An Efficient QoS-Aware Computational Resource Allocation Scheme in C-RAN

Mojgan Barahman and Luis M. Correia
INESC-ID / INOV / IST, University of Lisbon
Lisbon, Portugal
{mojgan.barahman, luis.m.correia}@tecnico.ulisboa.pt

Lúcio S. Ferreira
Universidade Lusiada de Lisboa / INESC-ID
Lisbon, Portugal
lucio.ferreira@inov.pt

Abstract — **In this paper, one proposes an approach to optimize the computational resource utilization of baseband unit pools in a Cloud Radio Access Network. The problem of resource allocation is formulated and solved as a constrained nonlinear optimization one, based on the concept of bargaining in cooperative game theory. The goal is to minimize resource usage by on-demand resource allocation, per instantaneous requirements of base stations, whilst taking Quality of Service into account. In the event of a shortage of resources, implying that not all demand can be served at the same time, baseband units are prioritized with a weighting policy. Real-time requirements and the priority of services being run on a baseband unit are the two contributors in calculating the weight in a timeslot. Lower prior baseband units, however, are always guaranteed to receive a minimum of resources to prevent them from crashes. Simulation results in a heterogeneous services environment show a minimum 83% improvement in allocation efficiency, compared to a fixed resource allocation scheme based on peak-hour traffic demands. Results also confirm that, in case of a resource shortage, 100% of the resources are fairly distributed among baseband units, fairness being governed by the weight of the baseband units in the pool.**

Keywords - Wireless Communications, Cloud-RAN, Computational Resource Utilization, Resource Allocation Efficiency, Nonlinear Optimization.

I. Introduction

The proliferation of high data rate applications in conjunction with high mobile terminals usage nowadays has triggered a drastic increase in data rate demands [1]. Therefore, wireless network providers must continuously improve their infrastructure to serve data demand accordingly. The challenge is even more difficult, since resource allocation in conventional Radio Access Networks (RANs) is inefficient, as it is based on peak-hour traffic requirements. However, since users demand is time varying, traffic is not always in the peak level, and may be up to 10 times lower in off-peak hours [2]; thus, a fixed allocation scheme leaves idle resources in various times/areas.

Cloud-RAN (C-RAN) has emerged as a centralized paradigm to provide a solution for higher data rates and capacity demands in a cost-efficient way [2]. C-RAN splits a traditional Base Station (BS) into the BaseBand Unit (BBU) and the Radio Remote Head (RRH), which are connected by a high-speed link. It enables the central management of computational resource, by integrating the BBU resources of multiple BSs in a data centre, designated as a BBU-pool. Centralizing BSs with different peak-hour traffic (e.g., by mixing residential and business regions) in the same BBU-pool balances resource usage, as the resources of under-loaded BSs at a given time instant can be shared with overloaded ones. It is then expected to have lower peak resource requirements in a pool than the sum of peak requirements of individual BSs.

However, to take advantage of C-RAN benefits, efficient strategies should be applied in order to distribute the resources of the BBU-pool among BBUs. An efficient resource provisioning scheme should minimize not only the resource idle times but also the BBUs' overloading. There are a number of works in the literature proposing a resource management strategy in BBU-pools aiming at optimizing computational resource usage.

The authors in [3], [4], [5], [6] propose models aiming to counterbalance the load among the BBUs in a BBU-pool, by migrating the load from overloaded BBUs to the under-loaded ones in a way that the load will be balanced ultimately. Al-Dulaimi et al. [7] took one-step further, based on a graph-colouring model, proposing a model to optimize the utilization of C-RAN resource by turning off the BBUs with low traffic and diverting their load to the neighbouring under-loaded BBUs. W. Chien et al. [8] propose a similar strategy to minimize the number of used BBU-pools, suggesting a model to optimize C-RAN resource utilization by assigning RRHs to BBU-pools appropriately, so that the number of used BBU-pools is decreased; their model predicts the traffic of an individual RRH in the network and then exploits a genetic algorithm to allocate the BBUs.

The authors in [9] and [10] propose a different strategy for optimizing resource utilization in C-RAN. Instead of load migration, their approach reconfigures the computational capability of the BBUs according to BBUs' instantaneous requirements elastically. In [9], a demand-aware resource provisioning scheme is proposed for the BBU-pool. Resource sharing is achieved by virtualization techniques, where BSs functionalities are implemented on Virtual Machines (VMs) on top of general-purpose servers. Their model predicts the amount of required resources based on a given pattern and changes the VMs' computational capacity accordingly; however, no management strategy is proposed in case of a resource shortage, when all of the BBUs' demand cannot be served at the same time. Y. Liao et al. [10] show the relationship between the required computational resource of the BBUs and the signal that is transmitted to users. In the case of the Coordinated MultiPoint (CoMP) deployment, they minimize the computational resource usage by reducing the number of RRHs that serve a user at a time; meanwhile, Quality of Service (QoS) constraints are taken into account.

Load migration strategies impose extra burden to the network, due to the increased data exchanges among BBUs. The data transmission cost is even higher in the case of small cells in dense areas since CoMP, handover and interference occur more often [11]. The more efficient resource allocation

978-1-7281-3106-1/20/$31.00 ©2020 IEEE

approach here is to assign the same BBU to the coordinated RRHs for CoMP, then, reconfiguring the BBU computing resource elastically, based on the instantaneous requirement of the RRHs traffic [12]. As a result, the focus in this work is on the latter strategy for optimizing the computational resource utilization of the BBU-pool, i.e. resizing BBUs' computing capability, rather than the load migration.

Unlike previous works, the resource allocation scheme proposed in this paper considers the computing resource constraints in the BBU-pool. The proposed model allocates the computational resource of the BBU-pool in case of resource shortage (when not all the BBUs' demand can be served at the same time) by prioritizing BBUs based on the type of services that are active and QoS limitations. QoS constraint and type of active services are observed as real-time parameters in the process of resource provisioning, which is of great significance for upcoming service-based wireless technologies, namely, 5G. Moreover, the model guarantees that minimum computational resource is allocated to an individual BBU in order to prevent it from crashes.

To this end, the priority based computational resource allocation is formulated as a cooperative bargaining game. Players, i.e. BBUs, choose strategies to maximize their own computational resource under a consensus-based decision-making process. The twofold solution not only maximizes processing speed for an individual BBU, but also maximizes BBU-pool efficiency with respect to the priority of services.

The remainder of this article is organized as follows. The next section provides the main assumptions and the selected architecture. Section III discusses the model for estimating the amount of Required Computational Capacity (RCC) in the BBU followed by an approach for reformulating the computational resource allocation as a centralized optimization problem. Section IV describes the performance metrics. Section V presents a case study followed by the analyses of the results in Section VI. Conclusions are drawn in Section VII.

II. Network Architecture and Assumptions

Fig.1 presents the general C-RAN architecture, with the BS split into the BBU and the RRH. The BBUs of several BSs are co-located in a central place to form a BBU-pool. BBUs are connected via high-speed links to their associated RRH. Several BBU-pools are connected to each other at a higher level, through high-speed optical links.

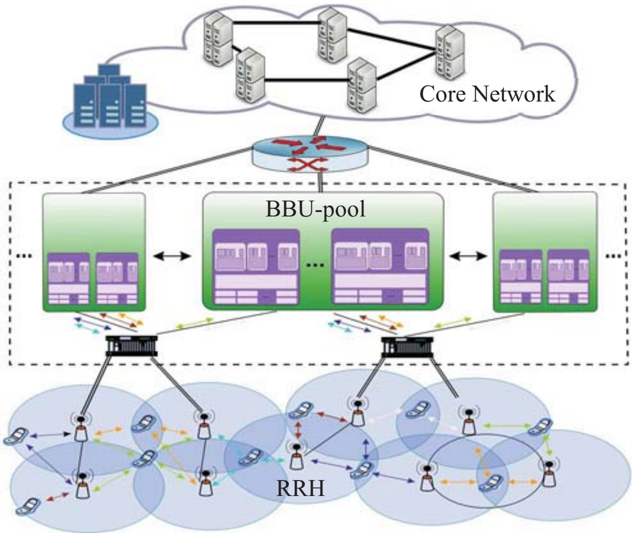

Fig. 1. C–RAN architecture.

Although a BBU can serve multiple RRHs [2], for the sake of simplicity, however, it is considered that there is a one to one connection in between the BBU and the RRH of a BS.

The BS functionalities are split in between the RRH and the BBU, amongst which, channel de/coding, de/modulation, MIMO de/pre-coding, channel estimation, OFDMA and SC-FDMA processing are the main ones that are considered to be processed in the BBU.

A BBU-pool is composed of several general-purpose computing servers, where the computing resource of multiple BSs are deployed on them. The baseband resources are shared and are flexibly allocated to the BBUs in the pool by the aid of virtualization [13]. BBU instances are hosted in VMs over servers, baseband functionalities being executed as software applications. The amount of computational and storage resource that are allocated to an individual VM is managed by a hypervisor. Resource provisioning is dynamic and demand-aware, meaning that resources are allocated and released to the VMs elastically, based on the instantaneous requirement of BBU instances. In case a BBU does not have enough computational resource to meet user-processing needs, it will be provisioned with more computing resources.

III. Analytical Model

A. Resource Demand Estimation

The proposed computational resource allocator assigns resources to BBUs according to their instantaneous RCC. The RCC of a BBU is considered as the amount of computational capacity that it requires in such a way that no computational delay is imposed on signal processing. The estimation of the amount of RCC of a BBU is based on [14] and [15]. Taking some well-defined operating scenarios, the RCC per information bit transmission is estimated by counting the number of mathematical operations that each processing step performs. The achieved values are then scaled for any desired scenario, considering variations of system/user parameters that affect the result, i.e. bandwidth, number of antennas, etc. Both estimated RCC values and the scaling rules have been interpreted from various sources either from scientific research or from empirical experiments [14].

As it is described in [16], in order to be able to apply parallel processing, leading to shortened execution times, the processing of the BBU is split per user, i.e., User Process (UP), each being processed independently and in parallel to the others, where possible. A UP contains the processing steps of a BBU that can be split per user, e.g. channel de/coding and de/modulation. Common Processing (CP), however, cannot be split per user, as it does the processing of a signal resulting from the combination of all users' signals, e.g. OFDMA and SC-FDMA. The total RCC of BBU b at a time instant t_k is composed of the total computational capacity that is required for CP and all of the UPs, given by

$$\begin{aligned} C^R_{b,t_k[\text{GOPS}]} &= \sum_{i=1}^{N^{CP}} C^R_{C\,b,P_i,t_k[\text{GOPS}]} \\ &\quad + \sum_{u=1}^{N^U_{b,t_k}} \sum_{j=1}^{N^{UP}} C^R_{U\,u,P_j,t_k[\text{GOPS}]} \end{aligned} \tag{1}$$

where:

- N^{CP}: number of CP steps,
- $C^R_{C\,b,P_i,t_k}$: RCC of BBU b for CP step i, P_i, at t_k, (measured in Giga Operations Per Second),
- N^U_{b,t_k}: number of active users in BBU b at t_k,
- N^{UP}: number of UP steps,
- $C^R_{U\,u,P_j,t_k}$: RCC of user u for UP step j, P_j, at t_k.

In case that several users are active at any given time instant, the computational capacity that is required for the CP should be met; otherwise, none of the user's data can be transferred. Therefore, a minimum computational capacity should be guaranteed, given by

$$C^{R_{min}}_{b,t_k[\mathrm{GOPS}]} = C^{R}_{C\,b,t_k[\mathrm{GOPS}]}\,, \qquad b = 1,2,\dots,N_B \tag{2}$$

where:

- $C^{R}_{C\,b,t_k}$: amount of RCC for CP in BBU b at t_k,
- N_B: number of BBUs in the BBU-pool.

B. Game Formulation

The problem of computational resource allocation inside a BBU-pool is formulated as a bargaining game in cooperative game theory [17], that is described as follows:

- Players are N_B BBUs, strategically competing against each other for computational resource.
- A strategy, which is also a feasible solution, is a vector $\boldsymbol{C}^{\boldsymbol{Al}}_{t_k}$ given by

$$\boldsymbol{C}^{\boldsymbol{Al}}_{t_k} = [C^{Al}_{1,t_k[\mathrm{GOPS}]}, C^{Al}_{2,t_k[\mathrm{GOPS}]}, \dots, C^{Al}_{N_B,t_k[\mathrm{GOPS}]}]^{\mathrm{T}} \tag{3}$$

  where $C^{Al}_{b,t_k}$is the amount of Allocated Computational Capacity (AlCC) to BBU b at t_k.
- The utility of a player, i.e. BBU b at t_k, is given by function $\mathcal{U}_{b,t_k}: \mathbb{R}^{N_B} \to \mathbb{R}$, which maps the given strategy onto a value in [0, 1], such that

$$\mathcal{U}_{b,t_k}(\boldsymbol{C}^{\boldsymbol{Al}}_{t_k}) = C^{Al}_{b,t_k[\mathrm{GOPS}]} / C^{R}_{b,t_k[\mathrm{GOPS}]}, \quad b = 1,2,\dots,N_B \tag{4}$$

  The utility of a BBU reflects the portion of the BBU's demand that is served; hence, BBUs aim to maximize their own utility functions by joining the game.

However, two constraints should be considered in the process of resource allocation. The sum of the AlCC to the BBUs should not be more than the Available Computational Capacity (AvCC) of the BBU-pool. In addition, an individual BBU's AlCC should not exceed its RCC at a given time instant. A solution should obey both constraints, therefore, the feasible solution set, $S^{FS}_{t_k}$, of the defined resource management problem is bounded as

$$S^{FS}_{t_k} = \{\boldsymbol{C}^{\boldsymbol{Al}}_{t_k} \,|\, \textstyle\sum_{b=1}^{N_B} C^{Al}_{b,t_k[\mathrm{GOPS}]} \le C^{Av}_{BP\,t_k[\mathrm{GOPS}]}, C^{Al}_{b,t_k[\mathrm{GOPS}]} \le C^{R}_{b,t_k[\mathrm{GOPS}]}\}, \qquad b = 1,2,\dots,N_B \tag{5}$$

where $C^{Av}_{BP\,t_k}$ is the amount of the AvCC of BBU-pool at t_k.

On the other hand, in order to boost users' satisfaction and enhance resource allocation's fairness and efficiency, QoS limitations should be considered in the process of computational resource allocation. To this end, an individual BBU is assigned with a Bargaining Power (BP). In case of resource shortage that all BBUs demand cannot be served at the same time, BBUs BP express their priority with regard to the QoS constraint and real-time demand. The idea is that, after allocating the minimum computational capacity that should be guaranteed to BBUs to prevent crashes, i.e. $C^{R}_{C\,b,t_k}$, the rest of the resources are assigned to BBUs proportionally to their instantaneous requirements and the active services' priority level. In this regard, each type of services is allocated with a weight, which is derived from the priority level that 3GPP has assigned to an individual service [18], being normalized as

$$w^{srv}_{s} = 1 + 99\,\frac{P^{srv}_{max} - P^{srv}_{s}}{P^{srv}_{max} - P^{srv}_{min}} \tag{6}$$

where:

- P^{srv}_{max}: maximum value of 3GPP services' priority levels,
- P^{srv}_{s}: priority level that 3GPP assigns to the desired service s,
- P^{srv}_{min}: minimum value of 3GPP services priority levels.

The value of a service's weight ranges within [1, 100]. By normalizing the services' weight, the effect of service weights in the model is more controlled. Correspondingly, the average weight of the active services in BBU b at a given time instant t_k is achieved by

$$\overline{w^{srv}_{b,t_k}} = \left(\textstyle\sum_{s=1}^{N^{srv}} N^{U}_{b,s,t_k}\, w^{srv}_{s}\right) / N^{U}_{b,t_k} \tag{7}$$

where:

- N^{srv}: number of different types of services,
- N^{U}_{b,s,t_k}: number of users that are running service s in BBU b at time instant t_k.

Exploiting (1) and (7), the BBUs RCC and the services' weight are combined to construct the BBUs' BP. As a result, the BP of BBU b at time instant t_k, is

$$B_{b,t_k} = \frac{\overline{w^{srv}_{b,t_k}}\left(C^{R}_{b,t_k[\mathrm{GOPS}]} - C^{R}_{C\,b,t_k[\mathrm{GOPS}]}\right)}{\sum_{l=1}^{N_B}\left(\overline{w^{srv}_{l,t_k}}\left(C^{R}_{l,t_k[\mathrm{GOPS}]} - C^{R}_{C\,l,t_k[\mathrm{GOPS}]}\right)\right)} \tag{8}$$

The value of a BBU's BP is positive, built in a way that the sum of all BBUs' BPs is equal to one,

$$\textstyle\sum_{b=1}^{N_B} B_{b,t_k} = 1 \tag{9}$$

C. Nash Bargaining Solution

Since the computational resource management problem in the BBU-pool is formulated as a bargaining game, the Generalized Nash Bargaining Solution (GNBS) strategy [17] is exploited to find the optimum resource provisioning that maximizes the BBU-pool's computational resource utilization.

GNBS maximizes the product of the BBUs' utility functions weighted by the BBUs' BP. In case of 1) convexity of the utility function, $\mathcal{U}_{b,t_k}$, given by (4) and 2) convexity and closeness of $S^{FS}_{t_k}$, the GNBS is a unique solution of the bargaining problem $\left(S^{FS}_{t_k} \cup \{\boldsymbol{C}^{\boldsymbol{R}}_{\boldsymbol{C}\,t_k}\}, \boldsymbol{\mathcal{U}}_{t_k[N_B\times 1]}(\boldsymbol{C}^{\boldsymbol{Al}}_{t_k})\right)$ that satisfies Nash axioms as the attributes that any rational solution should meet to come up with fairness and efficiency [17].

$\mathcal{U}_{b,t_k}$ is convex, as for any $\boldsymbol{C}^{\boldsymbol{Al}_1}_{t_k[\mathrm{N_B}\times 1]}, \boldsymbol{C}^{\boldsymbol{Al}_2}_{t_k[\mathrm{N_B}\times 1]} \in S^{FS}_{t_k}$ and α with 0 ≤ α ≤ 1, the following inequality holds [19]:

$$\mathcal{U}_{b,t_k}\left(\alpha\boldsymbol{C}^{\boldsymbol{Al}_1}_{t_k} + (1-\alpha)\boldsymbol{C}^{\boldsymbol{Al}_2}_{t_k}\right) \le \mathcal{U}_{b,t_k}\left(\alpha\boldsymbol{C}^{\boldsymbol{Al}_1}_{t_k}\right) + \mathcal{U}_{b,t_k}\left((1-\alpha)\boldsymbol{C}^{\boldsymbol{Al}_2}_{t_k}\right) \tag{10}$$

$S^{FS}_{t_k}$ is also convex as all of points of the line that connects any desired pair of the points of the set, lies inside the set [19].

Since the convexity of both utility function and feasible solution set are guaranteed, the GNBS provides a unique solution $\boldsymbol{C}_{t_k}^{\boldsymbol{Al}^*}$ for the computational resource management problem of a BBU-pool as

$$\boldsymbol{C}_{t_k}^{\boldsymbol{Al}^*} = \underset{\boldsymbol{C}_{t_k}^{\boldsymbol{Al}}}{\operatorname{argmax}} \prod_{b=1}^{N_B} \left(\mathcal{U}_{b,t_k}\left(\boldsymbol{C}_{t_k}^{\boldsymbol{Al}}\right) - \mathcal{U}_{b,t_k}\left(\boldsymbol{C}_{\boldsymbol{C}\,t_k}^{\boldsymbol{R}}\right)\right)^{B_{b,t_k}}, \quad \forall \boldsymbol{C}_{t_k[N_B\times 1]}^{\boldsymbol{Al}} \in S_{t_k}^{FS} \cup \left\{\boldsymbol{C}_{\boldsymbol{C}\,t_k[N_B\times 1]}^{\boldsymbol{R}}\right\} \tag{11}$$

where the utilization function of the BBU-pool is defined as

$$\mathcal{U}_{GBP}\left(\boldsymbol{C}_{t_k}^{\boldsymbol{Al}}\right) = \prod_{b=1}^{N_B} \left(\mathcal{U}_{b,t_k}\left(\boldsymbol{C}_{t_k}^{\boldsymbol{Al}}\right) - \mathcal{U}_{b,t_k}\left(\boldsymbol{C}_{\boldsymbol{C}\,t_k}^{\boldsymbol{R}}\right)\right)^{B_{b,t_k}} \tag{12}$$

Regarding the constraints given in (5), the solution of the following optimization problem, acquires the GNBS:

$$\underset{\boldsymbol{C}_{t_k[N_B\times 1]}^{\boldsymbol{Al}}}{\text{maximize}} \quad \mathcal{U}_{GBP}\left(\boldsymbol{C}_{t_k[N_B\times 1]}^{\boldsymbol{Al}}\right) \tag{13a}$$

subject to:

$$\sum_{b=1}^{N_B} C_{b,t_k[\text{GOPS}]}^{Al} \leq C_{BP\;t_k[\text{GOPS}]}^{Av} \tag{13b}$$

$$C_{b,t_k[\text{GOPS}]}^{Al} \leq C_{b,t_k[\text{GOPS}]}^{R}, b = 1,2,\dots,N_B \tag{13c}$$

The defined optimization problem is convex, since the constraints given in (13b) and (13c) are linear inequalities, the Hessian matrix of the objective function is also negative semi-definite [19]. In order to find the solution of the given optimization problem, CVX [20] is used, which converges to the optimal solution efficiently using the primal-dual interior point method.

IV. Evaluation Metrics

Regardless of the minimum guaranteed computational capacity for the BBUs, for CP, the user fulfilment level, f_{b,t_k}^B, is a metric that measures the portion of the BBU's RCC that is served, given by

$$f_{b,t_k}^B = \frac{C_{b,t_k[\text{GOPS}]}^{Al} - C_{C\,b,t_k[\text{GOPS}]}^{R}}{C_{b,t_k[\text{GOPS}]}^{R} - C_{C\,b,t_k[\text{GOPS}]}^{R}} \tag{14}$$

The value of a BBU user fulfilment level is within [0, 1], with the higher value indicating that a greater portion of the computational resource demand is met.

The efficiency of the dynamic resource allocation is another metric achieved by comparing the total amount of AlCC that the model suggests a BBU-pool in a time instant, with the traditional approach, in which the computational capacity is allocated to the BBUs statically, based on the BBUs peak-hour demand. Hence, the efficiency of the proposed resource allocation is defined as

$$\eta_{t_k[\%]} = \left(1 - \frac{\sum_{b=1}^{N_B} C_{b,t_k[\text{GOPS}]}^{Al}}{\sum_{b=1}^{N_B} C_{b[\text{GOPS}]}^{R_{PEAK}}}\right) 100 \tag{15}$$

where $C_b^{R_{PEAK}}$ is the peak RCC of BBU b. A full 100% resource block usage, together with the highest modulation scheme, results in the peak RCC of a BBU [15]. A higher value of η_{t_k} indicates a more efficient resource allocation.

In addition, the resource allocation fairness is compared with the result of Jain's fairness indicator [21], which defines the closeness of the fulfilment level of the BBUs to the average weight of the active services in the BBUs, given by

$$F_{t_k} = \frac{\left(\sum_{b=1}^{N_B} f_{b,t_k}^B \big/ \overline{w_{b,t_k}^{srv}}\right)^2}{N_B \sum_{l=1}^{N_B}\left(f_{l,t_k}^B \big/ \overline{w_{l,t_k}^{srv}}\right)^2} \tag{16}$$

The range of $F_{t_k}$is within [0,1]. The higher the value of F_{t_k}, the higher the fairness of the resource allocation at time instant t_k.

V. Scenario Description

Proposed model's performance is evaluated for a scenario in which one BBU-pool serves 7 micro-cells: 4 cells are located in a residential area and the other 3 are in a business one. BSs are assumed to have a 8 × 8 MIMO order and operate on a 40 MHz channel bandwidth, with 24 bit of quantization resolution.

In order to evaluate the model, the performance metrics defined in Section IV were used under the condition that the AvCC in the BBU-pool is increased from 0.16 to 6 TOPS, the system/user parameters being summarized in TABLE I.

TABLE I. System /User parameters.

# Spatial Streams	8
Cell Type	Micro
Channel Bandwidth [MHz]	40
Quantization Resolution [b]	24
MIMO Order	8 × 8
# BBUs in the BBU-pool	7
BBU-pool AvCC [TOPS]	[0.16, 6]

BBUs perform multiple services, i.e. VoIP, video streaming, video calling and file transferring, the corresponding traffic following the model proposed in [22].

A single snapshot of the network is taken in order to evaluate performance. The service weights resulted from (6) are listed in TABLE II, together with the number of active users in an individual BBU asking for that specific service.

TABLE II. Service characteristics.

ID	Weight	# Active Users (N_{b,t_k}^U)						
		BBU index (b)						
		1	*2*	*3*	*4*	*5*	*6*	*7*
VoIP	83	20	63	5	21	0	2	6
Video Streaming	48	16	10	11	10	6	8	2
File Transfer	36	1	0	4	0	2	0	0
Video Calling	59	12	9	11	7	0	0	2

As mentioned before, the highest modulation scheme (i.e. 1024QAM as discussed in [23] for upcoming wireless technologies) together with the 100% resources block usage results in the peak RCC of a BBU [15]. Therefore, the peak RCC becomes 4.8 TOPS in the defined scenario.

VI. Analysis of Results

The application of the scenario to the model described in Section III, results in the values of RCC, average weight of active services, minimum guaranteed computational capacity and BPs for BBUs that are listed in TABLE III. It shows that the BBU with index 2 has the highest average of service weight, since the majority of its services are VoIP (highest service priority). On the opposite, the BBU with index 5, with no VoIP, has the lowest average weight.

Fig. 2 shows AlCC in a BBU when AvCC, $C_{BP\;t_k}^{Av}$, increases from 0.16 to 6 TOPS. When AvCC equals 0.16 TOPS only the minimum guaranteed computational capacity is allocated to the BBUs due to the resource shortage. None of the BBU demands can be fully met before Th₁, since the sum of the RCCs is higher than AvCC. Once the minimum guaranteed requirements are allocated, the rest of the resources are distributed among BBUs with respect to the priority of each BBU, i.e., BBU BP.

TABLE III. BBUS' RCC, AVERAGE WEIGHT OF ACTIVE SERVICES, MINIMUM GUARANTEED ALCC AND BP IN THE SNAPSHOT t_k.

RRH Traffic Type	**Residential**				**Business**		
BBU Index (b)	*1*	*2*	*3*	*4*	*5*	*6*	*7*
$C^R_{b,t_k[\text{TOPS}]}$	1.19	2.25	0.69	0.77	0.26	0.01	0.29
$\overline{w^{srv}_{b,t_k}}$	64.74	76.10	56.00	69.37	45.00	55.00	71.20
$C^R_{C\,b,t_k[\text{GOPS}]}$	34.89	32.11	30.93	21.18	20.80	4.47	11.24
$B_{b,t_k[\%]}$	20.58	46.48	10.15	14.37	2.90	0.12	5.39

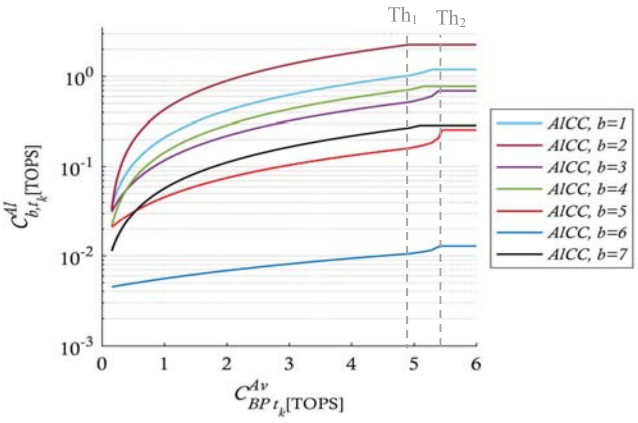

Fig. 2. BBUs' AlCC.

The effect of the BP is apparent when BBU index 5 is compared to BBU 7: the former receives more resources in the beginning, because its minimum guaranteed requirement is higher than the letter; however, when AvCC increases, the AlCC of BBU 7 exceeds the one of BBU 5 due the fact that BBU 7 has higher BP than BBU 5, hence, a higher priority in resource distribution in the pool. Fig. 2 also shows that the BBU minimum requirements are always guaranteed, and that the BBU AlCC never exceeds the RCC.

The BBUs' fulfilment level is presented in Fig. 3.

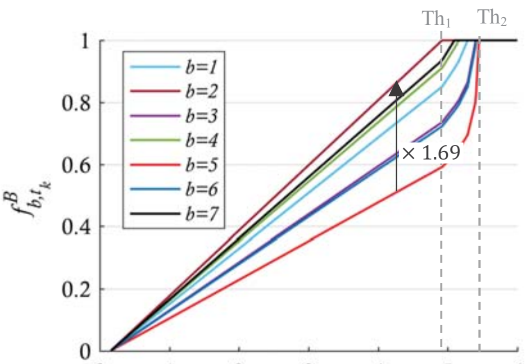

Fig. 3. BBUs fulfilment level.

By increasing AvCC, the BBU fulfilment level is also improved, proportionally to the average weight of the active services, before threshold Th₁. By comparing BBUs 5 and 2, it is confirmed that fulfilment levels have the same proportion of the average weights of active services, i.e. 1.69, up to Th₁. Between Th₁ and Th₂, however, the fulfilment level of BBU 5 grows fast, because the demand of BBUs with higher priority have already been met before Th₁. Since the allocated resources to the BBU cannot exceed the demand, with the increase of AvCC, the remaining resources become available to the lower prioritized BBUs.

It is also seen in Fig. 3 that BBU 2 is the first to receive 100% of its demand with the increase of AvCC, since it has the highest average service weight among other BBUs in the pool. On the contrary, BBU 5 is the last one that is fulfilled, since its active services have the lowest average weight.

The efficiency of the proposed resource allocation scheme is presented in Fig. 4. With the increase of AvCC, the efficiency decreases, as more resources are used. Although AvCC is still increasing beyond Th₂, the resource usage does not increase anymore. The reason is that the resource allocating scheme stops allocating more resources to the BBUs once their demand is fully met, hence, efficiency does not fall below 83%.

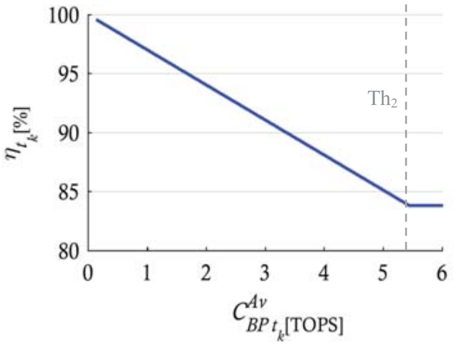

Fig. 4. Resource allocation efficiency.

Jain's fairness indicator, the last evaluation metric, is presented in Fig. 5. The allocation is defined to be fair if the fulfilment levels maintain the same proportion of the average weights of active services.

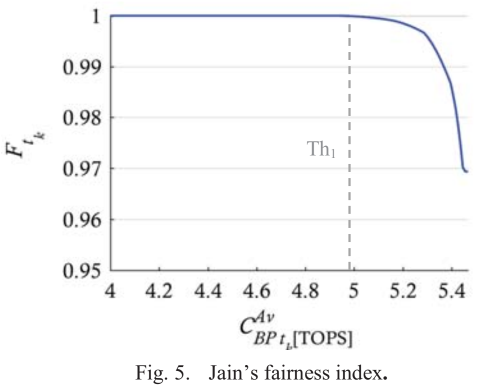

Fig. 5. Jain's fairness index.

As Fig. 5 shows, the fairness condition holds before Th₁. Beyond Th₁, however, the fairness indicator decreases, due the fact that the computational capacity proportional to the service weights is more than the RCC for the BBUs with high priority services. The resource allocation strategy bounds the AlCC in BBUs to their RCC so that remaining capacity is distributed among those with lower service priority. As a result, the AlCC of the BBUs with high priority services is less than the amount that is proportional to the average weight of their active services; on the other hand, the BBUs with lower service priority receive more than the average of their active services ratio. This will end up with the decrease of the defined fairness index, as the fairness condition does not hold. The decrease of defined fairness index confirms that the resource allocator takes not only the priority of services but also instantaneous requirement of the BBUs into account while distributing resources among them.

VII. Conclusions

A resource management strategy is proposed, aiming at maximizing the computational resource utilization in C-RAN. To this end, the computational resource allocation in the BBU-pool is modelled as a bargaining game, in which a BBU power in the bargaining step depends on the amount of instantaneous demand of the BBU and on the priority of active services. The bargaining powers specify the importance of BBUs while distributing resources among them.

BBUs fulfilment level, resource allocation efficiency and fairness are the metrics used to evaluate model performance. To this purpose, a scenario is defined in which one BBU-pool serves seven cells within a heterogeneous service environment. Results indicate that the proposed model allocates computational resource to the BBUs proportional to their bargaining power, the minimum required computational resource is always guaranteed, and the summation of the allocated resources never exceeds the available resources in the BBU-pool. In addition, it is shown that the BBUs that are processing services with a higher priority, receive higher fulfilment levels. The results also confirm that the resource allocation strategy is 100% fair in shortage conditions where none of the BBUs' demand can be fully met. Moreover, the model achieves a significant improvement in allocation efficiency, a minimum 83%, compared to the fixed resource allocation based on the peak-hour traffic.

Acknowledgment

This work was carried out within the framework of the COST Action CA15104 (IRACON).

References

[1] Cisco, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022,* White paper, Feb. 2019. [Online]. Available: https://www.cisco.com.

[2] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for Mobile Networks—A Technology Overview," *IEEE Communications Surveys & Tutorials,* vol. 17, no. 1, pp. 405-426, Firstquarter 2015.

[3] A. Beloglazov and R. Buyya, "Managing overload hosts for dynamic consolidation of virtual machines in cloud data centres under quality of service constraints," *IEEE Transactions on Parallel and Distributed Systems,* vol. 24, no. 7, p. 1366–1379, July 2013.

[4] Z. Xiao, W. Song and Q. Chen, "Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment," *IEEE Transactions on Parallel and Distributed Systems,* vol. 24, no. 6, p. 1107–1117, June 2013.

[5] T. Werthmann, H. Grob-Lipski, S. Scholz and B. Haberland, "Task assignment strategies for pools of baseband computation units in 4G cellular networks," in *ICC 2015 - IEEE International Conference on Communication Workshop*, London, UK, June 2015.

[6] B. J. R. Sahu, S. Dash, N. Saxena and A. Roy, "Energy-Efficient BBU Allocation for Green C-RAN," *IEEE Communications Letters,* vol. 21, no. 7, pp. 1637-1640, July 2017.

[7] A. Al-Dulaimi, S. Al-Rubaye and Q. Ni, "Energy Efficiency Using Cloud Management of LTE Networks Employing Fronthaul and Virtualized Baseband Processing Pool," *IEEE Transactions on Cloud Computing,* vol. 7, no. 2, pp. 403-414, June 2019.

[8] W. Chien, C. Lai and H. Chao, "Dynamic Resource Prediction and Allocation in C-RAN With Edge Artificial Intelligence," *IEEE Transactions on Industrial Informatics,* vol. 15, no. 7, pp. 4306-4314, July 2019.

[9] D. Pompili, A. Hajisami and T. X. Tran, "Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN," *IEEE Communications Magazine,* vol. 54, no. 1, pp. 26-32, Jan. 2016.

[10] Y. Liao, L. Song, Y. Li and Y. Angela Zhang, "How Much Computing Capability Is Enough to Run a Cloud Radio Access Network?," *IEEE Communications Letters,* vol. 21, no. 1, pp. 104-107, Jan. 2017.

[11] NGMN Allaiance, *Ran Evolution Project Comp Evaluation And Enhancement,* CoMP Work Stream, RAN Evolution Project, Final Deliverable, Mar. 2015. [Online]. Available: https://www.ngmn.org/wp-content/uploads/NGMN_RANEV_D3_CoMP_Evaluation_and_Enhancement_v2.0.pdf.

[12] J. Zhang, Y. Ji, S. Jia, H. Li, X. Yu and X. Wang, "Reconfigurable optical mobile fronthaul networks for coordinated multipoint transmission and reception in 5G," *IEEE Journal of Optical Communications and Networking,* vol. 9, no. 6, pp. 489-497, June 2017.

[13] I. Alyafawi, E. Schiller, T. Braun, D. Dimitrova, A. Gomes and N. Nikaein, "Critical issues of centralized and cloudified LTE-FDD Radio Access Networks," in *ICC 2015 - IEEE International Conference on Communication Workshop*, London, UK, June 2015.

[14] B. Debaillie, C. Desset and F. Louagie, "A Flexible and Future-Proof Power Model for Cellular Base Stations," in V*TC 2015 Spring - IEEE Vehicular Technology Conference*, Glasgow, Scotland, May 2015.

[15] MAMMOET - Massive MIMO for Efficient Transmission, [Online]. Available: https://mammoet-project.eu/. [Accessed Sep. 2019].

[16] M. Barahman, L. M. Correia and L. S. Ferreira, "A Fair Computational Resource Management Strategy in C-RAN," in *CobCom 2018 - International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications*, Graz, Austria, July 2018.

[17] R. B. Myerson, Game Theory: Analysis of Conflict, Cambridge, MA, USA: Harvard University Press, 1991.

[18] 3GPP, *Technical Specification Group Services and System Aspects; Policy and charging control architecture,* Technical Specification TS 23.203 V16.1.0, Sep. 2019.

[19] S. P. Boyd, L. Vandenberghe, Convex Optimization, West Nyack, NY, USA: Cambridge University Press, 2004.

[20] CVX – Software for Disciplined Convex Programming, [Online]. Available: http://cvxr.com. [Accessed Sep. 2019].

[21] R. K. Jain, D. M. Chiu and W. R. Hawe, *A quantitative measure of fairness and discrimination for resource allocation in shared systems,* DEC Technical Report TR-301, Digital Equipment Corporation, Sep. 1984. [Online]. Available: https://www.cse.wustl.edu/~jain/papers/ftp/fairness.pdf.

[22] NGMN Allaiance, *Next Generation Mobile Networks Radio Access Performance Evaluation Methodology,* White paper, Jan. 2008. [Online]. Available: https://www.ngmn.org/fileadmin/user_upload/NGMN_Radio_Access_Performance_Evaluation_Methodology.pdf.

[23] 3GPP, *Discussion on CQI and MCS table,* 3GPP TSG-RAN WG1 Meeting #91, R1-1719731, Nov. 2017. [Online]. Available: https://www.3gpp.org/ftp/TSG_RAN/WG1_RL1/TSGR1_91/Docs/.