BBU (intra) and between different MSS-BBUs (inter).

The present study is differentiated from [19] in several ways. First, we consider a heterogeneous DC system comprising GPPs able to process the BBU functions. To achieve this, an energy efficient resource allocation scheme is proposed that assigns the BBU functions to the appropriate server. The output of our experiments was used as a realistic input to our ILP model in order to evaluate the energy consumption requirements of the compute resources for the proposed and the traditional SW-BBU approach. As the proposed approach is based on the concept of disaggregation, we refer to it as the disaggregated SW-BBU (DSW-BBU) approach. ILP's results show that the proposed DSW-BBU approach can provide significant benefits in terms of energy consumption and as such operational expenditure associated with the BBU functions.

To address the computational complexity of the ILP, a heuristic algorithm was also developed, with the aim to also assign the construction elements of the BBU chain to the suitable servers hosted by the CU with minimum power consumption. The comparison of the numerical results produced by the different approaches (i.e. ILP and heuristic) confirms that the heuristic approach performs similarly with the ILP for low and high loading conditions. The rest of the paper is organized as follows. In Section 2, the problem formulation and the system model are presented. Section 3 describes the LTE PHY uplink benchmarking framework, the multistage ILP model, and the heuristic which were developed addressing the optimal BBU functions' placement. The numerical results of the ILP and the heuristic are presented and analyzed in Section 4. Finally, conclusions are drawn in Section 5.

2 Problem statement

We consider a generic 5G C-RAN where the processing requirements of a set $\mathcal{R}$ of R RRHs are supported by a set of compute resources located at the CU. Compute resources comprise a set $\mathcal{S}$ of S GPPs with individual processors of processing capacity C_s and performance per watt (measured in instructions per second—IPS) P_s. Servers are interconnected in accordance to a simple tree structure shown in Fig. 1 and are responsible to provide the required processing power for the support of FH services. This is achieved through the execution of baseband signal processing-related tasks required for the operation of RRHs. The compute requirements for baseband processing for RU $r, r \in \mathcal{R}$, can be calculated as the sum of all contributing computing elements responsible to perform the required functions, including single carrier-frequency division multiple access (SC-FDMA) demodulation (c_{r1}), subcarrier demapper (c_{r2}), frequency domain equalizer (c_{r3}), transform decoder (c_{r4}), constellation demapper (c_{r5}), descrambler (c_{r6}), rate matcher (c_{r7}), and turbo decoder (c_{r8}). As shown in Fig. 1, these functions need to be executed in a specific order.

The main objective of this work is to identify the optimal GPP server where each function can be allocated so that the total power consumption at the DC can be minimized satisfying, at the same time, the strict quality of service (QoS) constraints imposed by the CPRI protocol. To achieve this, we initially calculate the actual

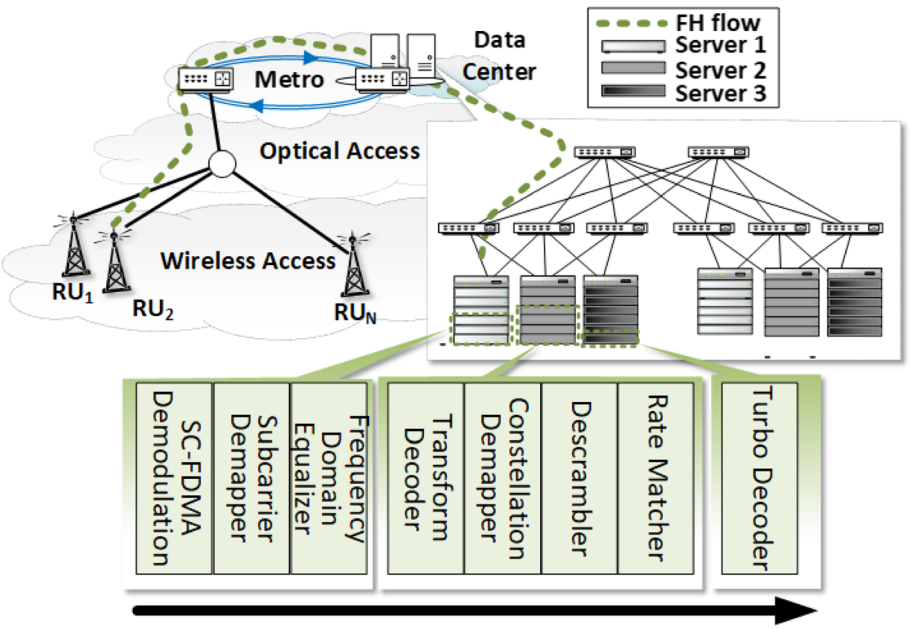

Fig. 1 Centralized processing of softwarised RAN functions on a data center hosting a different type of servers

processing requirements of each baseband processing function, in terms of total instructions, under various wireless access system configurations. These calculations are carried out using WiBench, an open source software implementation of the LTE protocol stack [14]. The processing requirements of each function are then used as input to a multistage ILP-based optimization framework that is able to assign each baseband subtask to the suitable GPP in an energy efficient manner. Although the problem of BBU placement has been studied by several authors [3, 4, 11, 20], the vast majority of these consider the BBU chain as a whole, without considering the specificities of its individual construction elements. In this study, however, it is shown that by (i) *disaggregating* the softwarized BBU into a set of smaller subtasks, (ii) *analyzing in depth the computational requirements* of each subtask, and (iii) *allocating* these subtasks to suitable GPPs as appropriate, significant benefits in terms of the operation efficiency of future 5G systems can be achieved. This study extends the state-of-the-art through the following:

- An extensive set of experiments used to characterize the processing requirements of the baseband functions as a function of the operational parameters of the wireless access network. These experiments led to the extraction of simple mathematical relations that can be used by network designers and operators to optimally allocate and size DC networks under various 5G network operational conditions.
- The development of an energy-aware multistage ILP-based optimization framework able to assign the BBU subtasks to a heterogeneous set of GPP elements reducing the DC power consumption by 20%.
- The development of a heuristic algorithm, with low computation complexity, able to route in real time the input BBU traffic in a heterogeneous set of GPP elements inside a DC.

3 Methods/experimentals

3.1 Benchmarking framework

3.1.1 Experimental platform description

For our experiments, we used WiBench, an open source suite for benchmarking wireless systems, and Intel's VTune Amplifier 2018 [21], a performance profiler for software performance analysis. WiBench provides various signal processing kernels. These kernels are configurable and can be used to build applications to model wireless protocols. The LTE PHY uplink that was used for the experiments was provided by the WiBench suite, and VTune was used to profile the LTE application. A summary of the BBU processing functions is presented below. This includes the following:

- The single carrier-frequency diversity multiple access that is a precoded orthogonal frequency diversity multiplexing (OFDM). It is preferred compared to OFDM, for the uplink transmission, as it is less susceptible to frequency offsets and has a lower peak-to-average power ratio. The SC-FDMA

demodulation function removes the cyclic prefix (CP) and performs N-point fast Fourier transform (FFT).
- The subcarrier demapper that extracts the data and the reference symbols from the subframes.
- The frequency domain equalizer that estimates the channel state information (CSI) by the received pilot signal through the least square estimation algorithm. It computes the channel coefficients, with the help of CSI, and equalizes the received data using a zero-forcing MIMO detector in the frequency domain as an equalizer.
- The transform decoder that performs M-point inverse fast Fourier transfer (IFFT).
- The constellation demapper that receives the signal and extracts the binary stream by generating logarithmic likelihood ratios (LLR).
- The descrambler that descrambles the input sequence.
- The rate matcher that separates the input stream into N streams, deinterleaves each code stream, and removes the redundant bits. For our experiments, one information bit is encoded into three transmitted bits, so N was constantly set to 3.
- The turbo decoder that takes soft information for each code, in our case LLR, and applies iteratively the soft-input soft-output (SISO) algorithm. The turbo decoder consists of two SISO decoders that perform the trellis traversal algorithm and one interleaver/deinterleaver. Higher number of iterations achieves an improved error correction performance, at the expense of higher computation cost. To address this issue, for the conducted experiments, 5 iterations were used [14].

3.1.2 Quantifying processing requirements of BBU functions

To increase the statistical validity of the results produced by the profiler, a thorough investigation between different numbers of subframes processing was conducted which resulted in setting the number of subframes to 1000. The set of experiments carried out was aiming at exploring the behavior of each processing function for different configurations of the LTE PHY uplink system. Figure 2 presents the dependence of the instructions performed on the data rate for different modulation schemes, when processing 1000 subframes by each function.

Taking into consideration the variance of the measurements, we can conclude that all functions present a linear dependence with the data rate. On the other hand, the influence of the modulation scheme, on the instructions number, differs for each function. More specifically, we observe that the modulation scheme does not affect the instruction number for SC-FDMA demodulation, subcarrier demapper, equalizer, and transform decoder. For the constellation demapper, an exponential dependence of the modulation scheme is observed, while the rate matcher and the turbo decoder exhibit linear dependence.

We observe that the turbo decoder performs a higher number of instructions, especially as the data rate increases, while the constellation demapper, the rate matcher, and the equalizer perform fewer instructions. This means that the turbo decoder, involving 1 to 4 orders of magnitude higher instructions compared to other functions, determines by large the total number of instructions needed to process a subframe and how this number depends on the data rate and the modulation scheme. Below are the linear expressions that fit the turbo decoder (Eq. (1)) and the total instructions (Eq. (2)) behavior.

$$\text{Instructions (million)} = 13747 \times \text{data rate (Mbps)} \quad (1)$$

$$\text{Instructions (million)} = 14575 \times \text{data rate (Mbps)} \quad (2)$$

3.2 Optional placement of softwarized RAN functions

Once the computational requirements c_{ri}, $i = 1, \ldots, 8$, of the dissagregated RAN functions for RU $r \in \mathcal{R}$ have been determined (Fig. 2), a multistage ILP modeling framework able to assign the construction elements of the BBU chain to the suitable servers is proposed. During stage 1, the first function of all fronthaul flows in the service chain (i.e., function SC-FDMA demodulation) reaching the CU will be assigned to servers $s \in S$ (Fig. 3). This is achieved by minimizing the total compute resource power consumption, approximated through the following cost function:

$$f_1(\mathbf{x}_1) = \sum\nolimits_{s \in S} \mathcal{E}_s \left(\sum\nolimits_{r \in R} x_{rs1} c_{r1} \right) \quad (3)$$

In Eq. (3), the summation $\sum_{r \in \mathcal{R}} x_{rs1} c_{r1}$ captures the total processing load of all c_{r1} functions processed at server s, x_{rs1} is a binary decision variable indicating whether function c_{r1} of RU $r \in \mathcal{R}$ is processed at server s or not, $\mathbf{x}_1$ is a vector containing all first stage decision variables x_{rs1}, and $\mathcal{E}_s$ is the power consumption model of server s.

Now let $Q_{\mathcal{R}1}$ be a set of paths for the FH flow of RU r, $r \in \mathcal{R}$ interconnecting the ingress node to server s, z_{rq} be the network capacity allocated to path $q \in Q_{\mathcal{R}1}$ for flow r, and h_{r1} be the transport network bandwidth requirements of function c_{r1}. h_{r1} can be directly estimated using the analysis [22]. Equation (3) should be minimized subject to a set of network and processing demand constraints described through the following set of equations:

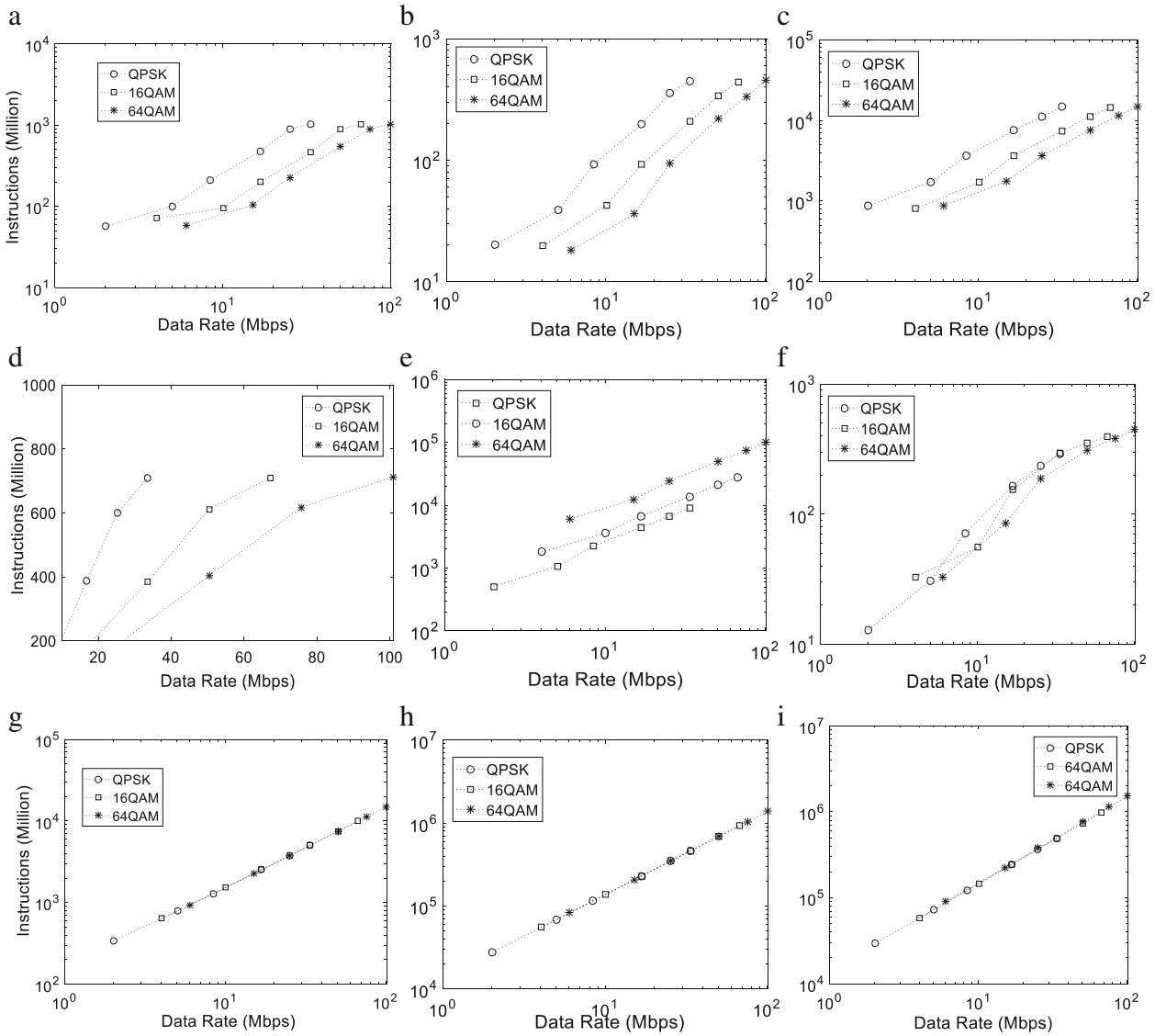

Fig. 2 Instructions per signal processing function under various data rates for **a** SC-FDMA demodulation, **b** subcarrier demapper, **c** equalizer, **d** transform decoder, **e** demodulation, **f** descrambler, **g** rate matcher, **h** turbo decoder, and **i** total instructions

$$\sum\nolimits_{s\in S} x_{rs1} = 1, \quad \forall r \in \mathcal{R} \tag{4}$$

$$\sum\nolimits_{r\in\mathcal{R}} x_{rs1} c_{r1} \le C_{s1}, \quad \forall s \in \mathcal{S} \tag{5}$$

$$\sum\nolimits_{s\in S} \sum\nolimits_{q\in Q_{\mathcal{R}1}} x_{rs1} z_{rq} = h_{r1}, \forall r \in \mathcal{R} \tag{6}$$

$$\sum\nolimits_{r\in\mathcal{R}} \sum\nolimits_{s\in S} \sum\nolimits_{q\in Q_{\mathcal{R}1}} \delta_{rqe} z_{rq} \le C_{e1}, \forall e \in E \tag{7}$$

Constraint (4) limits the number of servers where c_{r1} type of functions can be processed to one, and Eq. (5) indicates that the total number of tasks that can be assigned to server s, $s \in S$ cannot exceed its available processing capacity C_{s1} at stage 1, while Eqs. (6) and (7) introduce network demand and capacity constraints, respectively. In Eq. (7), δ_{rqe} is a binary coefficient taking values equal to 1 if e belongs to path $q \in Q_{R1}$ realizing demand c_{r1} at server s and C_{e1} is the available capacity of network link e at stage 1. After the solution of the first stage optimization problem, the remaining server and network capacity that can be used for the subsequent functions in the chain will be equal to:

$$C_{s1} - \sum\nolimits_{r\in R} x_{rs1} c_{r1} = C_{s2} \tag{8.1}$$

$$C_{e1} - \sum\nolimits_{r\in\mathcal{R}} \sum\nolimits_{s\in S} \sum\nolimits_{q\in Q_{\mathcal{R}1}} \delta_{rqe} z_{rq} = C_{e2} \tag{8.2}$$

The decision variables x_{rs2} of the second stage optimization problem responsible to forward (through a set of candidate paths $q \in Q_{\mathcal{R}2}$) and allocate the second

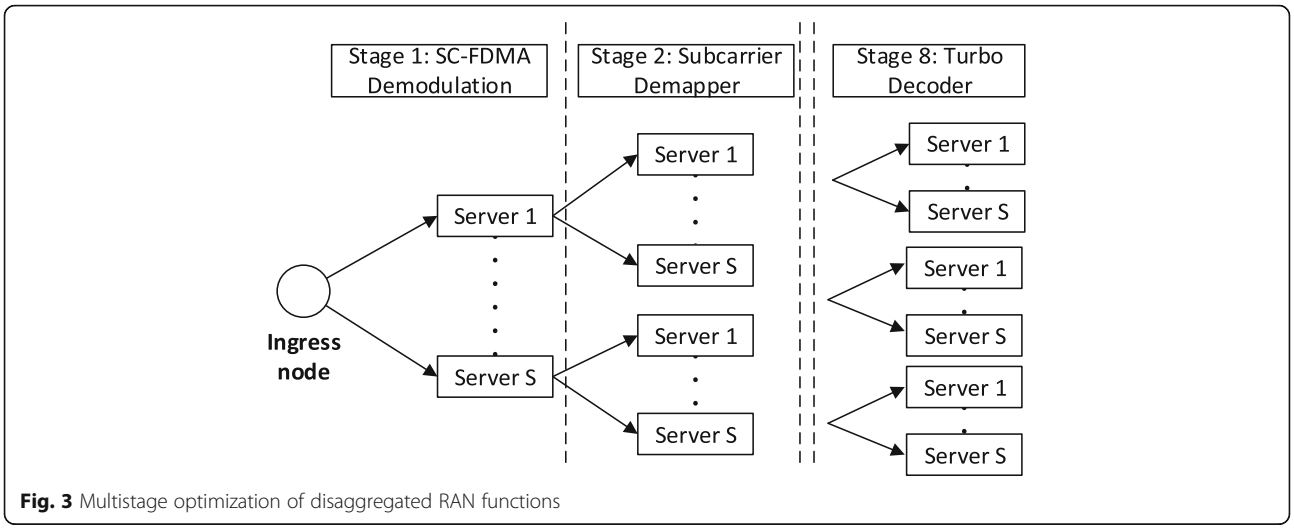

Fig. 3 Multistage optimization of disaggregated RAN functions

function of the FH service chain ($c_{r2}\equiv$ subcarrier demapper) to the optimal server s for processing depend on the results of the first stage problem. Typical example includes the lists of paths Q_{R2} that can be used to forward the output of the first function in the chain to the subsequent one (i.e., c_{r1} to c_{r2}). This set depends on the decisions taken by the first stage problem regarding the servers where c_{r1} functions can be placed. Other examples include the available capacity at the servers and network links. All this unknown information captured through data vectors ξ_t, $t=2,\ldots,8$, is revealed gradually as we proceed deeper in the processing of the service chain. The optimal compute resource assignment problem in disaggregated RAN environments can be solved through the minimization of the following nested cost function:

$$\min_{\mathbf{x}_1\in\mathcal{X}_1} f_1(\mathbf{x}_1)+\mathbb{E}\Big[\inf_{\mathbf{x}_2\in\mathcal{X}_{2(\mathbf{x}_1,\xi_2)}} f_2(\mathbf{x}_2,\xi_2)+\mathbb{E}\big[\cdots + \mathbb{E}\big[\inf_{\mathbf{x}_8\in\mathcal{X}_{8(\mathbf{x}_7,\xi_8)}} f_8(\mathbf{x}_8,\xi_8)\big]\big]\Big] \tag{9}$$

where extending Eq. (3) $f_t(\mathbf{x}_t,\xi_t)=\sum_{s\in S}\mathcal{E}_s\big(\sum_{r\in R}x_{rst}c_{rt}\big)$, $\xi_t=(C_{st},C_{et},h_{rt},z_{rq},)$ and based on Eqs. (4)–(8) $\mathcal{X}_t$, $t=2,\ldots,8$, can be described through the following constrains:

$$\mathcal{X}_t:=\Big\{\mathbf{x}_t:\sum\nolimits_{s\in S}x_{rst}=1, \tag{10.1}$$

$$\sum\nolimits_{r\in R}x_{rst}c_{rt}\le C_{st}, \tag{10.2}$$

$$\sum\nolimits_{s\in S}\sum\nolimits_{q\in Q_{Rt}}x_{rst}z_{rq}=h_{rt}, \tag{10.3}$$

$$\sum\nolimits_{r\in\mathcal{R}}\sum\nolimits_{s\in S}\sum\nolimits_{q\in Q_{Rt}}\delta_{rqe}z_{rq}\le C_{et}, \tag{10.4}$$

$$C_{st-1}-\sum\nolimits_{r\in R}x_{rst-1}c_{rt-1}=C_{st} \tag{10.5}$$

$$C_{et-1}-\sum\nolimits_{r\in\mathcal{R}}\sum\nolimits_{s\in S}\sum\nolimits_{q\in Q_{Rt-1}}\delta_{rqe}z_{rq}=C_{et}\Big\} \tag{10.6}$$

The multistage linear programming model Eqs. (3)–(9) can be decomposed into simpler subproblems using the duality theory [23]. After relaxing constraint (10.5), the Lagrangian function of Eq. (9) at stage t can be written in the following form:

$$L_t(\mathbf{x}_t,\xi_t)=f_t(\mathbf{x}_t,\xi_t)+\mathcal{Q}_{t+1}(\mathbf{x}_t,\xi_t)+\sum_{s=1}^{S}\pi_{st}^{\mathrm{T}}\Big(C_{st-1}-\sum\nolimits_{r\in R}x_{rst-1}c_{rt-1}-C_{st}\Big) \tag{11.1}$$

where

$$\mathcal{Q}_{t+1}(\mathbf{x}_t,\xi_t):=\mathbb{E}[Q_{t+1}(\mathbf{x}_t,\xi_{t+1})|\xi_t] \tag{11.2}$$

with

$$Q_t(\mathbf{x}_{t-1},\xi_t)=\inf_{\mathbf{x}_2\in\mathcal{X}_{2(\mathbf{x}_1,\xi_2)}}\Big[f_t(\mathbf{x}_t,\xi_t)+\mathcal{Q}_{t+1}(\mathbf{x}_t,\xi_t):C_{st-1}-\sum\nolimits_{r\in R}x_{rst-1}c_{rt-1}=C_{st}\Big] \tag{11.3}$$

The dual function is

$$\begin{aligned}\mathcal{D}_t(\pi_{st})&=\inf_{x_t}L_t(\mathbf{x}_t,\xi_t) \qquad (12)\\ &=-\sup_{x_t}\Big\{\sum_{s=1}^{S}\pi_{st}^{\mathrm{T}}C_{st}-\big[f_t(\mathbf{x}_t,\xi_t)+\mathcal{Q}_{t+1}(\mathbf{x}_t,\xi_t)\big]\Big\}\\ &\quad+\sum_{s=1}^{S}\pi_{st}^{\mathrm{T}}\Big(C_{st-1}-\sum\nolimits_{r\in R}x_{rst-1}c_{rt-1}\Big)\end{aligned}$$

and the dual problem can be stated as

$$\max_{\pi_{st}}\mathcal{D}_t(\pi_{st}) \tag{13}$$

Subject to Eqs. (10.1)–(10.4) and Eq. (10.6).

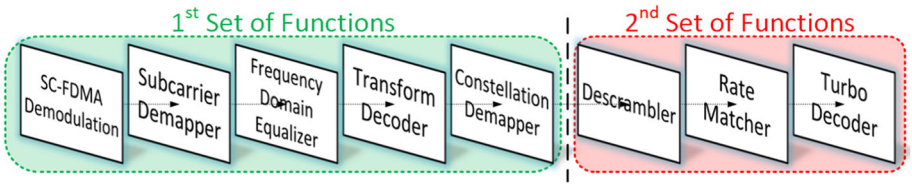

Fig. 4 Reduction of complexity through the grouping of BBU functions

Given that Eqs. (9) and (13) are convex problems, the duality gap between the original and the relaxed problem is zero [23].

3.3 Heuristic for optimal BBU placement description

Although the ILP model discussed in the previous subsection can be effectively used to identify the optimal placement of BBU functions within the data center, it suffers high computational complexity, thus making it unsuitable for real-time system deployments. To address this issue, a heuristic algorithm with low computational complexity is proposed that tries to identify the optimal compute resources required to support the most energy efficient processing of the BBU service chain within the data center.

To limit the complexity of the heuristic, we have defined 2 sets of functions (1st and 2nd set of functions) to which the 8 different BBU functions can be mapped (Fig. 4). To satisfy the requirements of the BBU service chain, the order of the various functions is always maintained within and across the 2 sets of functions defined. The first set comprises SC-FDMA, subcarrier demapper, frequency domain equalizer, transform decoder, and constellation demapper functions, while the second includes descramble, rate matcher, and turbo decoder functions. As shown in [24], the proposed grouping policy has been selected as it requires a relatively small amount of network resources for the interconnection of the first with the second set of functions while the computational requirements of the 2nd set are still very high.

The main objective of the heuristic is to allocate an input BBU service chain to the most energy efficient servers that have sufficient capacity to process it. The input service can be split and allocated to a set of servers,

```
for each RU:
    if a server that is already used can process it in time:
        assign it to that server
        update that server's capacity
    else:
        assign it to the most efficient server type (search between types 1-3) which can process it
        update that server's capacity
        if only the least energy efficient server type can process it:
            check the thresholds to decide if and which split option should be enabled.
            if split is enabled:
                create the 2 sets of functions
                find the closest appropriate servers (Dijkstra Algorithm)
                assign the sets to the appropriate servers.
                update the servers' capacity
            else:
                assign it to a least energy efficient server
                update the server's capacity
```

Fig. 5 Heuristic for BBU assignment problem

Table 1 Technical specifications of the servers used in the numerical evaluations

Server type	Computer/device	Servers	Chips	Cores	Threads	GOPS	Power (watt)	GOPS/watt	Idle (watt)
S0	SuperMicro X11DPi-N(T) SMC X11	2× Intel Xeon Platinum 8160	2	48	96	1071.37	360	2.98	53.40
S1	SuperMicro X11DPG-QT	2× Intel Xeon Gold 6140	2	36	72	888.52	336	2.64	52.40
S2	SuperMicro X10Dai SMC X10	2× Intel Xeon E5–2683 v4	2	32	64	700.94	288	2.43	81.00
S3	Sugon I908-G20	8× Intel Xeon E7–8860 v3	8	128	256	2510.56	1344	1.87	269.00

in case that splitting the service across servers is a more energy efficient option. A more detailed description regarding the server allocation process is provided in Fig. 5.

In our analysis, we were aiming at always serving the input traffic, independent of the volume of incoming data to be processed, satisfying at the same time, the time constraints associated with the service. Therefore, we are considering the ratio of the number of instructions required for the 2nd set of functions to be performed, over the number of instructions of the 1st set of functions. As it was also the case for the ILP analysis, four different types of servers randomly placed inside the DC racks are considered. These servers can be classified according to their energy efficiency, with type 1 server to be the most energy efficient, while type 4 the least energy efficient server. The technical specifications of which are provided in Table 1. Considering this assumption, we calculate the ratios of the capacity of the larger type of server (type 4 server, least energy efficient) over the capacities of the rest of the servers (type 1, type 2, and type 3). In addition, based on these ratios and the time constraints associated with the service (in total < 1 ms per subframe), we were able to define a set of thresholds that can be used to identify whether an incoming service can be split between the larger type server and any other of the smaller available type of servers.

For the specific functional split and the set of servers considered in this study, the numerical values of the thresholds we have identified are as follows: (a) 68% of type 4 server processing capacity if the 1st set of functions is allocated to server type 1, (b) 69% of type 4 server processing capacity if the 1st set of functions is allocated to server type 2, and (c) 70% of type 4 server processing capacity if the 1st set of functions is allocated to server type 3. It should be noted that in our calculations, suitable processing margins of the order of 2% have been allowed.

4 Results and discussion

4.1 Evaluation scenario

To quantify the benefits of the proposed softwarized RAN approach, the simple DC network topology of Fig. 1 is considered. This topology comprises 6 racks, each one packed with 48 servers. Connectivity between racks is provided with the switching solution provided in [25]. In the numerical calculations, we consider four types of servers, randomly placed inside the racks. The technical specifications of these servers are provided in Table 1, while their power consumption follows the linear stepwise function described in [26]. For the wireless access, we consider the topology described in [1] in which the served area is covered by a set of RRHs which forward their FH flows to the DCs for processing. Given that this study focuses on the computational aspects of the FH flows, we make the rational assumption that the transport network does not act as a bottleneck and, at the same time, it has sufficient capacity to transfer all flows to the DCs for processing.

4.2 ILP numerical results

Figure 6 compares the performance of the proposed optimization scheme (denoted DSW-BBU) in terms of power consumption with the traditional SW-BBU as a function of the served traffic. As expected, the power consumption at the DCs increases with the wireless access load. However, the DSW-BBU offers much better performance due to its increased ability to mix and match compute and network resources, leading to improved utilization of the servers and to higher energy efficiency. Specifically, numerical results show that when using the proposed approach, the overall system power consumption is reduced by 40% under high loading scenarios (from 50 to 30 kW)

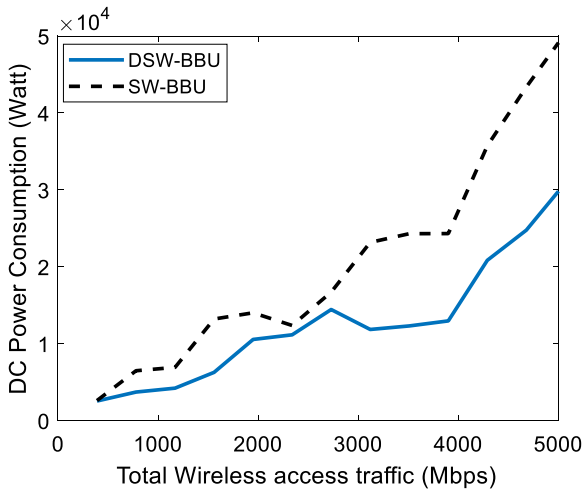

Fig. 6 DC power consumption for the traditional softwarised BBU (SW-BBB) and the disaggregated SW-BBU (DSW-BBU) as a function of the total wireless access traffic

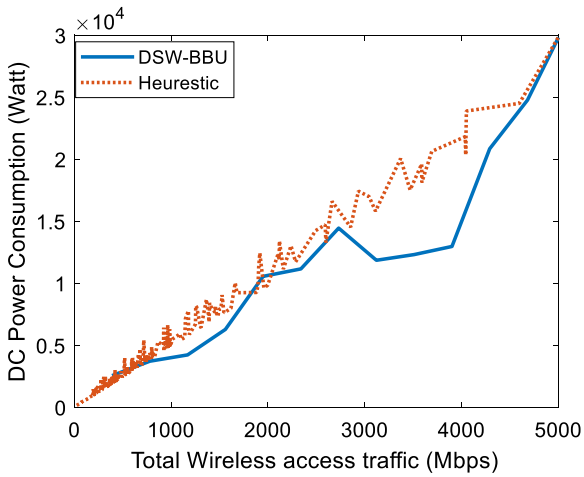

Fig. 7 Comparison between the ILP model and the heuristic algorithm

compared to the traditional solution where all BBU functions are hosted in the same physical servers.

4.3 Heuristic for optimal BBU placement numerical results

To evaluate the performance of the proposed heuristic, we have calculated the total power consumption of the compute resources required to support the set of services assumed in Section 4.1 and compare it with the results produced through the ILP approach also described in the previous subsection.

The total power consumption of the compute resources as a function of the total wireless access traffic load is shown in Fig. 7 for both the heuristic and ILP approaches. Relatively good agreement between the heuristic and ILP approaches is achieved particularly for low and high loading conditions. In these areas, both schemes allocate the BBU functions in the same machines, and therefore, the optimality gap is less than 5%.

In Fig. 7, fluctuations of the DC power consumption values calculated by the heuristic can be observed. This is due to that the heuristic approach taken based on the greedy approach introduces resource fragmentation depending on the input traffic statistics and sequence, which

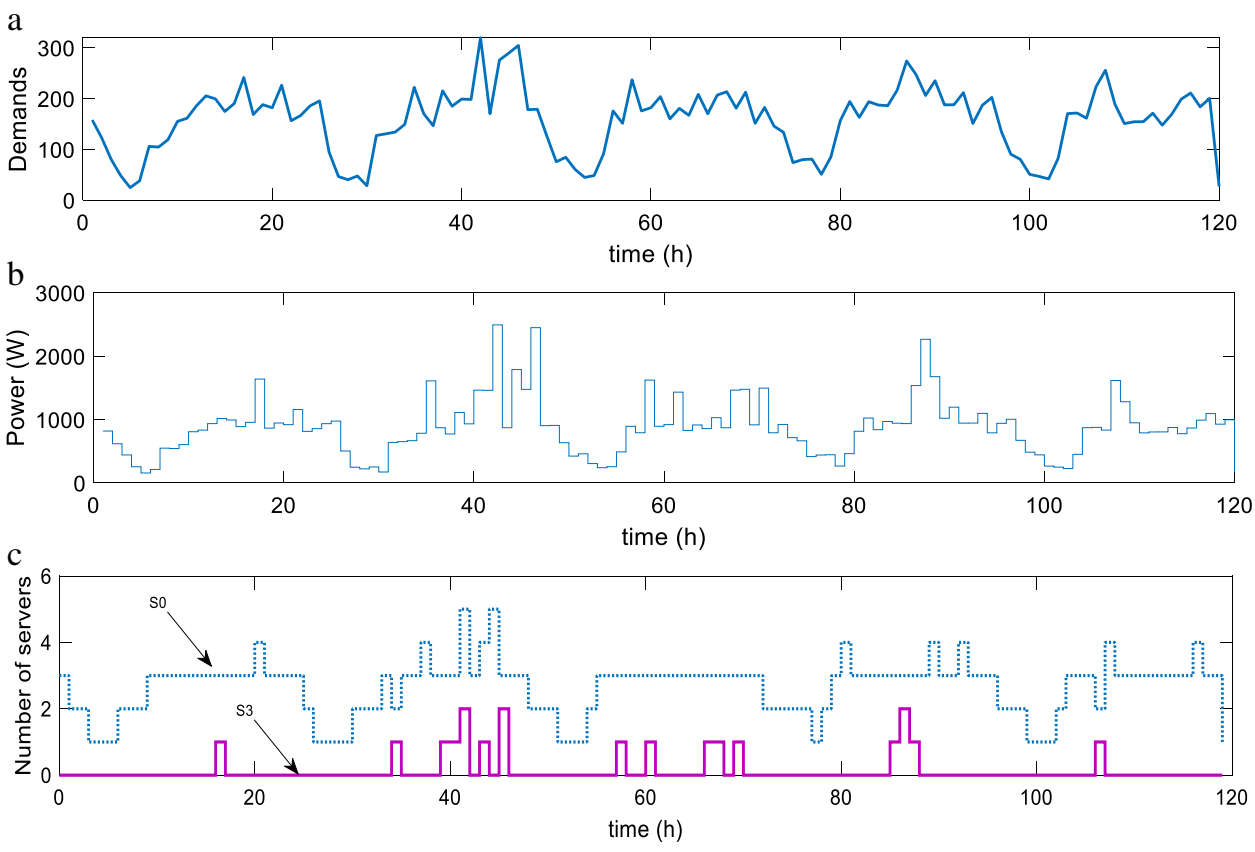

Fig. 8 **a** Wireless access demands as a function of time. **b** Evolution of power consumption over time. **c** Number of active servers supporting BBU processing

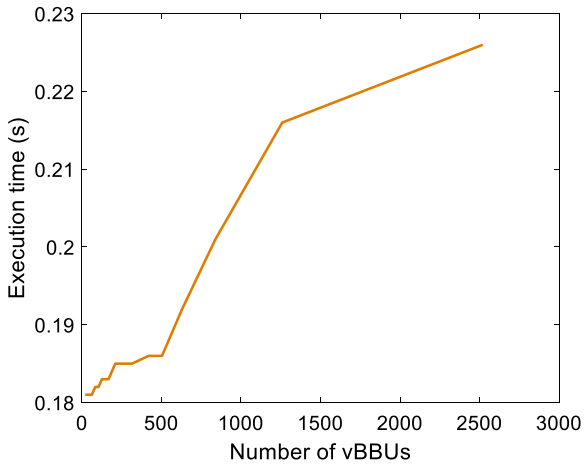

Fig. 9 Processing time for the heuristic scheme as a function of the number of vBBUs

is not an issue in the case of ILP identifying globally optimal solutions for all traffic demands at once.

The impact of traffic variation on the number and type of servers that are employed to accommodate the relevant BBU requirements is presented in Fig. 8a–c. Specifically, the evolution of the wireless access demands as a function of time for the measurements described in [27] is shown in Fig. 8a. These wireless access requirements are then mapped to BBU processing requirements using the analysis presented in Section 3.1. Once the processing requirements have been determined, the BBU SCs are constructed and mapped to the suitable server. A snapshot of the overall DC power consumption as a function of time for the traffic load shown in Fig. 8a is provided in Fig. 8b. As expected, under peak hours, the DC power consumption increases due to the activation of additional servers that are necessary for the processing of the very high BBU demands, whereas under off-peak hours, the power consumption is minimized. This is also verified in Fig. 8c where the evolution of the number of active servers supporting BBU processing as a function of time is illustrated. An interesting observation is that under high loading scenarios, not only additional DC servers of the same type are activated ("S0") but also servers which are less efficient ("S3").

Finally, the execution time of the heuristic scheme as a function of the number of BBU SCs that need to be constructed is shown in Fig. 9. We observe the execution time ranges from 0.18 s when a small number of SC is created up to 0.23 s under high traffic scenarios.

5 Conclusions

This paper focused on the concept of compute resource disaggregation in centralized softwarised RANs to allow the individual allocation of processing functions to different servers depending on the nature and volume of their processing requirements. Our experimental results have shown that from the whole BBU service chain, the turbo decoder, constellation demapper, rate matcher, and frequency domain equalizer functions introduce much higher demands, at the LTE PHY uplink, in terms of computational resources (approximately 90% for very high data rates). For these functions, the number of instructions executed has a linear dependence of the data rate and by extending the whole LTE PHY uplink presents the same dependence from the data rate. Since the baseband processing workload can be split into several functions, the overall system performance can be optimized, in our case in terms of overall power consumption by appropriate allocation of each function to a suitable server. This was performed using a purposely developed multistage ILP optimization framework, in the scenario of a heterogeneous DC, which can lead to better utilization of the servers and to higher energy efficiency. Specifically, the overall system power consumption is reduced by 40% under high loading scenarios compared to the traditional solution where all BBU functions are hosted in the same physical servers. A heuristic was also developed to address the computation complexity associated with the ILP approach, demonstrating good performance when compared with the ILP results (the gap was less than 5%), particularly for low and high loading conditions.

Abbreviations

BBU: Baseband unit; BS: Base station; CP: Cyclic prefix; CPRI: Common Public Radio Interface; C-RAN: Cloud radio access network; CSI: Channel state information; CU: Central unit; DC: Data center; DCC: Decentralized Cloud Controller; DSW-BBU: Disaggregated SW-BBU; FFT: Fast Fourier transform; FH: Fronthaul; GPP: General purpose processor; IFFT: Inverse fast Fourier transform; ILP: Integer linear programming; IPS: Instructions per second; LLR: Logarithmic likelihood ratios; MSS-BBU: Multiside/standard baseband unit; OFDM: Orthogonal frequency diversity multiplexing; QoS: Quality of service; RAN: Radio access network; RRH: Remote radio head; RU: Radio unit; SC: Service chain; SC-FDMA: Single carrier-frequency division multiple access; SISO: Soft-input soft-output; SW-BBU: Softwarized BBU; UP: User processing

Authors' contributions

The authors have contributed jointly to the manuscript. All authors have read and approved the final manuscript.

Funding

This work has been financially supported by the EU Horizon 2020 project 5G-PICTURE under grant agreement No 762057, the EU Horizon 2020 project SmartNet under grant agreement No 691405 and the EU Horizon 2020 project IN2DREAMS under grant agreement No 777596.

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]National and Kapodistrian University of Athens, Athens, Greece. [2]HPN Group, University of Bristol, Bristol, UK.

Received: 23 November 2018 Accepted: 7 June 2019
Published online: 18 July 2019

References

1. A. Tzanakaki et al., in *IEEE Communications Magazine, vol. 55, no. 10.* Wireless-optical network convergence: enabling the 5G architecture to support operational and end-user services (2017), pp. 184–192
2. N. Gkatzios, M. Anastasopoulos, A. Tzanakaki, D. Simeonidou, in *2018 European Conference on Networks and Communications (EuCNC), Ljubljana, Slovenia.* Compute resource disaggregation: an enabler for efficient 5G RAN softwarisation (2018), pp. 1–5
3. F. Musumeci, C. Bellanzon, N. Carapellese, M. Tornatore, A. Pattavina, S. Gosselin, Optimal BBU placement for 5G C-RAN deployment over WDM aggregation networks. J. Lightwave Technol. **34**(8), 1963–1970 (2016)
4. C. Colman-Meixner, G.B. Figueiredo, M. Fiorani, M. Tornatore, B. Mukherjee, in *2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Bangalore.* Resilient cloud network mapping with virtualized BBU placement for cloud-RAN (2016), pp. 1–3
5. A. Checko et al., "Cloud RAN for mobile networks—a technology overview," in IEEE Communications Surveys & Tutorials. 17. 1. 405–426, 2015
6. M. Fiorani, B. Skubic, J. Mårtensson, L. Valcarenghi, P. Castoldi, L. Wosinska, P. Monti, On the design of 5G transport networks. Photon Netw. Commun. **30**(3), 403–415 (2015)
7. C. Ranaweera et al., in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Honolulu, HI.* Optical X-haul options for 5G fixed wireless access: which one to choose? (2018), pp. 1–2
8. M. Ruffini, F. Slyne, Moving the network to the cloud: the cloud central office revolution and its implications for the optical layer. J. Lightwave Technol. **37**(7), 1706–1716 (2019)
9. D. Hisano, Y. Nakayama, K. Maruta, A. Maruta, in *2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates.* Deployment design of functional split base station in fixed and wireless multihop fronthaul (2018), pp. 1–6
10. C. Song, M. Zhang, Y. Zhan, D. Wang, L. Guan, W. Liu, L. Zhang, S. Xu, Hierarchical edge cloud enabling network slicing for 5G optical fronthaul. J. Opt. Commun. Netw. **11**, B60–B70 (2019)
11. D. Harutyunyan, R. Riggio, in *Proceedings of the 10th IFIP WG 6.6 International Conference on Management and Security in the Age of Hyperconnectivity.* Functional decomposition in 5G networks, vol 9701 (Springer-Verlag New York, Inc., New York, 2016), pp. 62–67
12. J. Duan, X. Lagrange, F. Guilloud, in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), Nanjing.* Performance analysis of several functional splits in C-RAN (2016), pp. 1–5
13. C. Desset et al., in *2012 IEEE Wireless Communications and Networking Conference (WCNC), Shanghai.* Flexible power modeling of LTE base stations (2012), pp. 2858–2862
14. Q. Zheng et al., in *2013 IEEE International Symposium on Workload Characterization (IISWC), Portland, OR.* WiBench: an open source kernel suite for benchmarking wireless systems (2013), pp. 123–132
15. [Online] WiBench, An Open Source Kernel Suite for Benchmarking Wireless Systems. http://wibench.eecs.umich.edu/about.html. Accessed 22 May 2019
16. S. Bhaumik et al., in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking (Mobicom '12).* CloudIQ: a framework for processing base stations in a data center (ACM, New York, 2012), pp. 125–136
17. [Online], OpenAirInterface, https://gitlab.eurecom.fr/oai/openairinterface5g/wikis/OpenAirUsage. Accessed 22 May 2019
18. [Online] Software Radio Systems (SRS). srsLTE, https://github.com/srsLTE/srsLTE. Accessed 22 May 2019
19. B. Haberland et al., Radio base stations in the cloud. Bell Labs Tech. J. **18**(1), 129–152 (2013)
20. R. Riggio, D. Harutyunyan, A. Bradai, S. Kuklinski, T. Ahmed, in *2016 12th International Conference on Network and Service Management (CNSM), Montreal, QC.* SWAN: base-band units placement over reconfigurable wireless front-hauls (2016), pp. 28–36
21. [Online], Vtune Amplifier, https://software.intel.com/en-us/vtune. Accessed 22 May 2019
22. U. Dötsch, M. Doll, H. Mayer, F. Schaich, J. Segel, P. Sehier, Quantitative analysis of split base station processing and determination of advantageous architectures for LTE. Bell Labs Tech. J. **18**(1), 105–128 (2013)
23. A. Shapiro, D. Dentcheva, A. Ruszczyński, in *MOS-SIAM Series on Optimization, ISBN: 978-0-89871-687-0.* Lectures on stochastic programming: modeling and theory (2009)
24. A. Tzanakaki, M. Anastasopoulos, D. Simeonidou, in *2018 in Optical Fiber Communication Conference (OFC), San Diego, CA.* Converged access/metro infastructures for 5G services (2018), pp. 1–3
25. J. Perelló et al., All-optical packet/circuit switching-based data center network for enhanced scalability, latency, and throughput. IEEE Netw. **27**(6), 14–22 (2013)
26. A. Tzanakaki et al., in *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Shanghai.* Energy efficiency in integrated IT and optical network infrastructures: the GEYSERS approach (2011), pp. 343–348
27. X. Chen, Y. Jin, S. Qiang, W. Hu, K. Jiang, in *IEEE Int. Conf. on Communications Workshops (ICC).* Analyzing and modeling spatio-temporal dependence of cellular traffic at city scale (2015), pp. 3585–3591

6 Publisher's Note