Received 9 October 2020; accepted 21 October 2020. Date of publication 29 October 2020; date of current version 9 December 2020.

Digital Object Identifier 10.1109/OJCOMS.2020.3034297

A QoS-Demand-Aware Computing Resource Management Scheme in Cloud-RAN

MOJGAN BARAHMAN [1] (Member, IEEE), LUIS M. CORREIA [1] (Senior Member, IEEE), AND LÚCIO STUDER FERREIRA [2] (Senior Member, IEEE)

[1] INESC-ID/Instituto Superior Técnico, University of Lisbon, 1000-029 Lisbon, Portugal

[2] INESC-ID/COPELABS, Universidade Lusófona, 1749-024 Lisbon, Portugal

CORRESPONDING AUTHOR: M. BARAHMAN (e-mail: mojgan.barahman@tecnico.ulisboa.pt)

ABSTRACT This article focuses on computing resource allocation in Cloud Radio Access Networks. A game-based optimization algorithm was developed to distribute the computing resources among BaseBand Units (BBUs) in a BBU-pool whereby resources utilization is maximized. The model allocates computing resources on-demand, based on the instantaneous requests of BBUs, using a game-theory bargaining approach. In the case that the available resources are not sufficient to fulfil all instantiation requests, BBUs are prioritized to ensure the adequate Quality of Service, low-priority ones being always guaranteed a minimum computing resource to avoid them to crash. The performance of the proposed model is observed over time, concerning resource usage, BBU fulfilment level and efficiency. Simulations in a group of cells with a mixture of heterogeneous services in tidal traffic conditions show that resources allocated to BBUs are consistent with the priority of ongoing services and in line with real-time demand. Results also show that improving the average fulfilment level from 98% to 100% requires doubling the available resources at the cost of the average resources usage being cut in half.

INDEX TERMS Cloud-RAN, computing resource utilization, resource allocation efficiency, wireless communications.

I. INTRODUCTION

Cloud Radio Access Networks (C-RANs) emerged in response to the need for higher data rates and capacity in upcoming mobile network generations, e.g., 5G [1]: BaseBand processing Units (BBUs) of Base Stations (BSs) are decoupled from the radio units, known as Radio Remote Heads (RRHs); software-based BBUs are then centralized and consolidated in the servers of a data center, known as BBU-pools. C-RAN is a critical enabling technology of 5G [2], providing not only higher data rates but also lower network latencies, by multiplexing the BBU resources in the pool. Resources multiplexing enables overloaded BBUs to use residual resources left by the underutilized ones, hence, utilization is improved, and fewer resources are required rather than the sum of stand-alone BBU demands [3], [4].

Although the consolidation of resources in C-RAN reduces the number of the required resources in the network, there are still critical challenges for data centers, such as power consumption [5]–[7]: a medium-sized one with 930 m^2 and 288 racks can consume 4 MW in the traffic peak [8]. Since computing resources, i.e., servers, are the most energy-intensive entities in data centers, it is worthwhile to apply efficient resource management strategies to maximize their utilization and reduce the number of idle ones; an idle server has no productivity, but still consumes 60% of its peak power usage [8].

However, designing efficient resource management strategies is a complicated process for cloud providers. Due to the variety of network services, user arrival rates and channel conditions, BBU resources demand fluctuate significantly throughout the day. On the one hand, a BBU computing capacity should suffice peak demands; on the other hand, provisioning fixed resources based on peak requirements leads to idle resources in the rest of the day.

As a result, an efficient resource management strategy in a BBU-pool should allocate the computing capacity dynamically, in accordance with the BBUs' instantaneous demand, while efficiently handling the resources in the case of a shortage. Resource shortages are time instants in which the BBU-pool's available resources are less than demand spikes, and come into play in two circumstances: when the objective is intentionally to design the pool with minimum computing resources; or, even if there are more computing resources, they cannot be initialized at a rate similar to the one of demand fluctuations (in the scale of milliseconds), due to hardware limitations.

In this work, a BBU-pool computing resource allocation scheme is proposed within a dynamic traffic demand environment. The proposed model estimates the BBUs' demands and reconfigures BBUs' Allocated Computing Capacity (AlCC) accordingly. The main objective is to maximize the utilization of BBU-pool computing resources, which is crucial to guarantee low power consumption in the network. The novelty of the proposed scheme is the consideration of the limits of the BBU-pool computing resources and the prioritization of BBUs in bottlenecks based on the characteristics of their ongoing services and Quality of Service (QoS) constraints. At the same time, the model guarantees all BBUs with a minimum computing resources to avoid crashing; furthermore, contrary to existing works, the proposed model has a low complexity and provides fairness of resource allocation and system efficiency, which makes it applicable in practical implementations.

In this context, computing resource allocation in a BBU-pool is modeled as a game-theory bargaining game. Players, i.e., BBUs, compete for the limited computing resources of the BBU-pool to maximize their processing speed. The Generalized Nash Bargaining Solution (GNBS) with adaptive bargaining powers [9] is applied in order to find a solution for the bargaining game. The two-fold solution maximizes both the BBU-pool computing resource utilization and the processing speed of the BBUs. QoS constraints are taken into account and service characteristics are monitored in real-time, which is essential not only in 4G deployments but also for the upcoming service-oriented 5G.

In [10], the first draft of the proposed model was presented, still limited to a single time instant, being then improved in [11], by addressing time-varying traffic and demand. The current paper provides an extension of the above concepts, by evaluating model performance considering real-time network traffic in a tidal channel condition. The work compares the performance of the proposed model against equal and demand-proportional resource allocation schemes, which can be found in the literature as common allocation approaches, in terms of resource usage and BBU fulfilment level. Moreover, it studies how the limit of the Available Computing Capacity (AvCC) is correlated with the fulfilment level of BBU demands. In general terms, the more resources are available, the better the fulfilment level and the lower the resource usage occur. However, above certain levels, the provisioning of more resources degrades the average resource usage dramatically, while contributing very little (or nothing) to improving the demands' fulfilment level.

The rest of this article is organized as follows. In Section II, related works are mentioned. The selected network architecture and the main assumptions are given in Section III. Section IV provides an overview of the proposed model and mentions the approach for estimating the computing resource demand of BBUs. Section V explains the resource allocation model.In Section VI, evaluation metrics are introduced, and in Section VII, a scenario is characterized, and results are analyzed. Finally, Section VIII concludes this article.

II. RELATED WORKS

Several resource management approaches have been proposed in the literature, aiming at maximizing C-RAN computing resource utilization. The suggested strategies can be classified into two main categories, depending on the computing capacity of the BBUs in the pool being allocated either in a fixed mode or in an adaptive one.

In fixed schemes, the computing capacity of the BBUs in the pool are not changeable. In the case of BBU overloading, the excess load is migrated to other underutilized active BBUs, enabling the overloaded BBU to use the extra resources left by the other ones in the pool, at a specific time instant. Consequently, the load becomes more balanced, leading to improved resource utilization and energy efficiency. Within this framework, Wang *et al.* [12] formulated the C-RAN resource management problem as a linear integer programming one. The proposed model re-assigns the processing tasks that cause BBU overloading to appropriate underutilized BBUs, so that the BBU-pool resource utilization is enhanced.

Additionally, load migration enables reducing the number of active BBUs by consolidating the processing task of multiple BBUs in a few ones in off-peak hours, when most of the BBUs in the pool are underutilized. Sundaresan *et al.* [13] suggested a dynamic RRH to BBU mapping framework, which enables a BBU to serve several RRHs at the same time: the goal was to minimize the idle resources by reducing the number of active BBUs when the traffic load is low and a single BBU is sufficient, showing a 50% improvement in resource usage, compared to the baseline one-to-one RRH to BBU mapping strategy. Similarly, Al-Dulaimi *et al.* [14] proposed a model based on graph coloring to switch off low traffic BBUs and divert their processing load to neighboring underloaded ones in the pool.

The authors in [15], [16] and [17] also formulated the BBU-pool resource allocation as a bin packing problem. BBUs are treated as bins with finite computing capabilities and the cell processing tasks as the items that should be packed in the bins so that fewer BBUs are used; they used heuristic algorithms to solve the defined problem. Chien *et al.* [18] went beyond the BBU-pool and proposed a resource management model to improve network resource usage by turning off the BBU-pools with low traffic and redirecting their RRHs in the network.

Many works in the literature focus mainly on load migration as a strategy for resource utilization optimization in

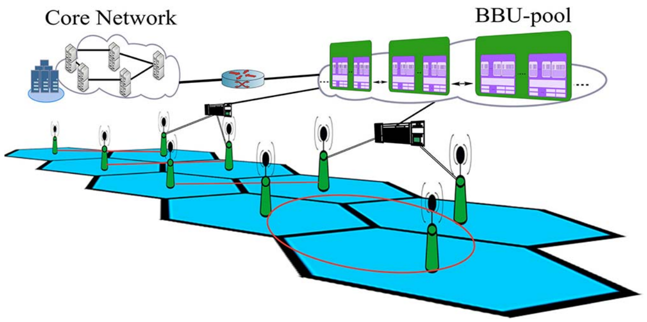


FIGURE 1. C-RAN architecture.

a BBU-pool. However, this policy imposes additional overheads to the network due to increased data exchanges between the source and the target BBUs [19]. The migration cost is higher in dense areas, since handover, Coordinated Multi-Point transmission/reception and interference occur more often among small cells [20]. One approach for reducing the data exchange burden is to serve coordinated RRHs with a single BBU [21], and the BBU computing capacity being elastically reconfigured according to its real-time demand. As a result, an adaptive computing capacity strategy is chosen in this article in order to optimize the computing resource utilization of the BBU-pool.

To the best of our knowledge, to date only a few works on BBU-pool resource management have considered adaptable computing resources for the BBUs. Pompili *et al.* [22] proposed a framework for elastic and on-demand computing resource allocation to the BBUs in the pool employing virtualization techniques. The BBU functions are performed on the Virtual Machines (VMs) that are reposed on top of general-purpose servers, and their model estimates BBU demands with reference to a given pattern, delivering the BBU-pool computing resources accordingly. Based on a similar platform, Yu *et al.* [23] proposed a model to improve the computing resource utilization of a BBU-pool by switching off the low traffic RRHs and their associated BBUs, diverting their processing load to the neighbors in the pool. If required, more resources are allocated to the target BBUs in order to improve their processing capability. The models proposed in [22] and [23] improve the computing resource utilization; however, both assume that there are always adequate resources in the pool to meet the peak demands and do not suggest a resource management strategy in the case of a resource shortage.

III. SYSTEM ARCHITECTURE AND ASSUMPTIONS

In this work, one considers a C-RAN architecture that can be used for both 4G and 5G. The selected architecture is presented in Fig. 1, where BBUs from multiple BSs are aggregated in a BBU-pool and each BBU is connected to its RRH through a high-speed optical link. The BBU-pools are linked in the upper level also via high-speed connections [2].

A BBU-pool is a centralized location, including computing resources of multiple BSs being consolidated in general-purpose servers, which are shared and flexibly allocated to the BBUs through virtualization techniques. The BBU process is performed as software applications on VMs that are reposed on top of servers [24]. The computing capacity of the BBUs in the pool is elastically allocable, meaning that resources are assigned and released to the BBUs adaptively, based on their real-time demand. Despite the fact that the computation resources of a server include Input/Output (I/O), storage, memory, Central Processing Unit (CPU), etc., for simplicity, only CPU (with the same configuration for all servers) is considered as the processing resource in the pool.

Although a BBU can transmit/receive a signal to/from several RRHs [2], for simplicity, it is assumed that each RRH is served by one BBU in the pool and that a BBU serves just one RRH. Besides, without loss of generality, only user plane data transmission is considered in this work, taking channel de/coding, de/modulation, Multiple-Input Multiple-Output (MIMO) de/pre-coding, channel estimation, and Orthogonal Frequency Division Multiple Access (OFDMA) and Single Carrier - Frequency Division Multiple Access (SC-FDMA) into account as the primary signal processing steps of the BBUs. However, by using a model similar to the one presented in Section IV-B, the proposed model can be fitted to the whole protocol stack layers and the control plane data transmission and signaling.

It is also assumed that in the case of a packet loss the transmitter resends the same packet under the Hybrid Automatic Repeat reQuest (HARQ) process, type I, and the retransmitted packet is treated as a new arrived one.

The BBU process is classified into two main groups [25]:

- *User Processing (UP):* it includes the signal processing steps that can be split per user and its set, S^{UP}, contains all UP steps in the BBU, such that

$$S^{UP} = \left\{P^{chc}, P^{chd}, P^{md}, P^{dm}, P^{mpc}, P^{mdc}, P^{che}\right\} \quad (1)$$

  where:
  - P^{chc}: channel coding,
  - P^{chd}: channel decoding,
  - P^{md}: modulation,
  - P^{dm}: demodulation,
  - P^{mpc}: MIMO pre-coding,
  - P^{mdc}: MIMO decoding,
  - P^{che}: channel estimation.
- *Common Processing (CP):* it includes the common signal processing steps among all users for a given RRH, the set, S^{CP}, containing all CP steps in the BBU, such that

$$S^{CP} = \left\{P^{OFDMA}, P^{SCFDMA}\right\} \quad (2)$$

  where:
  - P^{OFDMA}: OFDMA,
  - P^{SCFDMA}: SC-FDMA.

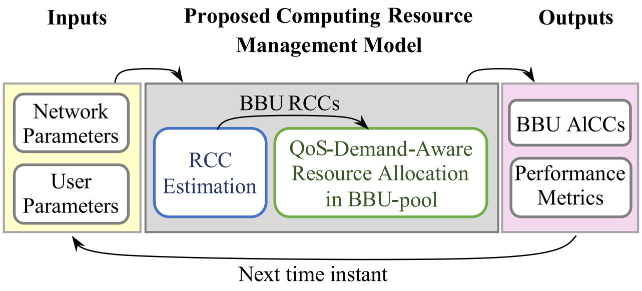


FIGURE 2. Global view of the QoS-demand-aware computing resource management model over time.

For the sake of clarity, all the notations that are used in this article are listed in the Appendix.

IV. MODEL OVERVIEW AND DEMAND ESTIMATION

A. GENERAL OVERVIEW

The aim of providing an efficient resource allocation strategy in a BBU-pool is to maximize resource utilization. To achieve this goal, resources should be allocated to BBUs based on their real-time demand such that QoS is maintained. Hence, the first step is traffic demand evaluation and the optimal solution for resource utilization can be found only afterwards. In this way, the proposed resource management algorithm comprises two components, Fig. 2:

- *Required Computing Capacity (RCC) Estimation:* Calculation of instantaneous demand (measured in Operations per Second [OPS]) of BBUs, according to the real-time network/user parameters.
- *QoS-Demand-Aware Resource Allocation in a BBU-Pool:* Obtaining the optimal on-demand computing resource allocation that maximizes both BBU-pool resource utilization and efficiency with respect to the required QoS.

Taking as inputs network and user parameters at a specific time instant, the estimation of the BBUs' RCC is based on a well-defined model [26] and [27]. The results are then fed to the computing resource allocation step in order to find the optimal AlCC to BBUs. To this end, the BBU-pool computing resource allocation is formulated as a game-theory based bargaining problem, which is solved by the corresponding axiomatic solutions.

In the next time instant, the resource management process is re-instantiated over new input parameters. Considering both QoS and BBU RCCs as real-time parameters, i.e., given on the basis of Time Transmission Intervals (TTIs), ensures that the BBU-pool is provisioned with an optimum configuration, consistent with BBU demands.

B. REQUIRED COMPUTING CAPACITY ESTIMATION

The RCC of a BBU is defined as the minimum computing capacity that it requires in order to adequately perform the instantaneous signal processing within a TTI. RCC estimation is done by a function of parameters that are effective on computing the complexity of signal processing, being listed in the set X,

TABLE 1. Reference values for effective parameters on computing complexity of signal processing (based on [27]).

Parameter (x)	$\Delta f_{BW[\text{MHz}]}$	$m_{u[\text{bit/symbol}]}$	r_u	N_u^{Str}	N_{MIM}	$Q_{[\text{bit}]}$	$\eta_b^{RB_B}$
Reference Value (x^{ref})	20	6	1			16, 24	1

$$X = \left\{ \Delta f_{BW[\text{MHz}]}, N_{MIM}, Q_{[\text{bit}]}, m_{u[\text{bit/symbol}]}, r_u, N_u^{Str} \right\} \tag{3}$$

where:

- Δf_{BW}: bandwidth (e.g., $\Delta f_{BW} \in \{20, 40, 100\}$ [MHz],
- N_{MIM}: MIMO order (e.g., $N_{MIM} \in \{1, 2, 4, 8\}$),
- Q: quantization resolution, (e.g., $Q \in \{16, 24\}$ [bit]),
- m_u: user u modulation (e.g., $m_u \in \{8, 10\}$ [bit/symbol]),
- r_u: user u coding ratio (e.g., $r_u \in [1/4, 1]$),
- N_u^{Str}: user u number of streams (up to the MIMO order).

Besides, Resource Block (RB) efficiency, η^{RB}, also affects the complexity of signal processing: for a single user at time instant t_k, $\eta_{u,t_k}^{RB_U}$ is the fraction of available RBs in the bandwidth being allocated to the user, given by

$$\eta_{u,t_k}^{RB_U} = \frac{N_{u,t_k}^{RB}}{N_{\Delta f}^{RB}} \tag{4}$$

where:

- N_{u,t_k}^{RB}: number of allocated RBs to user u at t_k,
- $N_{\Delta f}^{RB}$: total number of sub-frame RBs in a given bandwidth, (e.g., $N_{\Delta f}^{RB} = 200$, in a 20 MHz bandwidth).

The sum of all active users' RB efficiency in a BBU, at a specific time, states the BBU RB efficiency, so

$$\eta_{b,t_k}^{RB_B} = \sum_{\forall u \in S_{b,t_k}^U} \eta_{u,t_k}^{RB_U} \tag{5}$$

where S_{b,t_k}^U is the set of all active users in BBU b at t_k.

In order to estimate a BBU RCC, a reference value is given to each of the effective parameters. Accordingly, an algorithm is selected for every signal processing step. The RCC of a UP/CP step is then acquired, by counting the number of arithmetic operations that should be performed per information bit transmission. The reference values assigned to the parameters $x \ \in X$ and processing step RCCs obtained from them (which are the reference RCCs) are listed in TABLE 1 and TABLE 2, respectively.

The reference RCCs can then be scaled to any other desired value of x. For each UP processing step $p \in S^{UP}$ of user u, the scaling is given by

$$C_{u,p,t_k[\text{GOPS}]}^{R_{UP}} = C_{p[\text{GOPS}]}^{ref} \left(\eta_{u,t_k}^{RB_U} \right)^{E_{\eta^{RB},p}} \prod_{\forall x \in X} \left(\frac{x_{t_k}}{x^{ref}} \right)^{E_{x,p}} \tag{6}$$

where:

- C_p^{ref}: reference RCC of processing step p, TABLE 2,
- x_{t_k}: parameter x in the operating scenario at t_k,

TABLE 2. Reference RCCs and scaling exponents (based on [27]).

BS Processing Step (p)	Reference RCC ($C^{ref}_{p[\text{GOPS}]}$)	Effective Parameter (x) Scaling Exponent ($E_{x,p}$)						
		Δf_{BW}	m_u	r_u	N_u^{Str}	N_{MIM}	Q	η^{RB}
p^{SCFDMA}	2.7	1.0	-			1.0	1.2	0.5
p^{OFDMA}	1.3							
p^{che}	3.3		0.0		1.0			
p^{mpc}	1.3							1.0
p^{mdc}	2+3.3 N_{MIM}				0.0	2.0		
p^{dm}	2.7		1.5		1.0	0.0		
p^{md}	1.3							
p^{chd}	8.0		1.0					
p^{chc}	1.3							

- x^{ref}: reference value of x, TABLE 1,
- $E_{x,p}$: scaling exponents of x on step p, TABLE 2.

Accordingly, the total user processing RCC of BBU b is

$$C^{R}_{U\ b,t_k} = \sum_{\forall u\in S^{U}_{b,t_k}} \sum_{\forall p\in S^{UP}} C^{R_{UP}}_{u,p,t_k}. \tag{7}$$

On the other hand, for each CP step $p \in S^{CP}$ of BBU b, the scaling is

$$C^{R_{CP}}_{b,p,t_k[\text{GOPS}]} = C^{ref}_{p[\text{GOPS}]}\left(\eta^{RB_B}_{b,t_k}\right)^{E_{\eta^{RB},p}} \prod_{\forall x\in X}\left(\frac{x_{t_k}}{x^{ref}}\right)^{E_{x,p}}. \tag{8}$$

Finally, the total RCC of BBU b at a time instant t_k is achieved by summing the RCC of the CP steps and all users' UP steps,

$$C^{R}_{b,t_k} = \sum_{\forall p\in S^{CP}} C^{R_{CP}}_{b,p,t_k} + C^{R}_{U\ b,t_k}. \tag{9}$$

In the presence of several active users transmitting/ receiving signals in a cell, the RCC of the CP steps should be guaranteed, otherwise, none of the users' data can be transmitted/received. In other words, the minimum guaranteed RCC of BBU b is

$$C^{R_{\min}}_{b,t_k[\text{GOPS}]} = \sum_{\forall p\in S^{CP}} C^{R_{CP}}_{b,p,t_k[\text{GOPS}]}. \tag{10}$$

V. RESOURCE ALLOCATION

This section introduces the proposed QoS-Demand-Aware resource allocation Scheme (QDAS) in BBU-pool, including the formulation of the resource allocation problem as a bargaining game, Section V-A, and the application of the GNBS to find the optimal solution for the problem, Section V-B. Two more resource allocation schemes are also explained in Section V-C for comparison purposes.

A. GAME FORMULATION

The problem of finding an efficient resource allocation in the BBU-pool is comparable with a bargaining game in cooperative game-theory [9], [10]. BBUs are counted as players that are negotiating over a limited number of computing resources of the BBU-pool to increase their processing capacities, while taking resource utilization maximization as a mutual benefit into account. The outcome is an agreement on selecting one resource allocation strategy, i.e., a feasible solution from many possible choices.

A resource allocation strategy at a certain time instant t_k is given by vector $\mathbf{C}^{Al}_{t_k[N_B\ \times 1]}$,

$$\mathbf{C}^{Al}_{t_k} = \left[C^{Al}_{1,t_k[\text{GOPS}]}, C^{Al}_{2,t_k[\text{GOPS}]}, \ldots, C^{Al}_{N_B,t_k[\text{GOPS}]}\right]^{\mathrm{T}} \tag{11}$$

where:

- C^{Al}_{b,t_k}: BBU b AlCC at time instant t_k,
- N_B: number of BBUs in the pool.

Each BBU evaluates its preference over a selected strategy by its utility function individually. The utility of BBU b at t_k is defined by a function $\mathcal{U}_{b,t_k} : \mathbb{R}^{N_B} \to \mathbb{R}$ that reflects the portion of the BBU's request that is satisfied, given by

$$\mathcal{U}_{b,t_k}\left(\mathbf{C}^{Al}_{t_k}\right) = \frac{C^{Al}_{b,t_k[\text{GOPS}]}}{C^{R}_{b,t_k[\text{GOPS}]}}. \tag{12}$$

If the total computation demand is less than the available resources in the BBU-pool, then all BBUs' demands are satisfied; otherwise, a compromise solution is selected in which the minimum guaranteed RCC of BBUs, $\mathbf{C}^{R_{\min}}_{t_k[N_B\times 1]}$, are served,

$$\mathbf{C}^{R_{\min}}_{t_k} = \left[C^{R_{\min}}_{1,t_k[\text{GOPS}]}, C^{R_{\min}}_{2,t_k[\text{GOPS}]}, \ldots, C^{R_{\min}}_{N_B,t_k[\text{GOPS}]}\right]^{\mathrm{T}}. \tag{13}$$

During the bargaining, BBUs attempt to get more computing resources to increase their utility. However, three limitations are imposed:

1) the total AlCC in a feasible solution should not exceed the BBU-pool's AvCC,
2) the resource allocator must provide the minimum guaranteed RCC of an individual BBU, i.e., $C^{R_{\min}}_{b,t_k}$,
3) each BBU may not ask more capacity than its RCC at a specific time.

As a result of these critical constraints, the feasible solution set is bounded as

$$S^{FS}_{t_k} = \left\{\mathbf{C}^{Al}_{t_k} | \sum_{b=1}^{N_B} C^{Al}_{b,t_k} \le C^{Av}_{BP\ t_k}, \right.$$
$$\left. C^{R_{\min}}_{b,t_k} < C^{Al}_{b,t_k} \le C^{R}_{b,t_k}\right\} \tag{14}$$

for $b = 1, 2, \ldots, N_B$, where $C^{Av}_{BPt_k}$ is the BBU-pool AvCC at t_k. $S^{FS}_{t_k}$ is a convex set, because the line segment between any desired pair of points in the set lies entirely within the set [28]; in other words,

$$\forall \mathbf{C}^{Al_1}_{t_k}, \mathbf{C}^{Al_2}_{t_k} \in S^{FS}_{t_k},\quad \alpha\mathbf{C}^{Al_1}_{t_k} + (1-\alpha)\mathbf{C}^{Al_2}_{t_k} \in S^{FS}_{t_k},$$
$$0 \le \alpha \le 1. \tag{15}$$

The utility function $\mathcal{U}_{b,t_k}$ is also convex, since for any $\mathbf{C}^{Al_1}_{t_k}, \mathbf{C}^{Al_2}_{t_k} \in S^{FS}_{t_k}$ and β in $0 \le \beta \le 1$, the following inequality holds [28]:

$$\begin{aligned}&\mathcal{U}_{b,t_k}\left(\beta \mathbf{C}_{t_k}^{Al_1} + (1-\beta)\mathbf{C}_{t_k}^{Al_2}\right)\\ &\quad \leq \mathcal{U}_{b,t_k}\left(\beta \mathbf{C}_{t_k}^{Al_1}\right) + \mathcal{U}_{b,t_k}\left((1-\beta)\mathbf{C}_{t_k}^{Al_2}\right).\end{aligned} \tag{16}$$

Since both $\mathcal{U}_{b,t_k}$ and $S_{t_k}^{FS}$ are convex, the pair $(S_{t_k}^{FS} \cup \{\mathbf{C}_{t_k}^{R_{\min}}, \mathcal{U}_{t_k}(\mathbf{C}_{t_k}^{Al})\})$ defines the bargaining problem for the computing resource allocation in a BBU-pool [9].

Moreover, in order to maintain QoS requirements, BBUs are assigned with bargaining powers related to the weight of their ongoing services. Services weights result from the normalization of Priority Level of services defined in 3GPP, [29], in the range of [1, 100]. The rationale behind it, is that the Priority Level is a characteristic by which 3GPP specifies the QoS requirements and determines the packet forwarding treatment. The weight of service s is

$$w_s^{srv} = 1 + \frac{99\left(P_{\max}^{srv} - P_s^{srv}\right)}{\left(P_{\max}^{srv} - P_{\min}^{srv}\right)} \tag{17}$$

where:

- P_s^{srv}: the Priority Level of the service s, given by 3GPP,
- $P_{\min,\max}^{srv}$: the minimum/maximum of 3GPP service Priority Levels,
- 99 is used as normalization factor, being the difference between the maximum and the minimum parameters values.

Accordingly, the average weight of ongoing services in a BBU is denoted as

$$\overline{w_{b,t_k}^{srv}} = \frac{\sum_{s=1}^{N^{srv}} N_{b,s,t_k}^{U} w_s^{srv}}{N_{b,t_k}^{U}} \tag{18}$$

where:

- N_{b,s,t_k}^{U}: number of users of service s in BBU b at t_k,
- N_{b,t_k}^{U}: total number of users in BBU b at t_k.

Finally, the combination of BBUs' RCCs and the average weight of services defines the BBU bargaining powers as

$$B_{b,t_k} = \frac{\overline{w_{b,t_k}^{srv}}\left(C_{b,t_k}^{R} - C_{b,t_k}^{R_{\min}}\right)}{\sum_{l=1}^{N_B} \overline{w_{l,t_k}^{srv}}\left(C_{l,t_k}^{R} - C_{l,t_k}^{R_{\min}}\right)}. \tag{19}$$

A BBU bargaining power is a positive value within [0, 1], and the sum of all BBU bargaining powers in a time instant is always equal to one,

$$\sum_{b=1}^{N_B} B_{b,t_k} = 1. \tag{20}$$

Equation (19) implies that once the minimum guaranteed RCC is allocated to BBUs, i.e., $C_{b,t_k}^{R_{\min}}$, the rest of the resources are distributed such that QoS is maintained. In order to maintain QoS, services with a higher priority should be allocated with more resources. In this context, in the next sections, maintaining the QoS is equivalent to BBU prioritization based on service weights.

```
Input: $\mathbf{C}_{t_k}^{R}, \mathbf{C}_{t_k}^{R_{\min}}, \overline{\boldsymbol{w}_{t_k}^{srv}}, C_{BP\ t_k}^{Av}$
Output: $\mathbf{C}_{t_k}^{Al^*}$
1:  for $b$=1 to $N_B$ do
2:      $B_{b,t_k} \leftarrow \left(\overline{w_{b,t_k}^{srv}}\left(C_{b,t_k}^{R} - C_{b,t_k}^{R_{\min}}\right)\right) / \sum_{l=1}^{N_B}\left(\overline{w_{l,t_k}^{srv}}\left(C_{l,t_k}^{R} - C_{l,t_k}^{R_{\min}}\right)\right)$
3:  end for
4:  if $\sum_{b=1}^{N_B} C_{b,t_k}^{R} \leq C_{BP\ t_k}^{Av}$
5:      $\mathbf{C}_{t_k}^{Al^*} \leftarrow \mathbf{C}_{t_k}^{\boldsymbol{R}}$
6:  else
7:      $\mathbf{C}_{t_k}^{Al^*} \leftarrow \underset{\forall \mathbf{C}_{t_k}^{Al} \in S_{t_k}^{FS} \cup \{\mathbf{C}_{t_k}^{R_{\min}}\}}{\operatorname{argmax}} \mathcal{U}_{BP}\left(\mathbf{C}_{t_k}^{Al}\right)$
8:  end if
9:  return $\mathbf{C}_{t_k}^{Al^*}$
```

FIGURE 3. Algorithm for QoS-demand-aware computing resource allocation in the BBU-pool at time instant t_k.

B. GENERALIZED NASH BARGAINING SOLUTION

By modeling the BBU-pool computing resource allocation as a bargaining game, the GNBS can be used as the unique fair Pareto optimal solution among all feasible ones existing in $S_{t_k}^{FS}$. GNBS satisfies Nash axioms as the attributes that any rational solution should meet to come up with fairness and efficiency, and is achieved by maximizing the product of the BBU utility functions weighted by the BBU bargaining powers [9]. By defining $\mathcal{U}_{BP}(\mathbf{C}_{t_k}^{Al})$ as the utility function of the BBU-pool,

$$\mathcal{U}_{BP}\left(\mathbf{C}_{t_k}^{Al}\right) = \prod_{b=1}^{N_B}\left(\mathcal{U}_{b,t_k}\left(\mathbf{C}_{t_k}^{Al}\right) - \mathcal{U}_{b,t_k}\left(\mathbf{C}_{t_k}^{R_{\min}}\right)\right)^{B_{b,t_k}}. \tag{21}$$

GNBS provides a unique solution $\mathbf{C}_{t_k}^{Al^*}$ for the defined bargaining game by solving the following optimization problem:

$$\mathbf{C}_{t_k}^{Al^*} = \underset{\forall \mathbf{C}_{t_k}^{Al} \in S_{t_k}^{FS} \cup \left\{\mathbf{C}_{t_k}^{\boldsymbol{R}_{\mathbf{min}}}\right\}}{\operatorname{argmax}} \mathcal{U}_{BP}\left(\mathbf{C}_{t_k}^{Al}\right). \tag{22}$$

For clarity, the process of the computing resource allocation in a BBU-pool is shown in Fig. 3. Given the BBUs' RCCs, minimum guaranteed RCCs, the average weight of ongoing services and the BBU-pool AvCC as inputs, all BBU bargaining powers are calculated in the first step, line 2. In the case that the total resource demand is less than or equal to the AvCC, all BBUs are allocated with the computing resources fulfilling their demands, line 5, otherwise, GNBS is achieved as an optimal compromise solution by solving (22), line 7.

In order to solve (22), first the following optimization problem is put forward:

$$\underset{C_{t_k}^{Al}}{\text{maximize}}\quad \mathcal{U}_{BP}\left(C_{t_k}^{Al}\right) \tag{23a}$$

$$\text{subject to}\quad \sum_{b=1}^{N_B} C_{b,t_k}^{Al} \leq C_{BP\ t_k}^{Av} \tag{23b}$$

$$C_{b,t_k}^{R_{\min}} < C_{b,t_k}^{Al} \tag{23c}$$

$$C_{b,t_k}^{Al} \leq C_{b,t_k}^{R} \tag{23d}$$

where (23b) to (23d) take the constraints given in $S_{t_k}^{FS}$ into account. The objective function is then transformed to the logarithmic form in order to facilitate the solving of the problem. Due to the monotonic behavior of the logarithm function, the logarithm of $\mathcal{U}_{BP}(\mathbf{C}_{t_k}^{Al})$, does not change the result [28]; thus, the objective function can be rewritten as:

$$\mathcal{U}_{LBP}\left(\mathbf{C}_{\mathrm{t_k}}^{Al}\right) = \sum_{b=1}^{N_B} B_{b,t_k} \log\left(C_{b,t_k}^{Al} - C_{b,t_k}^{R_{\min}}\right) \tag{24}$$

$\mathcal{U}_{LBP}(\mathbf{C}_{\mathrm{t_k}}^{Al})$ tends to $-\infty$ when C_{b,t_k}^{Al} approaches $C_{b,t_k}^{R_{\min}}$, hence, constraint (23c) is automatically satisfied. It can then be relaxed and the optimization problem for $b = 1, 2, \ldots, N_B$ is rewritten as:

$$\underset{C_{t_k}^{Al}}{\text{maximize}} \quad \mathcal{U}_{LBP}\left(C_{t_k}^{Al}\right) \tag{25a}$$

$$\text{subject to} \quad \sum_{b=1}^{N_B} C_{b,t_k}^{Al} \le C_{BP\ t_k}^{Av} \tag{25b}$$

$$C_{b,t_k}^{Al} \le C_{b,t_k}^{R}. \tag{25c}$$

Equation (25) is convex, since all constraints are linear inequalities, and the objective function is the sum of the concave functions [28], therefore, there is a unique optimal solution for (25) and it can be solved by CVX [30] (a modeling system for constructing and solving disciplined convex programs, developed by Stanford University), converging to the global optimum efficiently. One can find a detailed discussion on solving the problem in [31] with linear time complexity in the order of $O(N_B)$.

C. OTHER MODEL'S APPROACHES

In order to evaluate the performance of the proposed model, other resource allocation schemes found in the literature were also implemented, hence enabling a comparison, the equal and demand-proportional allocation approaches having been taken, [32]–[34]:

- *Equal Resource Allocation Scheme (EAS):* a simple resource allocation scheme that equally distributes computing resources among the BBUs, regardless of BBU demands and active services' priority, leading to

$$C_{b,t_k}^{Al_{EAS}} = \frac{C_{BP\ t_k}^{Av}}{N_B}. \tag{26}$$

- *Demand-Proportional Resource Allocation Scheme (DAS):* an allocation scheme ensuring minimum guaranteed demands to each BBU and distributing the remaining resources among them proportionally to their user processing demands, leading to

$$C_{b,t_k}^{Al_{DAS}} = \left(C_{BP\ t_k}^{Av} - \sum_{l=1}^{N_B} C_{l,t_k}^{R_{\min}}\right) \frac{C_{U\ b,t_k}^{R}}{\sum_{l=1}^{N_B} C_{U\ l,t_k}^{R}} + C_{b,t_k}^{R_{\min}} \tag{27}$$

DAS is more complex than EAS, since BBU demands should be achieved prior to resource provisioning.

These resource allocation schemes do not provide any optimization, still they serve as a good comparison.

VI. PERFORMANCE METRICS

As explained before, the aim of resource management is to enhance resource utilization while maintaining QoS, for which resource allocation should uphold the priority of ongoing services. Two metrics are defined to assess the performance of the proposed model, and in addition, another enables to compare the efficiency of the proposed model with the one of the other approaches:

- *Resource Usage:* a value within [0, 100]% indicating the proportion of BBU-pool AlCC that are used for signal processing to the existing computing capacity, given by

$$U_{t_k[\%]} = \frac{\sum_{b=1}^{N_B} \min\left\{C_{b,t_k[\mathrm{GOPS}]}^{Al}, C_{b,t_k[\mathrm{GOPS}]}^{R}\right\}}{C_{BP[\mathrm{GOPS}]}} 100 \tag{28}$$

where C_{BP} is the BBU-pool existing resources; higher values of U_{t_k} indicate a lower resource wastage, i.e., a larger portion of the existing computing capacity is used.

- *BBU Fulfilment Level:* a value within [0, 1] measuring the fraction of UP RCC of BBU b that is satisfied without any processing delay, given by

$$f_{b,t_k}^{B} = \frac{\min\left\{C_{b,t_k[\mathrm{GOPS}]}^{Al}, C_{b,t_k[\mathrm{GOPS}]}^{R}\right\} - C_{b,t_k[\mathrm{GOPS}]}^{R_{\min}}}{C_{b,t_k[\mathrm{GOPS}]}^{R} - C_{b,t_k[\mathrm{GOPS}]}^{R_{\min}}} \tag{29}$$

higher values of f_{b,t_k}^{B} indicate that a larger portion of the BBU UP demands is met.

- *Dynamic Resource Allocation Efficiency:* a value within [0, 100]% comparing the BBU-pool AlCC resulting from the proposed model with the one from the comparison approach, in which BSs are provisioned for peak load, given by

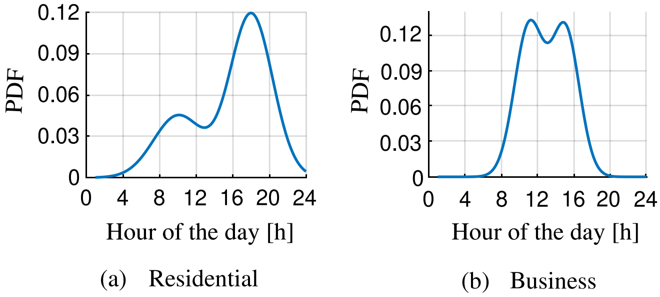
$$\eta_{t_k[\%]} = \left(1 - \frac{\sum_{b=1}^{N_B} C_{b,t_k[\mathrm{GOPS}]}^{Al}}{\sum_{b=1}^{N_B} C_{b[\mathrm{GOPS}]}^{R_{PEAK}}}\right) 100 \tag{30}$$

where $C_b^{R_{PEAK}}$ is the BBU b peak hour RCC; following (9), a full 100% usage of the RBs bandwidth, while users' modulation and coding schemes are at the highest level, leads to the peak level of a BBU RCC, hence, the higher is the value of η_{t_k}, the more efficient is the proposed resource provisioning scheme.

VII. ANALYSIS OF RESULTS

A. REFERENCE SCENARIO

This section presents the scenario setup. It includes 4 micro-cell RRHs in a residential area and other 3 in a business



(a) Residential (b) Business

FIGURE 4. User arrival rate for residential and business areas.

TABLE 3. Service characteristics.

Service	Service Parameter		Distribution	Mean	Std. Dev.
VoIP	Packet Inter-Arrival Time		Deterministic	20 ms	-
	Duration		Poisson	120 s	11 s
Video	Frame Packets Inter-Arrival Time		Pareto	6.1 ms	3.6 ms
	Duration		Poisson	300 s	17.3 s
	Packet Volume		Pareto	1.3 MB	257 B
File Transfer	File Size		Lognormal	2 MB	700 B
E-Mail				1.3 MB	380 B
Web Browsing	Packet Inter-Arrival Time	Reading Time	Exponential	30 s	
		Parsing Time		130 ms	
	Duration		Poisson	420 s	20.5 s
	Packet Volume	Main Object Size	Lognormal	11 MB	25.3 MB
		Embedded Objects Size		8.2 MB	47.3 MB
		Number of Embedded Objects / Page	Pareto	7.6	10.4

TABLE 4. Traffic mixture.

Service ID	Service Weight	Service Penetration [%] / BBU Index (b)					
		VoIP dominance (RV, BV)		File transfer dominance (RF, BF)		Without service dominance (RM1, RM2, BM)	
		DL	UL	DL	UL	DL	UL
VoIP	83	60	71	2	3	15	18
Video Calling	59	1	1	1	1	1	1
Video Streaming	48	1	-	1	-	1	-
File Transfer	36	22	26	80	94	67	79
E-mail		2	2	2	2	2	2
Web Browsing		14	-	14	-	14	-

one, the RRHs being driven by 7 instances of BBUs, co-located in a single BBU-pool, where each BBU instance in the pool is associated with a single RRH. All BSs are configured with down- and uplink bandwidths of 40 MHz, 24 bit quantization resolution and support for 8 × 8 MIMO.

On the user side, the equipment is a smartphone with 8 spatial streams. The user's Signal-to-Noise-Ratio (SNR) is represented by a random variable taken uniformly in [1, 35] dB at each time instant; accordingly, the modulation and coding ratio used to serve the user is extracted from [35]. One should note that 1024 QAM is assumed to be the highest modulation offered by the network, leading to a BBU peak RCC of 4.8 TOPS for the proposed scenario, based on (9). The aforementioned parameters are required for BBUs' RCC estimations based on (6) and (8).

The user arrival rate follows a mixture of two truncated normal distributions for both residential and business areas: for the former, the mean values are at 10 AM and 6 PM, standard deviations are 160 min and 140 min, and mixing proportions are 30% and 70% for the first and second distribution, respectively; in the latter mean values are at 11 AM and 3 PM for the first and second distribution respectively, both with the standard deviation of 95 min and 50% of mixing proportion, Fig. 4. User peak hours for residential and business areas are taken from [36] and the traffic outside peak hours is selected in such a way that it gradually increases until the peak and then decreases.

To generate traffic demand, an attempt has been done to emulate a typical day of operation in cellular networks, however, due to hardware limitations, this resulted in a too long simulation time, hence, only 10 minutes of network time was simulated, starting at 6 PM (one of the peaks), with a time granularity of 1 ms. The simulation includes a combination of heterogeneous services, i.e., VoIP, video calling/streaming, file transfer, email and Web browsing. The 5 types of services was chosen according to the estimation that, until 2025, more than 90% of mobile traffic will be composed of the proposed service mix [37] (social networking and software down- and upload are considered as file transfer).

For simplicity, it is assumed that users request only one type of service at a time. Service durations are randomly generated for VoIP, video calling/streaming and Web browsing, based on Poisson distributions with the mean values of 120 s, 300 s and 420 s, respectively. File transfer and email service durations, however, rely on the user file size and network data rate. Moreover, traffic generation is done at the packet level, where packet size and flow are characterized by stochastic models defined exclusively for each service [38]. According to the user modulation and coding ratio, the number of RBs that are required to transfer the generated packet is extracted from [35]. Service characteristics are summarized in TABLE 3.

The traffic mixture per cell is summarized in TABLE 4, profiles with a dominance of VoIP (V) or File transfer (F) and Mixed without dominance (M) being used. The traffic mixture in TABLE 4 was designed so that each BBU has a different type of service as the highest ratio of running service, so that one can analyze model performance in allocating resources based on service priorities.

The BBUs might serve RRHs being in a Residential (R) or Business (B) areas. BBU names in TABLE 4 denote both the area location and service dominance:

- *RV:* Residential area with VoIP dominance,
- *BV:* Business area with VoIP dominance,
- *RF:* Residential area with File transfer dominance,
- *BF:* Business area with File transfer dominance,

- *RM1/RM2:* Residential area without service dominance (Mix) (area location and traffic mixture are considered the same for these BBUs, the goal being to analyze the model behavior for BBUs with equal conditions),
- *BM:* Business area without service dominance (Mix).

Regarding the BBU-pool, the centralization of BBUs' computing capacity, which are provisioned for peak demands, i.e., 4.8 TOPS per BBU, together with an extra 17% for signaling overhead and saturation prevention, results in a BBU-pool with 39 TOPS computing capacity. However, in order to analyze the impact of computing capacity on BBU fulfilment levels, resource usage and efficiency, the resource allocation phase is repeated with the computing capacity of BBU-pool taken in [0.2, 39] TOPS. Each run is done for the 10 minutes simulated traffic, and model performance is assessed.

The evaluation of the proposed model is also done by comparing its performance against EAS and DAS. To this end, 8.5 TOPS is taken as the existing resources in the BBU-pool and the evaluation is done accordingly. The maximum resources that all BBUs are allowed to utilize in all experiments is 83%, in order to avoid data-center saturation.

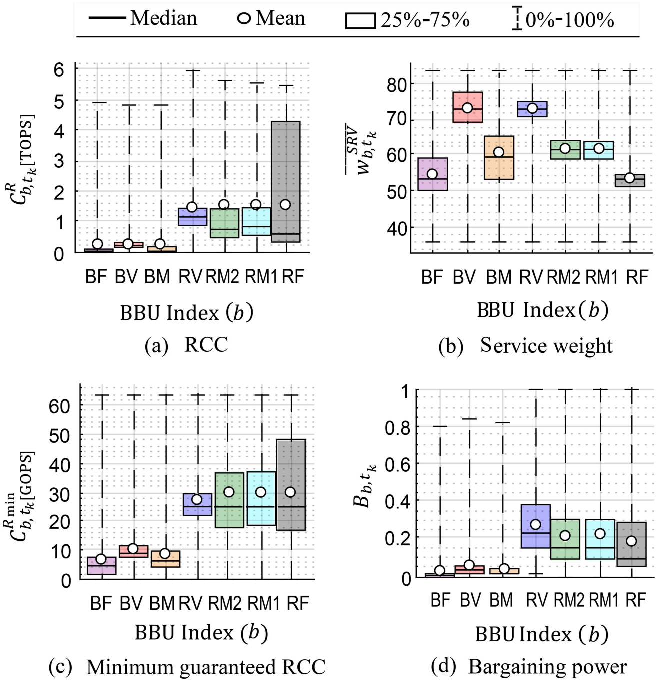
It should also be noted that all simulations were implemented in MATLAB on a desktop PC with Intel®Core^{TM} i3-4150 3.50 GHz with a two-core processor and 8 GB of memory, in which CVX runs for 6 ms on average, to find the solution of (25) as the optimal resource allocation in a given time instant of the simulated scenario.

B. RCC ESTIMATION RESULTS

The average of BBUs' RCCs, ongoing service weights, minimum guaranteed RCCs, and bargaining powers were computed with the proposed model for the given scenario. Simulation results are shown in Fig. 5, where BBUs are sorted by demand, i.e., from the lowest to the highest one.

Fig. 5(a) presents a higher RCC for residential BBUs compared to business ones, since the chosen scenario has a residential peak traffic demand leading to a higher number of active users, hence, a higher RCC for serving BBUs. For the same reason, the minimum guaranteed RCC of BBUs follows a similar pattern, Fig. 5(c); the values are relatively smaller, since the minimum guaranteed RCC accounts only for the CP processing steps' demands.

Regardless of the RRH type, simulation results show a reasonably equal average service weights for BBUs with the same traffic mixture, Fig. 5(b). In addition, since VoIP has the highest service weight, BBUs RV and BV, with the highest proportion of VoIP, have the highest average among all other BBUs. Despite the same average of service weights, BBU RV has a bargaining power higher than the BBU BV one, Fig. 5(d), which is due to the fact that it is a function of both service weights and RCC, thus, an unequal RCC may lead to different BBU bargaining powers.



(a) RCC

(b) Service weight

(c) Minimum guaranteed RCC

(d) Bargaining power

FIGURE 5. BBU demands, service weights, and bargaining powers variations in the simulated scenario from 6:00 PM to 6:10 PM.

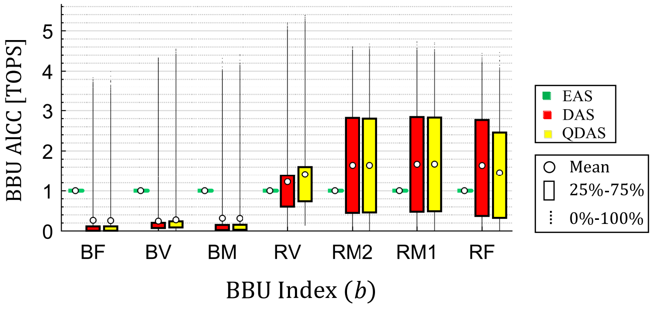


FIGURE 6. AlCC in different allocation schemes.

C. COMPARE WITH OTHER MODELS

This section compares the performance of the proposed model, QDAS, with the two other reference ones, EAS and DAS, in terms of resource usage and BBU fulfilment level. To this end, the resource allocation phase is repeated for each of the allocation schemes separately over 10 minutes of simulated network traffic. To narrow down the comparison, only the instances in which the total demand in the BBU-pool exceeds the available capacity are taken.

Fig. 6 compares BBUs' AlCCs in the three allocation schemes. EAS allocates resources equally among BBUs, regardless of service priorities or BBU demands. Although EAS is a fast resource allocation scheme without too much complexity, it leads to a waste of resources if the BBU demand is less than its share. In such cases, some allocated resources are unused while a neighboring BBU may experience shortage.

In contrast, DAS takes real-time demand of the BBUs into account. It allocates the minimum guaranteed resources,

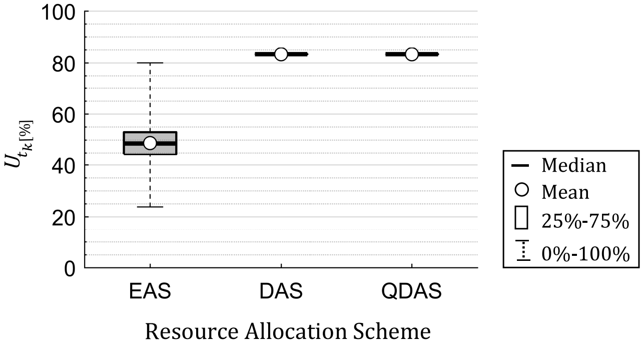


FIGURE 7. Resource usage in different resource allocation schemes.

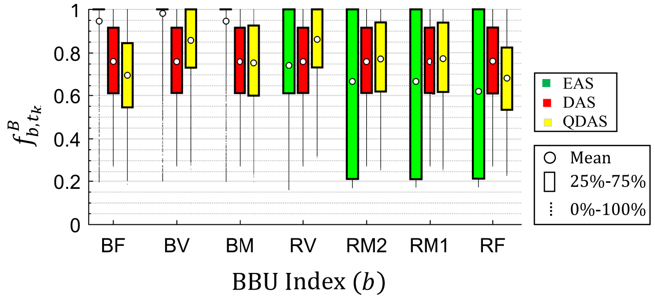


FIGURE 8. BBU fulfilment levels in different allocation schemes.

$C^{R_{\min}}_{b,t_k}$, to each BBU and distributes the remaining resources proportionally to their user processing requirements, $C^{R}_{U\ b,t_k}$. As a result, no BBU encounters a resource shortage in this scheme, while its neighboring BBUs are underutilized. For the same reason, as presented in Fig. 6, BBUs in the residential area, with higher demands, receive more resources than in the business one. Moreover, BBU RF and BF receive the highest and the lowest amount of resources, since they have the highest and the lowest demand among all BBUs, respectively, Fig. 5(a).

Similar to DAS, QDAS takes BBU demands into account, hence, resource allocation follows a similar pattern. However, the difference between these two approaches stems from the fact that, in addition to BBUs' demands, QDAS takes QoS, hence, service priorities, into account, thus, QDAS allocates more (less) resources to BBUs with higher (lower) average service weights, compared with DAS. The effect of service priority is apparent when comparing the AlCC of BBUs RV and RF in DAS with the one that QDAS assigns to them. As one can see in Fig. 6, DAS allocates 1.23 TOPS to RV while QDAS increases its AlCC to 1.41 TOPS, which is almost 15% more; in contrast, QDAS decreases the resources allocated to RF by 12% (from 1.64 TOPS to 1.45 TOPS) compared to DAS, since its services are not as critical as the ones in RV.

The overall resource usage is presented in Fig. 7. It can vary in the range of [0, 83]%, depending on the available resources being fully used or not.

Due to the dynamicity of the network, BBUs' demands fluctuate over time, hence, the BBU-pool's total demand may be less, equal, or more than the available resources at a given time instant. In the event that the total demand surpasses the available resources and none of the BBU's allocated resources exceed its demand, the available resources are fully utilized, hence, there is no wastage. In contrast, wastage may happen in two circumstances: when the available resources exceed the sum of all BBUs' demands, irrespective of the allocation policy; or, when the available resources are less or equal than the total demand, but a poor allocation policy distributes more resources to one (or more) BBUs than their demand.

The low resource usage in EAS, Fig. 7, is an example of resource wastage in the second circumstances, since it distributes resources evenly, regardless of the BBUs' demands, resulting that business BBUs are underutilized while residential ones are overloaded. On the other hand, DAS and QDAS take BBUs' demands into account, thus, resources are fully utilized in both, since none of allocated resources exceed their demands.

Fig. 8 illustrates that DAS fulfils all BBU's demands equally, irrespective of the priority of ongoing services, therefore, the resource allocation is not fair in the case of a shortage, because BBUs running critical services, i.e., services with lower delay budget and higher priorities [39], require more resources to keep up with QoS.

In contrast, QDAS supports QoS, so, BBUs with higher service priorities have higher fulfilment levels. One can see the effect of service priority by comparing BBUs RF and BV: although the RCC of RF is much higher than BV's, Fig. 5(a), its fulfilment level is smaller than BV since it has a lower average of ongoing service weight, Fig. 5(b). Moreover, BBUs RV and BV have the highest fulfilment level among all, as their services have the highest weights on average. By comparing with DAS, it is also apparent that QDAS shows a higher performance and increases the fulfilment level of BBUs RV and BV, by 13%, for the same reason.

Fig. 8 also shows that EAS fulfils more BBU demands in business areas than in the residential areas, given the uniform resource allocation. This is an example of resource wastage, because for BBUs in business areas, the demand is often less than the allocated resources, while, at the same time, BBUs in residential areas run into resource shortage, the outcome being a high (low) fulfilment level for the BBUs in the business (residential) areas.

D. RESOURCE ALLOCATION RESULTS

In this section the impact of computing capacity on BBU fulfilment levels, resource usage and efficiency is analyzed. To this end, the resource allocation phase is repeated with the BBU-pool computing capacity, C_{BP}, being increased from 0.2 to 39 TOPS.

Fig. 9 shows the capacity share of BBUs. BBUs with higher bargaining powers, i.e., higher priorities, are allocated with more resources in the presence of a resource shortage, i.e., before threshold Th_1 in Fig. 9. One can see the effect of the bargaining power by comparing BBUs RF and RM1:

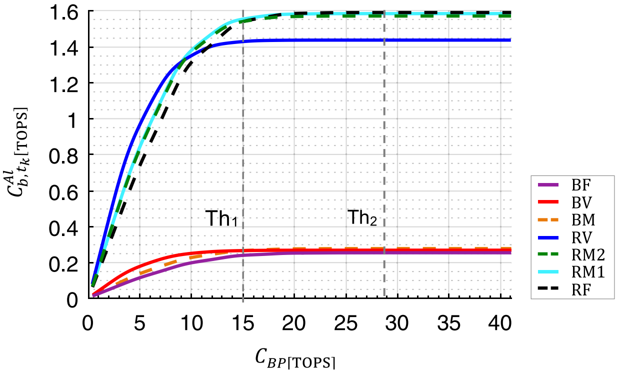


FIGURE 9. Average of the BBUs' AlCC.

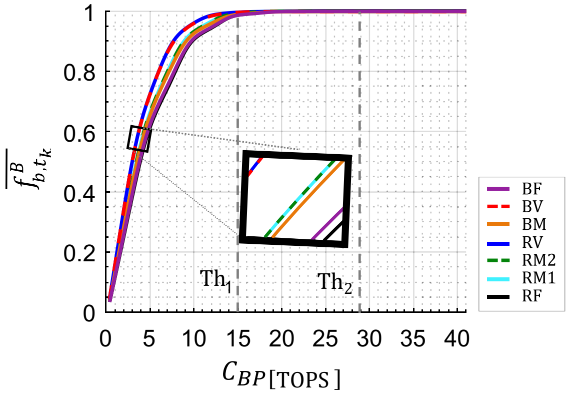


FIGURE 10. Average of the BBU fulfilment levels.

although their RCCs have similar mean values, Fig. 5(a), RM1 is allocated with more computing capacity before Th$_1$, since it has a higher bargaining power, hence, higher priority, while the computing resources are being allocated to BBUs. The resource allocator shrinks the capacity share of the lower priority BBUs in order to compensate for the higher priority BBU resource shortage. Beyond Th$_2$, 100% of BBU requests are served, since the available resources are more than the overall demand.

The impact of BBU-pool capacity variations on the fulfillment level of BBUs in the pool is presented in Fig. 10. Regardless of demand, BBUs with higher service priorities account for higher fulfillment levels in the presence of a resource shortage. Moreover, the fulfilment levels for BBUs with a similar average of service weights are equal, since the proposed resource allocator keeps BBU AlCCs proportional to the weight of their ongoing services. One can see the effect of the service weights by comparing BBUs RV and BV: although the RCC of RV is much higher than BV's, Fig. 5(a), both are fulfilled fairly equally. These BBUs also have the highest fulfilment level among all others, since they have the highest average of service weights.

One should also note that although increasing C_{BP} improves the average fulfillment level, as shown in Fig. 10, correlation is not linear. For instance, when C_{BP} is doubled from 15 to 30 TOPS, the average fulfilment level is improved by only 2%, from 98% to near 100%. This becomes more important when the same boost in C_{BP} incurs a near 20% drop in average resource usage, as depicted in Fig. 11. This behavior indicates that cloud providers should carefully consider the trade-off between BBU fulfilment levels and resource usage. An idea to decrease resource wastage, can be to reduce the available computing capacity in the BBU-pool, while degrading the capacity share of the delay-tolerant services in the BBU, to compensate for real-time services resource shortage.

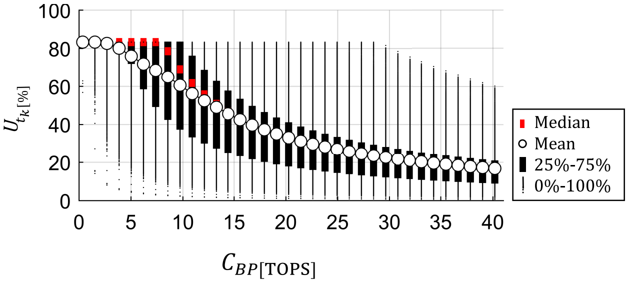


FIGURE 11. Resource usage.

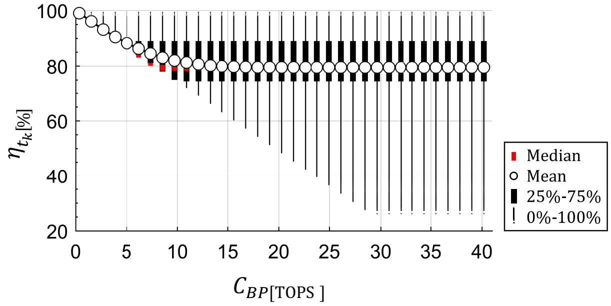


FIGURE 12. Resource allocation efficiency.

Fig. 11 also shows that due to the severe resource shortage in the beginning, when C_{BP} is small, the available resources of the BBU-pool are basically entirely allocated among BBUs. By increasing C_{BP}, the resources usage decreases: due to the dynamicity of the network, BBUs' demands fluctuate over time, leading to situations where, in some time instants, the total demand is less than the available resources, in these cases resources not being fully utilized, since the allocator bounds the BBU AlCCs to their real-time demands. When C_{BP} increases further, more resources remain unused, and hence, resources usage drops.

The efficiency of the proposed resource allocation model is another important metric, being in Fig. 12. The average efficiency of the pool decreases for an increasing C_{BP}, the decline being faster in the beginning when C_{BP} increases from 0.2 to 10 TOPS: in this range, there is a resource shortage, so the available resources are instantly allocated, the direct outcome being the decline in efficiency as more resources become available in the beginning. Once the requirements of BBUs are fully met, and there is no more shortage, the allocator stops assigning more resources to BBUs (due to the allocation strategy). Resources that

TABLE A. Used notations.

Notations	Definition
Δf_{BW}	Bandwidth
η_{t_k}	Resource allocation efficiency at t_k
η^{RB}	RB efficiency
$\eta_{u,t_k}^{RB_U}$	User u RB efficiency
$\eta_{b,t_k}^{RB_B}$	BBU b RB efficiency
B_{b,t_k}	Bargaining power of BBU b at t_k
C_{b,t_k}^{Al}	BBU b AlCC at t_k in QDAS
$C_{b,t_k}^{Al_{DAS}}$	BBU b AlCC at t_k in DAS
$C_{b,t_k}^{Al_{EAS}}$	BBU b AlCC at t_k in EAS
$\mathbf{C}_{t_k}^{Al}$	Resource allocation strategy vector at t_k
$C_{BP\ t_k}^{Av}$	AvCC of BBU-pool at t_k
C_{BP}	Existing resources of BBU-pool
C_{b,t_k}^{R}	Total RCC of BBU b at t_k
$C_{U\ b,t_k}^{R}$	Total users processing RCC of BBU b at t_k
$\mathbf{C}_{t_k}^{R_{\min}}$	Minimum guaranteed RCC vector at t_k
$C_{b,t_k}^{R_{\min}}$	Minimum guaranteed RCC of BBU b at t_k
$C_{b}^{R_{PEAK}}$	BBU b peak hour RCC
$C_{b,p,t_k}^{R_{CP}}$	RCC of CP step p of BBU b at t_k
$C_{u,p,t_k}^{R_{UP}}$	RCC of UP step p of user u at t_k
C_{p}^{ref}	Reference RCC of processing step p
$E_{\eta^{RB},p}$	Scaling exponents of RB efficiency on processing step p
$E_{x,p}$	Scaling exponents of parameter x on processing step p
f_{b,t_k}^{B}	Fulfilment level of BBU b at t_k
m_u	User u modulation
N_B	Number of BBUs in the BBU-pool
N_{MIM}	MIMO order
$N_{\Delta f}^{RB}$	Total number of sub-frame RBs in a given bandwidth
N_{u,t_k}^{RB}	Number of allocated RBs to user u at t_k
N^{srv}	Number of service types
N_{u}^{Str}	User u number of streams
N_{b,t_k}^{U}	Total number of active users in BBU b at t_k
N_{b,s,t_k}^{U}	Number of active users of service s in BBU b at t_k
P^{chc}	Channel coding processing
P^{chd}	Channel decoding processing
P^{che}	Channel estimation processing
P^{dm}	Demodulation processing
P^{mpc}	MIMO pre-coding processing
P^{md}	Modulation processing
P^{mdc}	MIMO decoding processing
P^{OFDMA}	OFDMA processing
P^{SCFDMA}	SC-FDMA processing
P_{s}^{srv}	Priority level of service s
$P_{\max\|\min}^{srv}$	Minimum/maximum of 3GPP service priority levels
Q	Quantization resolution
r_u	User u coding ratio
S^{CP}	Common processing set
$S_{t_k}^{FS}$	Feasible solution set
S_{b,t_k}^{U}	Set of all active users in BBU b at t_k
S^{UP}	User processing set
t_k	Time instant k
X	Set of effective parameters on signal processing complexity
x^{ref}	Reference value of parameter x
x_{t_k}	Value of parameter x in the operating scenario at t_k
U_{t_k}	Usage of the BBU-pool resources at t_k
$\boldsymbol{\mathcal{U}}_{t_k}$	Utility vector at t_k
$\mathcal{U}_{b,t_k}$	Utility of BBU b at t_k
$\mathcal{U}_{BP}$	Utilization of the BBU-pool resources
$\mathcal{U}_{LBP}$	Logarithmic form of utilization of BBU-pool resources
w_{s}^{srv}	Weight of service s
$\overline{w_{b,t_k}^{srv}}$	Average weigh of ongoing services in BBU b at t_k

become available afterwards, remain un-allocated and efficiency drops slower beyond 10 TOPS. However, the average efficiency never drops below 83%, which is the total demand divided by the sum of separate peak demand of BBUs.

VIII. CONCLUSION

In this article, a real-time resource management model is proposed, to maximize the computing resource utilization of a C-RAN BBU-pool. To this end, the allocation of resources among the BBUs is modeled as a game-theory bargaining problem and a Generalized Nash Bargaining Solution, with adaptive bargaining powers, is applied to find the optimal solution. In the event of resource shortage, where there are unsatisfied requests, the model prioritizes BBUs according to the QoS requirements of the services being processed. BBU priorities are indicated by a mixture of ongoing service priorities and real-time demands. Low-priority BBUs are always guaranteed a minimum computing resource to avoid them from crashing.

The performance of the proposed model is evaluated in terms of resource usage, BBU fulfilment level and efficiency over time. To this end, a scenario is defined in which a BBU-pool includes seven BBUs offering heterogeneous services with tidal traffic flows. Results confirm that resources provided to the BBUs are consistent with their real-time demands and proportional to the priority of ongoing services. Results also demonstrate that improving the average fulfilment level from 98% to 100% requires doubling the available resources at the cost of average resource usage being cut in half.

The evaluation of the proposed model's performance is also done by comparing results with two other schemes: equal and demand proportional resource allocation. The results confirm that the proposed model manages bottlenecks effectively and shows a higher performance, e.g., by increasing the fulfillment level of high prioritized BBUs by 13%.

In future work, the authors plan to improve the BBUs' prioritization policy by considering imposed delays to packets and expanding the work to a joint computing and radio resource management model.

APPENDIX LIST OF NOTATIONS

For the sake of clarity, all notations used in this article are listed TABLE A.

REFERENCES

[1] “Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022,” Cisco, San Jose, CA, USA, White Paper, Feb. 2019. [Online]. Available: https://www.cisco.com

[2] A. Checko *et al.*, “Cloud RAN for mobile networks—A technology overview,” *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.

[3] A. M. Abdalla, J. Rodriguez, I. Elfergani, and A. Teixeira, “Energy efficiency in the cloud radio access network (C-RAN) for 5G mobile networks,” in *Optical and Wireless Convergence for 5G Networks*. Hoboken, NJ, USA: Wiley, 2019.

[4] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, “Statistical multiplexing gain analysis of heterogeneous virtual base station pools in cloud radio access networks,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5681–5694, Aug. 2016.

[5] M. Dayarathna, Y. Wen, and R. Fan, “Data center energy consumption modeling: A survey,” *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 732–794, 1st Quart., 2016.

[6] M. F. Hossain, A. U. Mahin, T. Debnath, F. B. Mosharrof, and K. Z. Islam, “Recent research in cloud radio access network (C-RAN) for 5G cellular systems—A survey,” *J. Netw. Comput. Appl.*, vol. 139, pp. 31–48, Aug. 2019.

[7] T. Werthmann, H. Grob-Lipski, and M. Proebster, “Multiplexing gains achieved in pools of baseband computation units in 4G cellular networks,” in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, London, U.K., Sep. 2013, pp. 3328–3333.

[8] S. Pelley, D. Meisner, P. Zandevakili, T. F. Wenisch, and J. Underwood, “Power routing: Dynamic power provisioning in the data center,” in *Proc. 15th Int. Conf. Conf. Archit. Support Program. Lang. Oper. Syst. (ASPLOS)*, Pittsburgh, PA, USA, Mar. 2010, pp. 232–242.

[9] R. B. Myerson, *Game Theory: Analysis of Conflict*. Cambridge, MA, USA: Harvard Univ. Press, 1991.

[10] M. Barahman, L. M. Correia, and L. S. Ferreira, “An efficient QoS-aware computational resource allocation scheme in C-RAN,” in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Seoul, South Korea, Apr. 2020, pp. 1–6.

[11] M. Barahman, L. M. Correia, and L. S. Ferreira, “A real-time QoS-demand-aware computational resource sharing approach in C-RAN,” in *Proc.Eur. Conf. Netw. Commun. (EuCNC)*, Dubrovnik, Croatia, Jun. 2020, pp. 236–241.

[12] X. Wang *et al.*, “Energy-efficient virtual base station formation in optical-access-enabled cloud-RAN,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1130–1139, May 2016.

[13] K. Sundaresan, M. Y. Arslan, S. Singh, S. Rangarajan, and S. V. Krishnamurthy, “FluidNet: A flexible cloud-based radio access network for small cells,” *IEEE/ACM Trans. Netw.*, vol. 24, no. 2, pp. 915–928, Apr. 2016.

[14] 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 Trans. Cloud Comput.*, vol. 7, no. 2, pp. 403–414, Apr./Jun. 2019.

[15] M. Qian, W. Hardjawana, J. Shi, and B. Vucetic, “Baseband processing units virtualization for cloud radio access networks,” *IEEE Wireless Commun. Lett.*, vol. 4, no. 2, pp. 189–192, Apr. 2015.

[16] F. Zhang, J. Zheng, Y. Zhang, and L. Chu, “An efficient and balanced BBU computing resource allocation algorithm for cloud radio access networks,” in *Proc. IEEE Veh. Technol. Conf. (VTC-Spring)*, Sydney NSW, Australia, Jun. 2017, pp. 1–5.

[17] A. Younis, T. X. Tran, and D. Pompili, “Bandwidth and energy-aware resource allocation for cloud radio access networks,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6487–6500, Oct. 2018.

[18] W. Chien, C. Lai, and H. Chao, “Dynamic resource prediction and allocation in C-RAN with edge artificial intelligence,” *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 4306–4314, Jul. 2019.

[19] C. Clark *et al.*, “Live migration of virtual machines,” in *Proc. Symp. Netw. Syst. Design Implement. (NSDI)*, Boston, MA, USA, May 2005, pp. 273–286.

[20] *RAN Evolution Project Comp Evaluation and Enhancement*, (CoMP Work Stream, RAN Evolution Project, Final Deliverable), NGMN Alliance, Frankfurt, Germany, Mar. 2015. [Online]. Available: https://www.ngmn.org/wp-content/uploads/NGMN_RANEV_D3_CoMP_Evaluation_and_Enhancement_v2.0.pdf

[21] 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 J. Opt. Commun. Netw.*, vol. 9, pp. 489–497, 2017.

[22] D. Pompili, A. Hajisami, and T. X. Tran, “Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN,” *IEEE Commun. Mag.*, vol. 54, no. 1, pp. 26–32, Jan. 2016.

[23] N. Yu, Z. Song, H. Du, H. Huang, and X. Jia, “Dynamic resource provisioning for energy efficient cloud radio access networks,” *IEEE Trans. Cloud Comput.*, vol. 7, no. 4, pp. 964–974, Oct./Dec. 2019.

[24] 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 *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 5523–5528.

[25] M. Barahman, L. M. Correia, and L. S. Ferreira, “A fair computational resource management strategy in C-RAN,” in *Proc. Int. Conf. Broadband Commun. Next Gener. Netw. Multimedia Appl. (CobCom)*, Graz, Austria, Jul. 2018, pp. 1–6.

[26] *MAMMOET—Massive MIMO for Efficient Transmission*. Accessed: Mar. 2020. [Online]. Available: https://mammoet-project.eu

[27] B. Debaillie, C. Desset, and F. Louagie, “A flexible and future-proof power model for cellular base stations,” in *Proc. IEEE Veh. Technol. Conf. (VTC-Spring)*, Glasgow, Scotland, May 2015, pp. 1–7.

[28] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, West Nyack, NY, USA: Cambridge Univ. Press, 2004.

[29] *Technical Specification Group Services and System Aspects; Policy and Charging Control Architecture, V16.1.0*, 3GPP Standard TS 23.203, Sep. 2019.

[30] *CVX—Software for Disciplined Convex Programming*. Accessed: Mar. 2020. [Online]. Available: http://cvxr.com

[31] S. Khakurel, C. Leung, and T. Le-Ngoc, “A generalized water-filling algorithm with linear complexity and finite convergence time,” *IEEE Wireless Commun. Lett.*, vol. 3, no. 2, pp. 225–228, Apr. 2014.

[32] F. Fossati, S. Moretti, P. Perny, and S. Secci, “Multi-resource allocation for network slicing,” *IEEE/ACM Trans. Netw.*, vol. 28, no. 3, pp. 1311–1324, Jun. 2020.

[33] U. C. Kozat and A. C. K. Soong, “On the impact of slicing granularity on the availability and scalability of 5G networks,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–7.

[34] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, “Rate control for communication networks: Shadow prices, proportional fairness and stability,” *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, Apr. 1998.

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

[36] C. Peng, S. B. Lee, S. Lu, H. Luo, and H. Li, “Traffic-driven power saving in operational 3G cellular networks,” in *Proc. ACM MobiCom Int. Conf. Mobile Comput. Netw.*, Las Vegas, NV, USA, Sep. 2011, pp. 121–132.

[37] P. Cerwall *et al.*, “The Ericsson mobility report,” Ericsson, Stockholm, Sweden, Rep. EAB-20:004467 Uen, Jun. 2020. [Online]. Avalable: https://www.ericsson.com/49da93/assets/local/mobility-report/documents/2020/june2020-ericsson-mobility-report.pdf

[38] *Next Generation Mobile Networks Radio Access Performance Evaluation Methodology*, NGMN Alliance, Frankfurt, Germany, White Paper, Jan. 2008. [Online]. Available: https://www.ngmn.org/fileadmin/user_upload/NGMN_Radio_Access_Performance_Evaluation_Methodology.pdf

[39] *3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; System Architecture for the 5G System (5GS); Stage 2; (Release 16), V16.3.0*, 3GPP Standard TS 23.501, Dec. 2019.

MOJGAN BARAHMAN (Member, IEEE) received the M.Sc. degree in computer science from the Shahid Bahonar University of Kerman, Iran, in 2010. She is currently pursuing the Ph.D. degree with the Instituto Superior Tecnico, University of Lisbon. She was with the Azad University of Kerman as a Teaching and Research Assistant from 2010 to 2014. She is currently a Research Assistant with INESC-ID. Her research interests are wireless communication systems with a particular interest in cloud/edge radio access network and resource allocation. She was actively participating in the COST Action CA15104 (IRACON), to which she has contributed with several technical documents.

LUIS M. CORREIA (Senior Member, IEEE) was born in Portugal, in 1958. He received the Ph.D. degree in electrical and computer engineering from IST, University of Lisbon, in 1991, where he is currently a Professor in Telecommunications, with his work focused on wireless and mobile communications in the areas of propagation, channel characterization, radio networks, traffic, and applications, with the research activities developed in the INESC-ID Institute. He has acted as a consultant for the Portuguese telecommunications operators and regulator, besides other public and private entities, and has been in the Board of Directors of a telecommunications company. Besides being responsible for research projects at the national level, he has participated in 32 projects within European frameworks, having coordinated six and taken leadership responsibilities at various levels in many others. He has supervised more than 200 M.Sc./Ph.D. students, having edited six books, contribute to European strategic documents, and authored more than 500 papers in international and national journals and conferences, for which served also as a reviewer, a editor, and a board member. Internationally, he was part of 37 Ph.D. juries, and 68 research projects and institutions evaluation committees for funding agencies in 12 countries, and the European COST and Commission. He has been the Chairman of Conference, of the Technical Programme Committee and of the Steering Committee of various major conferences, besides other several duties. He was a National Delegate to the COST Domain Committee on ICT. He was active in the European Net!Works platform, by being an elected member of its Expert Advisory Group and of its Steering Board, and the Chairman of its Working Group on Applications, and was also elected to the European 5G PPP Association. He has launched and served as Chairman of the IEEE Communications Society Portugal Chapter.

LÚCIO STUDER FERREIRA (Senior Member, IEEE) received the Licenciado and Ph.D. degrees in electrical and computer engineering from the IST/Technical University of Lisbon, Portugal, in 1997 and 2013, respectively. He is a Lecturer, a Researcher, and a Project Manager working in the fields of wireless communications and computer science. As a Researcher, he worked with the Deutsche Telekom Innovation Laboratories, IST, Instituto de Telecomunicações, INOV-INESC, INESC-ID and Multivision Consulting. He participated in 17 projects within European frameworks as a Researcher and a Project Manager. As an Assistant Professor, he worked with the Universidade Lusiada de Lisboa, Universidade da Beira Interior, ISTEC, and Universidade Lusófona. He is/was supervisor of three Ph.D. and 14 M.Sc. thesis in IST, UBI, and ISCTE. He authored more than 60 papers in international and national journals and conferences, for which he also served as a reviewer, an editor, and a board member.