IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON RECENT ADVANCES ON RADIO ACCESS AND SECURITY METHODS IN 5G NETWORKS

Received May 7, 2017, accepted June 1, 2017, date of publication June 5, 2017, date of current version August 14, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2712138

Adaptive Resource Balancing for Serviceability Maximization in Fog Radio Access Networks

NHU-NGOC DAO[1], JUNWOOK LEE[1], DUC-NGHIA VU[1], JEONGYEUP PAEK[1], (Member, IEEE), JOONGHEON KIM[1], SUNGRAE CHO[1], KI-SOOK CHUNG[2], AND CHANGSUP KEUM[2]

[1]School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, Republic of Korea
[2]Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea

Corresponding author: Changsup Keum (cskeum@etri.re.kr)

This work was supported by the ETRI R&D Program through the Government of Korea, Development of Technologies for Proximity, Real-time, and Smart Service Recommendation Platform, under Grant 17ZH1510.

ABSTRACT Serviceability is the ability of a network to serve user equipments (UEs) within desired requirements (e.g., throughput, delay, and packet loss). High serviceability is considered as one of the key foundational criteria toward a successful fog radio access infrastructure satisfying the Internet of Things paradigm in the 5G era. In this paper, we propose an adaptive resource balancing (ARB) scheme for serviceability maximization in fog radio access networks wherein the resource block (RB) utilization among remote radio heads (RRHs) are balanced using the backpressure algorithm with respect to a time-varying network topology issued by potential RRH mobilities. The optimal UE selection for service migration from a high-RB-utilization RRH to its neighboring low-RB-utilization RRHs is determined by the Hungarian method to minimize RB occupation after moving the service. Analytical results reveal that the proposed ARB scheme provides substantial gains compared with the standalone capacity-aware, max-rate, and cache-aware UE association approaches in terms of serviceability, availability, and throughput.

INDEX TERMS Fog radio access network (F-RAN), mobile remote radio head (RRH), resource balancing, serviceability maximization, Hungarian method, backpressure algorithm.

I. INTRODUCTION

The harmonization between centralized cloud computing and distributed fog radio access networks (F-RANs) is considered a promising paradigm for fifth generation (5G) mobile systems [1]. Centralized cloud computing processes heavy operations in the base band unit (BBU) pool to provide high performance, while F-RANs, which geographically distribute multiple remote radio heads (RRHs), have been proposed to achieve high throughput, spectral efficiency, energy efficiency, as well as low latency maintenance [2], [3]. In F-RANs, high power nodes (HPNs) are deployed for wide-area coverage and perform control operations. On the upper side, the HPNs connect to the BBU pool via the backhaul links. On the peer side, HPNs are coordinated via standardized S1/X2 interfaces under the supervision of fog orchestrators. On the lower side, there are RRHs that are light-weight units which consists of multiple antennas and possible cache support (a.k.a, eRRH), and the fog orchestrator located at the HPNs manages the communication and radio resource allocation of the RRHs. The RRHs operate in the user plane and harmonize data traffic delivery for user equipments (UEs) with the BBU over the fronthaul links. In the F-RAN model, an UE can transmit/receive data to/from multiple RRHs simultaneously [4]; see Fig. 1. In order to evaluate a successful fog radio access infrastructure, serviceability is considered as one of the key foundational criteria. The serviceability in a mobile network is defined as *the ability of the network to serve UEs within the desired requirements (e.g., throughput, delay, and packet loss)*, referring to the concept of serviceability for cloudlets in [5]. To be more specific, the serviceability is the percentage of UEs that are served within the desired requirements by the network per the cumulative arrival of UEs during a specific time interval under pre-determined mean UE arrival and departure rates.

Although F-RANs can bring impressive performance in terms of their high data rate and low latency due to short-range dense deployment and pre-fetched cache of the fog-computing enabled RRHs [6]–[8], ability to support a superhigh connection density (up to 1,000,000 devices per square kilometer) following the international mobile telecommunications-2020 (IMT-2020) standard requirements specified for 5G mobile systems issued by the

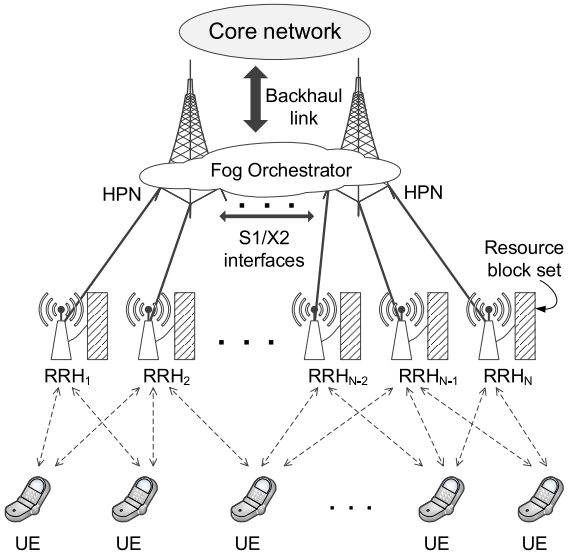


FIGURE 1. System architecture of F-RAN: Fog orchestrator located at the high power nodes (HPNs) manages local remote radio heads (RRHs) in terms of communication and radio resource allocation.

international telecommunication union (ITU) [9] remains an open challenge. Moreover, since the max-rate algorithms and cache-aware algorithms are mostly preferred for the user association of RRHs [10], [11], the *interesting* RRHs (in the UE association criteria point-of-view), which have a high signal-to-interference-plus-noise ratio (SINR) and/or relevant cached contents, might consistently suffer from overcapacity problem. These circumstances lead to an unfair situation wherein new incoming UEs will have only a small possibility to be served by these RRHs. In the worst case, a group of these closely related RRHs may virtually form a *black hole* area that attracts new incoming UEs according to signal strength and favorite contents, but that cannot provide services due to overcapacity. To cope with this scenario, the capacity-aware approach focuses on directing the incoming UE associations using a greedy strategy with respect to the available capacity of the RRHs. Although the capacity-aware approach achieves better resource balance and fairness among the RRHs, the spectral efficiency is unconsidered. Therefore, an efficient resource balancing scheduler is needed to alleviate the burden of the interesting RRHs as well as to support the spectral efficiency, resulting in better serviceability.

To overcome this challenge, we propose an adaptive resource balancing (ARB) scheme for serviceability maximization in F-RANs wherein the resource block (RB) utilization among RRHs is balanced using the *backpressure algorithm* with respect to a time-varying network topology. Moreover, RRH mobility is considered by the proposed scheme to better adapt to scenarios where mobile RRHs are implemented to alleviate the user serving burden of the interesting RRHs. The optimal UE selection for service movement from a high-RB-utilization RRH to its neighboring low-RB-utilization RRHs is determined by the *Hungarian method* [12] to minimize RB occupation after moving the service. Thereafter, user associations are coordinated among the neighboring RRHs to share the burden of serving UEs. Therefore, the overcapacity black hole problem and the unfairness of association possibility are both addressed, resulting in a significant improvement of the serviceability in F-RANs. The main contributions of this paper are three-fold:

- We have anatomized the F-RAN system model and formulated the characteristics that affect the serviceability. The problem statement representing our target of serviceability maximization is identified along with the constraints.
- We have proposed the ARB scheme for RRHs which uses the backpressure algorithm to distribute UE association and service to address the black hole of overcapacity problem and unfairness of user associations. The optimal UE selection for service movement is determined by the Hungarian method to achieve the minimum RB occupation after moving the service.
- Evaluations have been performed to verify the superior performance of our proposed scheme over the capacity-aware, max-rate, and cache-aware algorithms in terms of the serviceability, availability, and throughput.

The remainder of this paper is organized as follows. Section II describes a literature review of the state-of-the-art approaches. Section III presents the system model and problem statement wherein the related characteristics of F-RANs are identified and the constraints are formulated. Section IV introduces our approach based on the backpressure algorithm and Hungarian method. Performance evaluations and analysis in Section V show our proposed scheme's effectiveness over the capacity-aware, max-rate, and cache-aware approaches. Finally, Section VI concludes the paper.

II. LITERATURE REVIEW

As aforementioned in Section I, cell association in F-RANs is mostly driven by two descending priorities: RRHs which have interesting cached contents (i.e., cache-aware approaches) and RRHs which have high SINR (i.e., max-rate approaches). The cache-aware approaches drive user associations according to the relevant cached contents that user devices are interested in [13] and [14]. In conjunction with the content pre-fetching algorithms [4], [15], cache-aware approaches aim at providing low end-to-end latency as well as service stability for better user experience. There are a variety of studies that have been proposed based on these approaches recently, particularly in 5G systems [16]. Since the cache-aware approaches prioritize content delivery, they have disadvantage in spectral efficiency, thus leading to low serviceability for a massive user devices.

In max-rate approaches, spectral efficiency is preferred for user associations in order to achieve the maximum data rate. Since the spectral efficiency is well known typical metric for network evaluation, several researchers have

focused on improving this feature by using optimization techniques [17], [18], game theoretical models [19], interference managements [20], etc., as indicated by the intensive survey in [21]. Although the max-rate approaches provide better spectral utilization and throughput, they cannot address the issues of unfairness and unbalancing among cells [22] which negative impact to the network serviceability.

To balance the load among cells in term of user serving, the capacity-aware approaches are introduced in the literature. In the capacity-aware approaches, user devices consider the serving capacities of cells for association [23], [24]. In spite that the existing capacity-aware approaches can resolve the unbalance among cells, the mobility of the cells as well as the ability of user devices to simultaneously associate with multiple cells (as featured in F-RAN) have not been considered.

Comprehensive analysis and evaluation are performed by Yan *et al.* [22] to illustrate the impact of cell association with respect to edge-computing node locations, cache sizes, and user access modes. The numerical results of the ergodic rate reveal proportional trends between the number of connecting devices per RRH and the number of edge-computing nodes, edge-computing node density, and cache size. In other words, this phenomenon illustrates the problem with overcapacity in the interesting RRHs. Fortunately, D2D communications may alleviate the burden of RRHs since UEs prefer to associate in peer edge-computing devices due to various advantages (e.g., low latency, energy efficiency, and interference mitigation) thanks to the evolutionary game-based approach [11]. In spite of this, D2D communication should only be considered as an additional utilization in terms of serviceability improvement due to the small cache size in edge-computing devices. Meanwhile, needy efficient resource balancing remains its key responsibility.

III. PROBLEM STATEMENT

Considering the F-RAN system model illustrated in Fig. 1, the RRHs are dynamically deployed depending on the local traffic interests, resulting in geography-varying RRH densities. In addition, mobile RRHs could be used in order to alleviate the rush-hour traffic and the temple burst data transmissions. Without loss of generality, the arrival rate and departure rate of UEs to/from the network are assumed to follow a Poisson process with mean value λ and an exponential process with mean value μ, respectively. Meanwhile, since the locations of UEs are driven by local traffic interests, the number of UEs associated in an RRH can be modeled by using the Zipf distribution [25]. The notation used in this paper is summarized in Table 1.

Given the desired data rate v_j of the *j-th* UE for service satisfaction, the number of resource blocks (RBs) r_{ij} that the *i-th* RRH must assign to the *j-th* UE according to the data rate v_j [26] is determined by

$$r_{ij} = \left\lceil \frac{v_j}{\Delta f \log_2(1 + \text{SINR}_{ij})} \right\rceil, \tag{1}$$

TABLE 1. Notation definitions.

Symbol	Meaning
N	Number of RRHs
R	Radius of RRH coverage area
C_i	Capacity of *i-th* RRH (unit: RB)
$Q_i(t)$	Current occupied capacity of *i-th* RRH assigned to UEs at time slot t
$S_i(t)$, $S(t)$	Current serviceability of *i-th* RRH and the network at time slot t, respectively
$U_i^*(t)$	Set of UEs intending to associate in the *i-th* RRH at time slot t
$U_i(t)$	Set of UEs associated in the *i-th* RRH at time slot t
λ	Mean arrival rate of UEs
μ	Mean departure rate of UEs
p_i	Probability that a UE associates in the *i-th* RRH
v_j	Desired data rate of *j-th* UE
$\hat{v}$	Minimum data rate required by the UEs
r_{ij}	Number of resource blocks required from the *i-th* RRH to satisfy the data rate v_j of the *j-th* UE
ψ	Distance bound defining neighboring relations among RRHs
$\delta_{ik}(t)$	Feasible service movement indicator from RRH i to RRH k at time slot t
$\Psi_i(t)$	Neighboring RRH set of *i-th* RRH determined by ψ and $\delta_{ik}(t)$ at time slot t
$u_{ik}^*(t)$	Optimal UE for service movement from RRH i to RRH k at time slot t
$W_{ik}^{(j)}(t)$	Weight of moving *j-th* UE service from RRH i to RRH k at time slot t

where $r_{ij} \in \mathbb{N}$, Δf is the bandwidth that 1 RB utilizes during 1 ms (i.e., 180 KHz [27]), and SINR_{ij} is the signal-to-interference-plus-noise ratio on the data channel between the *i-th* RRH and the *j-th* UE.

Following [25], the interest of a UE to associate with an RRH depends on the popularity of the cached data contents in the RRH as long as the signal quality is above an acceptable threshold. In other words, the probability p_i that a UE intends to associate in the *i-th* RRH is given by

$$p_i = f_{\xi_i}(\sigma, \Xi) = \frac{\frac{1}{\xi_i^\sigma}}{\sum_{k=1}^{\Xi} \frac{1}{k^\sigma}}, \tag{2}$$

where Ξ is the number of content topics, ξ_i is the popularity rank of the cached data contents in *i-th* RRH, and the Zipf exponent σ ($\sigma > 0$) controls the relative popularity of the data [11]. Since the mean arrival rate λ and mean departure rate μ of the UEs to/from the network are the same for the whole network, the number of UEs arriving at an RRH heavily depends on its popularity with respect to the interesting cached contents. Let $U_i^*(t)$ be a set of UEs that intend to associate with the *i-th* RRH at time slot t. Hence, the expected number of such UEs can be obtained as

$$\mathbb{E}[|U_i^*(t)|] = \sum_{j=0}^{\infty} j e^{-\lambda} \frac{\lambda^j}{j!} p_i = \lambda p_i. \tag{3}$$

Because the mean departure rate μ is constant, the number $|U_i(t)|$ of UEs being served by the *i-th* RRH at time slot t proportionally increases with $|U_i^*(t)|$. On the other hand, the current occupied capacity $Q_i(t)$ of the *i-th* RRH assigned to

its UEs is given as

$$Q_i(t) = \sum_{j=0}^{|U_i(t)|} r_{ij}. \tag{4}$$

In other words, the current occupied capacity of the *i-th* RRH $Q_i(t) \propto \{\lambda, \xi_i\}$. According to the serviceability definition mentioned in Section I, the serviceability of the *i-th* RRH at time slot t (i.e., during $[0, t]$ duration) is derived as

$$S_i(t) \triangleq \frac{|\bigcup_{\tau=0}^{t} U_i(\tau)|}{\sum_{\tau=0}^{t} |U_i^*(\tau)|}, \quad \tau = 1, 2, \cdots, t. \tag{5}$$

Therefore, the serviceability of the network is obtained by

$$S(t) \triangleq \frac{\sum_{i=0}^{N} |\bigcup_{\tau=0}^{t} U_i(\tau)|}{\sum_{i=0}^{N} \sum_{\tau=0}^{t} |U_i^*(\tau)|}, \quad \tau = 1, 2, \cdots, t, \tag{6}$$

where N is the number of RRHs in the network.

It is observed that acceleration of the current occupied capacity directly affects the serviceability of an RRH. When the occupied capacity exceeds maximum capacity, the serviceability constraint is violated. Assuming that same conditions are applied to every UE, since the UE arrival rates and departure rates are objective, the effects of the popularity rank of the cached data contents in an RRH (refer to Equation 2) should be relaxed by emigrating user services to neighboring RRHs, which are assumed to have similar conditions.

As aforementioned, the network serviceability can be improved by emigrating user services to appropriate neighboring RRHs. To achieve this, we propose the ARB scheme wherein RB utilization among neighboring RRHs are balanced using the backpressure algorithm with respect to a time-varying network topology that is affected by RRH mobilities. A distance threshold ψ is introduced in order to define a bounded radius for determining the neighboring relations between two RRHs. Suppose that the coverage area of every RRH is disc-like with a radius of R. Given the fact that a UE can only move services between two RRHs if the UE is located in the overlapped area of these two RRHs, the maximum value of ψ is obtained by doubling R, resulting in $\psi = \alpha R$, where $0 \leq \alpha \leq 2$. According to the distance threshold ψ, each *i-th* RRH has a set of possible neighboring RRHs at time slot t considering RRH mobility.

For each relation between the *i-th* RRH and its *k-th* possible neighboring RRH, a feasible service movement indicator $\delta_{ik}(t)$ is defined to ensure that the current occupied capacity (in percent) of RRH i is greater than the current occupied capacity of RRH k following the proposed scheme rationale of moving services from high-RB-utilization RRH to low-RB-utilization RRH,

$$\delta_{ik}(t) \triangleq \begin{cases} 1 & \text{if } \dfrac{Q_i(t)}{C_i} > \dfrac{Q_k(t)}{C_k} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The combination of the distance threshold ψ and feasible service movement indicators determines the corresponding neighboring RRH set $\Psi_i(t)$ of the *i-th* RRH. The neighboring RRH sets help form time-varying RRH relation matrices. The RRH relation matrix is then transformed into a directed graph model wherein RRHs and RRH relations are considered as nodes and directed edges, respectively.

The backpressure algorithm is applied in the directed graph model of RRHs for the movement of user services from higher-occupied-capacity RRHs to lower-occupied-capacity RRHs in order to achieve approximate balance statuses among the RRHs. The optimal UE selection for service movement in each edge is determined by the Hungarian method to minimize RB occupation to improve the spectral efficiency.

IV. ADAPTIVE RESOURCE BALANCING

A. THE BACKPRESSURE ALGORITHM

The backpressure algorithm is a method for routing and scheduling commodities (e.g., data packets and radio resources) around a queuing system. Since the original backpressure algorithm was introduced by Tassiulas and Ephremides [28], a variety of studies have shown that backpressure-based systems for wireless multihop architectures achieve good performance in terms of network throughput optimality, adaptive resource allocation, traffic load balancing, and simple implementation [29]. The rationale of the backpressure algorithm is a large-queue-backlog node (i.e., meaning high pressure) should deliver its commodities to neighboring small-queue-backlog nodes (i.e., meaning low pressure) in order to push commodities travelling around the system. The backpressure algorithm is mathematically constructed and verified using *Lyapunov drift* [30]. Assume that there is a graph model wherein each node owns a separate queuing buffer, then the algorithm includes two iterative main steps in every time slot t as follows:

Step 1 (Optimal commodity determination): On each link (i,k), an optimal commodity is selected that satisfies the targeted functions (e.g., throughput maximization, load balancing, and resource allocation).

Step 2 (Delivery amount calculation): The optimal active link matrix is obtained by $\arg\max \sum_{\forall(i,k)} W_{ik}(t)\, r_{ik}(t)$, where $W_{ik}(t)$ represents the backlog differential of the optimal commodity between node i and node k, and $r_{ik}(t)$ is the data rate of link (i,k) at time slot t. Optimal commodities are delivered on the active links within the amount of $\max[\min(W_{ik}(t), r_{ik}(t)), 0]$.

B. OPTIMAL UE SELECTION FOR SERVICE MOVEMENT

Suppose that the *i-th* RRH has a neighboring RRH set $\Psi_i(t)$ derived from the bounded distance threshold ψ of αR and the feasible service movement indicator $\delta_{ik}(t)$ at time slot t. Moreover, the RRH reaches a current occupied capacity of $Q_i(t)$ since it is serving the UE set $U_i(t)$ at this time. As aforementioned in Section III, the *i-th* RRH should move UE services to its neighboring RRHs whenever its occupied capacity is greater than those of the neighboring RRHs. The optimal UE selection for service movement between two

RRHs is determined via the *Hungarian method* to achieve the minimum RB occupation after moving the service. In order to consider the RB occupation after moving the service, the *weight* of service movement is proposed. The weight $W_{ik}^{(j)}(t)$ of service movement of the *j-th* UE from the *i-th* RRH to the *k-th* neighboring RRH in $\Psi_i(t)$ is defined by the ratio of RB utilization for the *j-th* UE between two such RRHs as follows:

$$W_{ik}^{(j)}(t) \triangleq \frac{\log_2(1+\text{SINR}_{ij})}{\log_2(1+\text{SINR}_{kj})}, \tag{8}$$

where $k = 1, 2, \cdots, |\Psi_i(t)|$. The optimal UE selection problem can be described by

$$\text{minimize}: \sum_{j=1}^{|U_i(t)|}\sum_{k=1}^{|\Psi_i(t)|} x_{jk} W_{ik}^{(j)}(t) \tag{9}$$

$$\text{s.t. } x_{jk} \in \{0, 1\}, \tag{10}$$

$$\sum_{j=1}^{|U_i(t)|} x_{jk} \leq 1, \ \forall k = 1, 2, \cdots, |\Psi_i(t)|, \tag{11}$$

$$\sum_{k=1}^{|\Psi_i(t)|} x_{jk} \leq 1, \ \forall j = 1, 2, \cdots, |U_i(t)|, \tag{12}$$

$$\sum_{j=1}^{|U_i(t)|}\sum_{k=1}^{|\Psi_i(t)|} x_{jk} = \min(|U_i(t)|, |\Psi_i(t)|), \tag{13}$$

where the indicator x_{jk} is given by

$$x_{jk}(t) \triangleq \begin{cases} 1 & \text{if service of } j\text{-th UE is moved to } k\text{-th RRH} \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Constraints 11 and 12 ensure that a potential UE in $U_i(t)$ could only match with at most one selective RRH in $\Psi_i(t)$ and vice versa. Constraint 13 to ensure that maximum matchings are established between $U_i(t)$ and $\Psi_i(t)$.

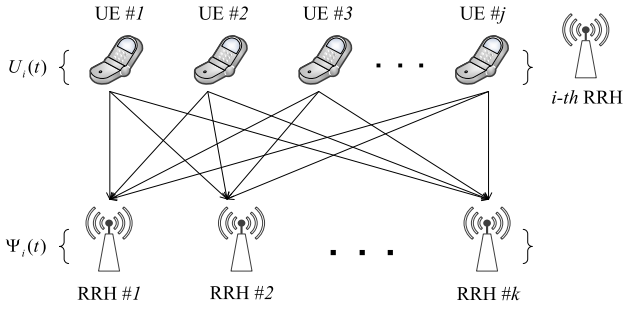


FIGURE 2. Optimal UE selection for service movement from *i-th* RRH to neighboring RRHs based on the Hungarian method.

In order to address this problem, we apply the Hungarian method on the bipartite graph from UE set $U_i(t)$ and the neighboring RRH set $\Psi_i(t)$ of the *i-th* RRH; see Fig. 2. The Hungarian method is performed as follows [12]:

Step 1: Augment the weight matrix of $W_{ik}^{(j)}(t)$ into a square matrix by adding additional entries of 0 (*Lines 1-6, Algorithm 1*).

Step 2: In the square weight matrix $[W_{ik}^{(j)}(t)]$, subtract the entries of each row by the smallest row entry. Consequently, subtract the entries of each column by the smallest column entry (*Lines 7-14, Algorithm 1*).

Step 3: Mark the minimum number of rows and columns of the weight matrix for which all the zero entries are covered. Such set of such rows and columns is referred to as an optimal 0-covered set (*Line 15, Algorithm 1*).

Step 4: If the size of the optimal 0-covered set is equal to $\max(|U_i(t)|, |\Psi_i(t)|)$, the solution is found wherein optimal matching between $U_i(t)$ and $\Psi_i(t)$ are obtained at zero entries resulting in a summation of 0, referred to as a 0-summed matching set. Otherwise, do Step 5 (*Lines 16-18, Algorithm 1*).

Step 5: Find the smallest unmarked entry. Subtract this entry from each uncovered row, and then add it to each covered column. Jump to Step 3 (*Lines 19-25, Algorithm 1*).

Algorithm 1 Optimal UE Selection

Input: $U_i(t)$, $\Psi_i(t)$ ▷ Bipartite graph
Output: $\{u_{ik}^*(t)|k = 1, 2, \cdots, |\Psi_i(t)|\}$ ▷ Optimal UEs

1: Initiate $[W] = \{\{W_{ik}^{(j)}(t)|j = 1, 2, \cdots, |U_i(t)|\}; k = 1, 2, \cdots, |\Psi_i(t)|\}$
2: $n = \max(|U_i(t)|, |\Psi_i(t)|)$
3: **if** $n == |U_i(t)|$ **then**
4: $\quad [W] = [[W]|[0]]_{nn} = [w_{jk}]$
5: **else**
6: $\quad [W]^T = [[W]^T|[0]]_{nn} = [w_{jk}]^T$
7: **for** $j = 1, j \leq n, j++$ **do**
8: $\quad x = \min\{w_{jk}|k = 1, 2, \cdots, n\}$
9: $\quad$ **for** $k = 1, k \leq n, k++$ **do**
10: $\quad\quad w_{jk} = w_{jk} - x$
11: **for** $k = 1, k \leq n, k++$ **do**
12: $\quad x = \min\{w_{jk}|j = 1, 2, \cdots, n\}$
13: $\quad$ **for** $j = 1, j \leq n, j++$ **do**
14: $\quad\quad w_{jk} = w_{jk} - x$
15: $S =$ optimal 0-covered set
16: **if** $|S| == n$ **then**
17: $\quad$ **for** $\forall w_{jk} \in$ 0-summed matching set **do**
18: $\quad\quad u_{ik}^*(t) = j$
$\quad$ **return**
19: **else**
20: $\quad z = \min\{\forall w_{jk} \notin S\}$
21: $\quad$ **for** $\forall w_{jk}$ from rows $\notin S$ **do**
22: $\quad\quad w_{jk} = w_{jk} - z$
23: $\quad$ **for** $\forall w_{jk}$ from columns $\in S$ **do**
24: $\quad\quad w_{jk} = w_{jk} + z$
25: $\quad$ **jump** to Line 15
26: **end**

Finally, a set $\{u^*_{ik}(t)|k = 1, 2, \cdots, |\Psi_i(t)|\}$ of optimal UEs for each service movement from the *i-th* RRH to its *k-th* neighboring RRH in $\Psi_i(t)$ is achieved.

As an example, suppose that the *i-th* RRH is serving a set of UE $U_i(t) = \{$UE #1, UE #2, UE #3, UE #4$\}$ and has a set of neighboring RRHs $\Psi_i(t) = \{$RRH #1, RRH #2, RRH #3$\}$ at time slot t; referred to Fig. 2. Based on Equation 8, assume that the weight of service movement of moving each UE belonging to $U_i(t)$ to an RRH belonging to $\Psi_i(t)$ is determined, resulting in the weight matrix W

$$W = \begin{bmatrix} 3 & 1 & 1 \\ 2 & 3 & 4 \\ 2 & 2 & 3 \\ 1 & 4 & 2 \end{bmatrix},$$

where the UEs and RRHs are indexed by rows and columns, respectively. Algorithm 1 is applied to the weight matrix W as follows:

$$\begin{bmatrix} 3 & 1 & 1 \\ 2 & 3 & 4 \\ 2 & 2 & 3 \\ 1 & 4 & 2 \end{bmatrix} \overset{(a)}{\rightarrow} \begin{bmatrix} 3 & 1 & 1 & \underline{0} \\ 2 & 3 & 4 & \underline{0} \\ 2 & 2 & 3 & \underline{0} \\ 1 & 4 & 2 & \underline{0} \end{bmatrix} \overset{(b)}{\rightarrow} \begin{bmatrix} 3 & \underline{1} & \underline{1} & \underline{0} \\ 2 & 3 & 4 & 0 \\ 2 & 2 & 3 & 0 \\ \underline{1} & 4 & 2 & 0 \end{bmatrix}$$
$$\overset{(c)}{\rightarrow} \begin{bmatrix} \mathbf{2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \underline{1} & 2 & 3 & \mathbf{0} \\ 1 & 1 & 2 & \mathbf{0} \\ \mathbf{0} & \mathbf{3} & \mathbf{1} & \mathbf{0} \end{bmatrix} \overset{(d)}{\rightarrow} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 3 & 1 & 0 \end{bmatrix} \overset{(e)}{\rightarrow} \begin{bmatrix} 2 & 0 & \mathbf{0} & 1 \\ 0 & 1 & 2 & \mathbf{0} \\ 0 & \mathbf{0} & 1 & 0 \\ \mathbf{0} & 3 & 1 & 1 \end{bmatrix}$$

According to the chain above, $(\overset{(a)}{\rightarrow})$ entries 0 are augmented to achieve a square matrix; $(\overset{(b)}{\rightarrow})$ subtract the entries in each row by the smallest row entry (i.e., 0s – marked by underlined entries); $(\overset{(c)}{\rightarrow})$ subtract the entries in each column by the smallest column entry (i.e., 1, 1, 1, and 0 – marked by underlined entries). Note that a combination of row #1, row #4, and column #4 covers all 0 entries (marked by bold entries). The number of marked rows and columns is equal to 3 less than the total number of matrix rows (or columns), i.e., 4. Therefore, $(\overset{(d)}{\rightarrow})$ the smallest entry among the unmarked entries (i.e., 1 – marked by the underlined entry) is subtracted from each uncovered row, and then $(\overset{(e)}{\rightarrow})$ adds this smallest entry to each covered column. Now, the combination of row #1, row #2, row #3, and column #1 covers all 0 entries and the number of such rows and columns is 4, which is equal to the total number of matrix rows (or columns). Therefore, the solution is found wherein the 0-summed matching set is $\{w_{13}, w_{24}, w_{32}, w_{41}\}$ (marked by bold entries). In other words, UE #1, UE #3, and UE #4 are the optimal UEs for moving service from the *i-th* RRH to RRH #3, RRH #2, and RRH #1 in $\Psi_i(t)$, respectively.

C. ADAPTIVE RESOURCE BALANCING

We consider an F-RAN wherein N RRHs are deployed. At time slot t, the neighboring RRH set $\Psi_i(t)$ of every *i-th* RRH is determined by the distance threshold ψ of αR and the feasible service movement indicator $\delta_{ik}(t)$. A directed graph model $G(V, E)$ is initiated, where each vertex x_i represents the *i-th* RRH. The directed edges are established based on the relations between the *i-th* RRH and its neighboring RRHs in $\Psi_i(t)$; see Algorithm 2.

Algorithm 2 Adaptive Resource Balancing

$\psi = \alpha R$
Find $\Psi_i(t)$ for each *i-th* RRH
Initiate a directed graph $G(V, E)$
Set *i-th* RRH as vertex x_i in V, $\forall i = 1, 2, \cdots, N$
for $i = 1, i \leq N, i++$ **do**
  for $k = 1, k \leq |\Psi_i(t)|, k++$ **do**
    Set a directed edge from x_i to x_k
  $S_i(t) = 1 - \frac{Q_i(t-1) + \sum r\_arr_{ij} - \sum r\_dep_{ij}}{C_i}$
for $i = 1, i \leq N, i++$ **do**
  Find $\{u^*_{ik}(t)|k = 1, 2, \cdots, |\Psi_i(t)|\}$ by Algorithm 1
  for $\forall u^*_{ik}(t) \in \{u^*_{ik}(t)\}$ **do**
    Find $r\_out_{(i\rightarrow k)j}$ and $r\_in_{(i\rightarrow k)j}$
    $Q_i(t) = Q_i(t) - r\_out_{(i\rightarrow k)j}$
    $Q_k(t) = Q_k(t) + r\_in_{(i\rightarrow k)j}$
end

According to the backpressure scheme, optimal UE selection for service movement should be determined for each direct edge (i, k) from the *i-th* RRH to its *k-th* neighboring RRH in $\Psi_i(t)$, which we can obtain via the Hungarian method described in Section IV-B. Assume that the current amount of RBs that the *i-th* RRH assigns to the *j-th* UE is $r_{ij}(t)$. For RB-utilization balancing purposes between the two vertices of edge (i, k), the maximum amount of RBs serving the optimal UE $u^*_{ik}(t)$ that the *i-th* RRH can release $(r\_out^*_{(i\rightarrow k)j})$ so they can be re-assigned by the *k-th* RRH afterward $(r\_in^*_{(i\rightarrow k)j})$ is calculated as follows:

$$r\_out^*_{(i\rightarrow k)j} = \min\left(\beta\left(S_k(t) - S_i(t)\right)C_i, r_{ij}\right), \quad (15)$$
$$r\_in^*_{(i\rightarrow k)j} = \beta\left(S_k(t) - S_i(t)\right)C_k, \quad (16)$$

where $0 < \beta \leq 1$ is a balance factor. Since the number of needy assigned RBs satisfying a given UE data rate depends on the SINR between the RRH and UE (refer to Equation 1), the amount of RBs serving the optimal UE $u^*_{ik}(t)$ that the *i-th* RRH should release $(r\_out_{(i\rightarrow k)j})$ according to the corresponding amount of RBs supported by *k-th* RRH $(r\_in_{(i\rightarrow k)j})$ cannot exceed the maximum values. Moreover, the ratio of $r\_out_{(i\rightarrow k)j}$ to $r\_in_{(i\rightarrow k)j}$ must equal the weight of service movement $W^{(j)}_{ik}(t)$ to support the same UE data rate. Therefore, we obtain

$$r\_out_{(i\rightarrow k)j} = \begin{cases} r\_out^*_{ij} & \text{if } \dfrac{r\_in^*_{kj}}{r\_out^*_{ij}} > W^{(j)}_{ik}(t) \\ \left\lceil \dfrac{r\_in^*_{kj}}{W^{(j)}_{ik}(t)} \right\rceil & \text{otherwise,} \end{cases} \quad (17)$$

$$r\_in_{(i\rightarrow k)j} = \begin{cases} \left\lceil r\_out^*_{ij} W^{(j)}_{ik}(t) \right\rceil & \text{if } \dfrac{r\_in^*_{kj}}{r\_out^*_{ij}} > W^{(j)}_{ik}(t) \\ r\_in^*_{kj} & \text{otherwise.} \end{cases} \quad (18)$$

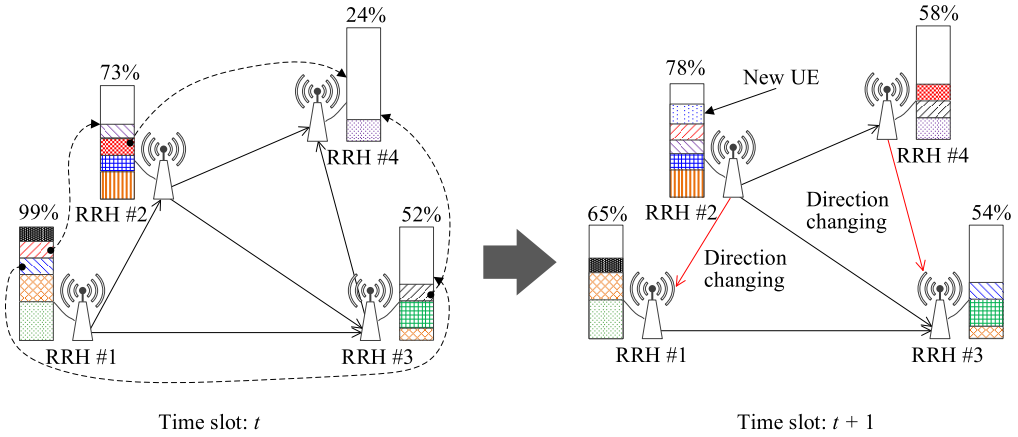


FIGURE 3. An example of the ARB scheme for resource balancing among neighboring RRHs by emigrating user services.

The occupied capacity $Q_i(t+1)$ of the *i-th* RRH at time slot $t+1$ is updated as follows:

$$Q_i(t+1) = Q_i(t) + \sum_{k=1}^{|\Psi_i(t)|} r\_in_{(k\to i)j} - \sum_{k=1}^{|\Psi_i(t)|} r\_out_{(i\to k)j} + \sum r\_arr_{ij} - \sum r\_dep_{ij}, \quad (19)$$

where $\sum r\_arr_{ij}$ and $\sum r\_dep_{ij}$ are the numbers of RBs assigned to the new incoming UEs and the UEs released during time schedule $[t, t+1)$ at the *i-th* RRH, respectively. Accordingly, the serviceability $S_i(t+1)$ of the *i-th* RRH and $S(t+1)$ of the network updated in Equation 5 and 6, respectively.

Fig. 3 illustrates an example of the ARB scheme for resource balancing among neighboring RRHs. At time slot t, RRH #1, RRH #2, RRH #3, and RRH #4 have occupied capacities of 99%, 73%, 52%, and 24%, respectively. Due to the bounded distance threshold and feasible service movement indicator, a directed graph model ($G(V, E)$) of the network can be derived in the form of a relation matrix as follows:

$$G(V,E) = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where rows and columns represent vertices, and an entry value of 1 indicates there is a relation between the two corresponding RRHs, and vice versa. Following our proposed scheme in Algorithm 1, optimal UEs are determined and the service movement is performed on the edges of (RRH #1 → RRH #2), (RRH #1 → RRH #3), (RRH #2 → RRH #4), and (RRH #3 → RRH #4). Suppose that during time schedule $[t, t+1)$, a new UE arrives at RRH #2. Finally, the occupied capacities of RRH #1, RRH #2, RRH #3, and RRH #4 are updated by 65%, 78%, 54%, and 58%, respectively, at time slot $t+1$. As a result, the directions on the edges of (RRH #1 → RRH #2), (RRH #3 → RRH #4) are reversed (indicated with red lines). The relation matrix is given by

$$G(V,E) = \begin{bmatrix} 0 & \mathbf{0} & 1 & 0 \\ \mathbf{1} & 0 & 1 & 1 \\ 0 & 0 & 0 & \mathbf{0} \\ 0 & 0 & \mathbf{1} & 0 \end{bmatrix}.$$

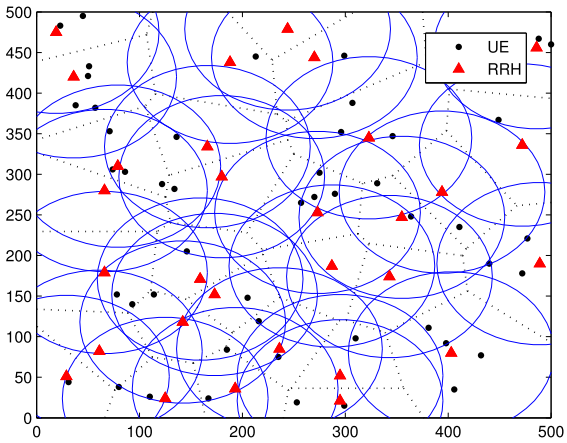


FIGURE 4. Distribution of the UEs and RRHs in the simulation scenario where the coverage areas are formed according to the distance threshold (circles centered at triangles) and Voronoi tessellation (dotted lines).

V. PERFORMANCE EVALUATION

In this section, we evaluate the performance of our proposed ARB scheme through a series of simulation studies. A network scenario has been developed including 30 RRHs (wherein 5 RRHs are mobile) randomly deployed over an area of 500×500 square meters. Fig. 4 depicts an example of a simulation scenario where the coverage areas are formed according to the distance threshold (circles centered at a triangle) and Voronoi tessellation (dotted lines). Detailed simulation parameters are summarized in Table 2. Since our proposed scheme executes independently of the UE association process, we performed simulations for six different UE association strategies, including (a) capacity-aware,

(b) capacity-aware + ARB, (c) max-rate, (d) max-rate + ARB, (e) cache-aware, and (f) cache-aware + ARB. In this way, we aim to identify the effects of the ARB scheme on individual UE association approaches in terms of serviceability, availability, and throughput. The availability is defined by the percentage of UEs that are served within the minimum requirements (e.g., throughput, delay, and packet loss) by the network per the cumulative arrival of UEs during a specific time interval under pre-determined UE arrival and departure rates [31]. In the scope of this paper, we consider the minimum requirement as a minimum throughput ($\hat{v}$) that is common among all UEs.

TABLE 2. Simulation parameters.

Parameter	Value
Network dimension	500m×500m
Total number of RRHs	30
Number of mobile RRHs	5
RRH coverage radius (R)	100 m
RRH bandwidth	{10, 15, 20} MHz
Mean arrival rate (λ)	10 UE/s
Mean departure rate (μ)	5 UE/s
Desired data rate (v_j)	1 – 2 Mbps
Minimum data rate ($\hat{v}$)	512 Kbps
Time slot duration	1 s

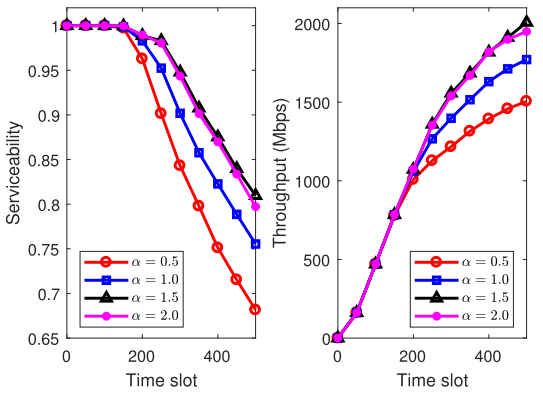


FIGURE 5. Performances of the ARB scheme depending on the bounded distance threshold ($\psi = \alpha R$).

As aforementioned in Section III, the balance factors α and β are utilized to control the strength of the bounded distance threshold defining the neighboring RRH relations and the amount of service movements between two neighboring RRHs, respectively. Fig. 5 illustrates the performances of the ARB scheme depending on the factor α. It is observed that the ARB scheme provides better network serviceability and throughput when $\alpha > 1$. Although a small difference exists, performances of the ARB scheme are approximately within the α values of 1.5 and 2. Since a higher α results in more complicated neighboring relations (i.e., larger number of neighboring RRHs), which in turn increases the computational cost of the ARB scheme, 1.5 is preferred for the value of α configuration. It is worth noting that the optimal value of α varies in a range of (1, 2) based on the network conditions. Similarly, better performance is achieved with higher values of β. Since higher β value leads to larger amounts of service movement, 0.75 is recommended for β configuration instead of 1 even though these two values contribute approximately the same effect; see Fig. 6.

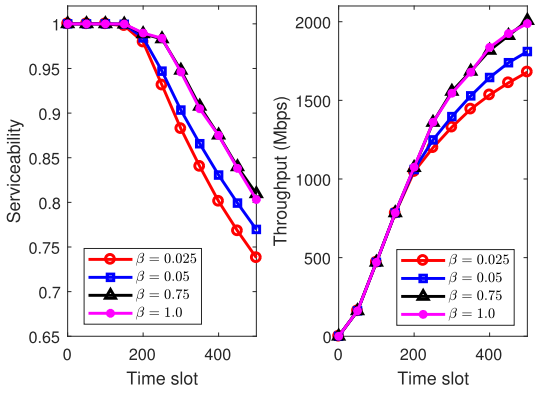


FIGURE 6. Performances of the ARB scheme depending on the balance factor β.

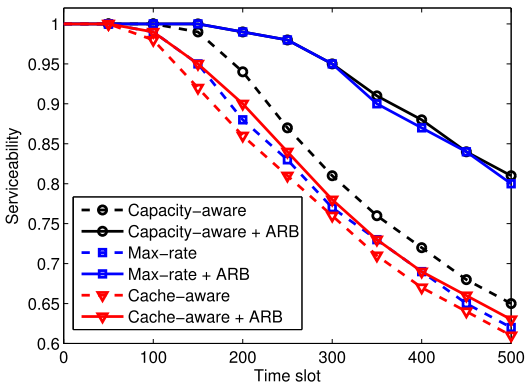


FIGURE 7. Network serviceability satisfying various desired UE data rates (v_j) of 1 – 2 Mbps.

Fig. 7 shows the network serviceability satisfying desired UE data rate ($\hat{v}$) randomized in range of 1 – 2 Mbps. During the early time slots (less than 50), almost all RRHs have enough capacity to serve the incoming UEs, representing 100% serviceability. However, since the number of simultaneous UE associations increases significantly with time ($\lambda > \mu$), some interesting RRHs reach overcapacity, resulting in a decrease of the network serviceability. The capacity-aware, max-rate, and cache-aware schemes suffer from decreasing serviceability after time slot 50. Among these approaches, the capacity-aware scheme obtains the best performance due to the consideration of available RRH capacities. On top of these UE associations, we also apply the ARB scheme in order to balance RB utilization among the RRHs. The simulation results show that the ARB scheme improves network serviceability by up to 25.18% and 29.47% compared to the standalone capacity-aware and max-rate approaches, respectively. For the cache-aware approach, the

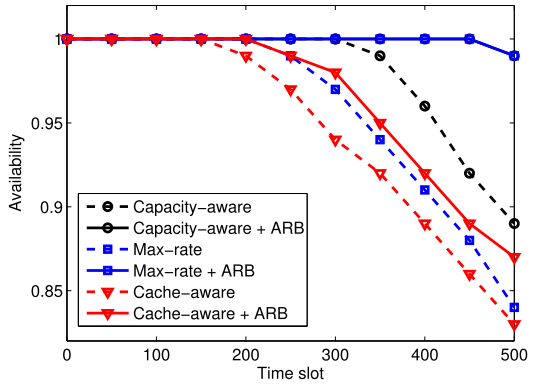


FIGURE 8. Network availability satisfying the minimum UE data rate ($\hat{v}$) 512 Kbps.

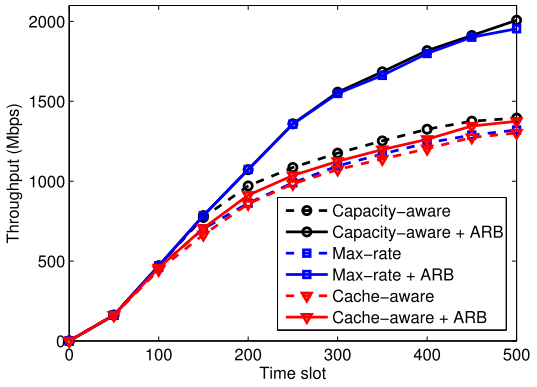


FIGURE 9. Total network throughput achieved by each of the six different UE association strategies.

ARB scheme only has a small advantage (4.07% increase) due to the consideration of cached content.

Similar to the network serviceability, the network availability is improved when applying the ARB scheme (Fig. 8) by 11.83%, 17.78%, and 4.11% compared to the standalone capacity-aware, max-rate, and cache-aware approaches, respectively. The network availabilities provided by capacity-aware + ARB and max-rate + ARB are close in value since the ARB scheme moves UE services with respect to optimal RB utilization (see Section IV-B, optimal UE selection criterion).

Regarding the total network throughput, Fig. 9 represents a comparison of the evaluated schemes. During the first 50 time slots, there is little difference in network throughput since almost all RRHs possess enough capacity to satisfy the desired UE data rates. Afterward, some RRHs reach overcapacity, resulting in UE dropped behavior. In the capacity-aware and cache-aware approaches, UE association prefers the RRHs that have a high available capacity and a large amount of cached contents, respectively. These approaches cannot achieve the best throughput performance due to the unfocused RB utilization. On the other hand, although the max-rate approach considers the highest SINR as the main criterion for UE association, its greedy behavior might be

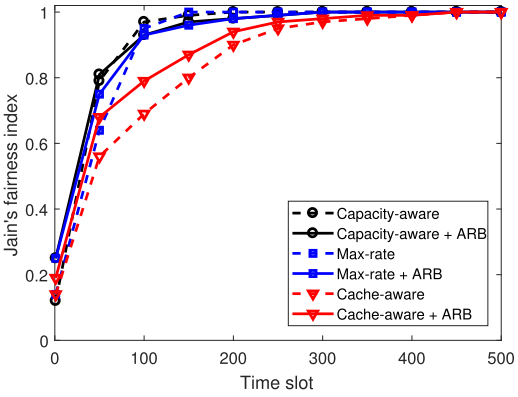


FIGURE 10. Jain's fairness index among RRHs in term of current occupied capacity (%). Plot starts from t = 1 since all nodes occupy no capacity at t = 0.

non-optimal for later incoming UEs, which must associate with lower-SINR RRHs since the best one will soon reach overcapacity. Fortunately, the ARB scheme can address the shortcomings of these UE association approaches by moving UE services between neighboring RRHs with respect to RB utilization. The simulation results reveal that applying the ARB scheme improves the total network throughput by up to 43.93%, 48.05%, and 5.63% compared to the standalone capacity-aware, max-rate, and cache-aware approaches, respectively.

Table 3 summarizes the numerical results of the simulation using popular communication indexes including: number of cumulative served UEs, number of cumulative dropped UEs, unserviceable start point (time slot), maximum throughput (Mbps), and average RB utilization (bit/RB). It is worth noting that the average RB utilization is significantly improved by up to 41.49%, 60.04%, and 1.78% when applying the ARB scheme over the capacity-aware, max-rate, and cache-aware approaches, respectively.

In order to evaluate the resource balancing effects when using the ARB scheme, we compared the Jain's fairness index [32] for each scheme. The Jain's fairness index determines the fairness of the occupied capacity as a ratio by

$$J\left(\frac{Q_i(t)}{C_i}, \forall i = 1, 2, \cdots, N\right) \triangleq \frac{\left(\sum_{i=1}^{N} \frac{Q_i(t)}{C_i}\right)^2}{N \sum_{i=1}^{N} \left(\frac{Q_i(t)}{C_i}\right)^2}. \quad (20)$$

It is observed that $0 < J(\cdot) \leq 1$ and a higher index represents better fairness among the competitors. In Fig. 10, the improved effects of the ARB scheme over existing UE association approaches is well illustrated during the first 50 time slots. Afterward, even though the Jain indexes of the standalone capacity-aware and max-rate approaches significantly increase, this actually has a negative effect since almost all RRHs reach overcapacity. On the contrary, applying the ARB scheme provides a high fairness index with an effective performance. The cache-aware algorithm obtains the worst fairness index since it is driven by interesting contents instead of communication-related parameters.

TABLE 3. Statistically numerical results.

Index	Capacity-aware	Capacity-aware + ARB	Max-rate	Max-rate + ARB	Cache-aware	Cache-aware + ARB
Cumulative served UEs	3,151	3,982	3,027	3,928	2,942	3,070
Cumulative dropped UEs	1,849	1,018	1,973	1,072	2,058	1,930
Unserviceable start point (time slot)	113	133	67	133	53	63
Maximum throughput (Mbps)	1,403.18	2,054.12	1,332.6	1,980.47	1,310.93	1,380.57
Average RB utilization (bit/RB)	481.88	681.83	424.23	678.95	495.86	504.67

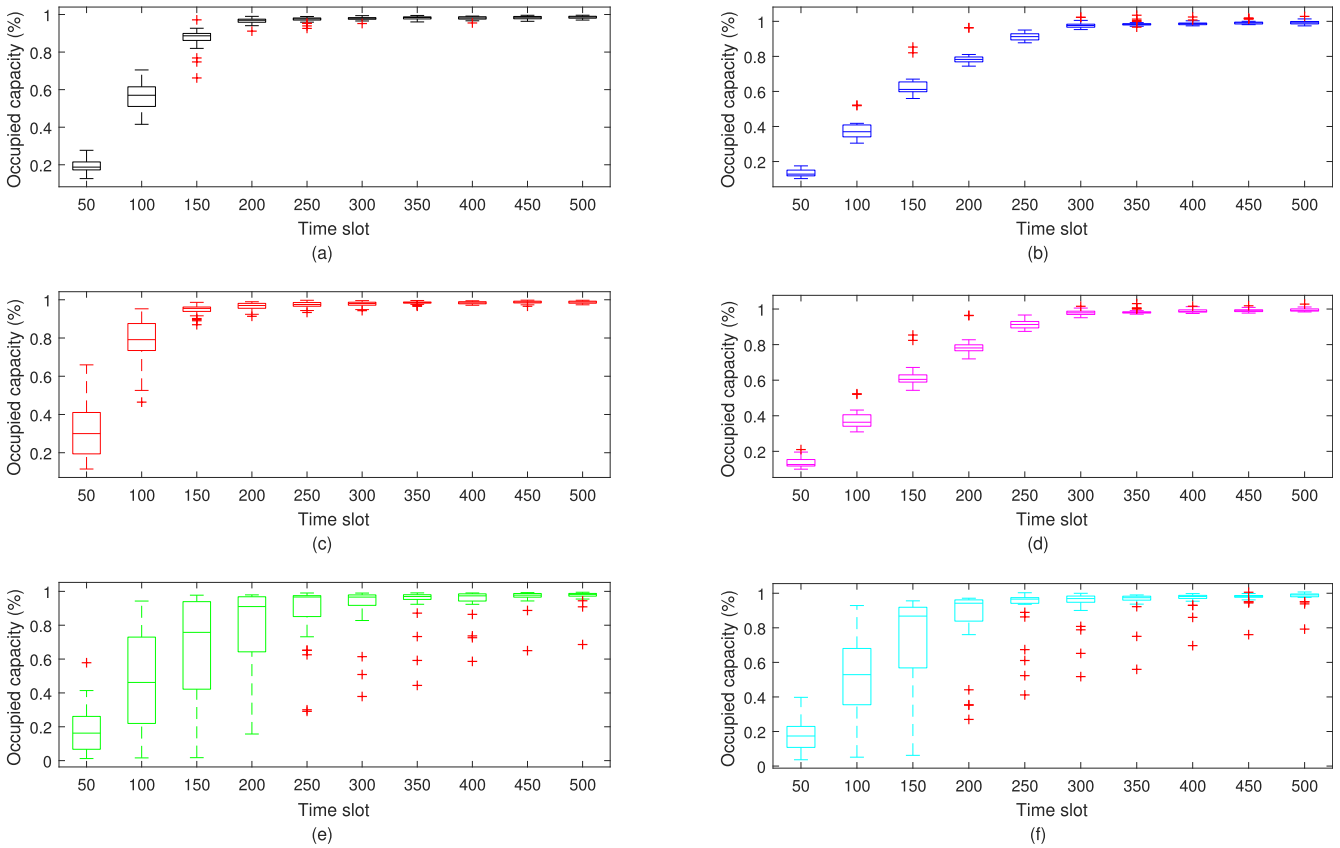


FIGURE 11. Distribution of current occupied capacity (%) in RRHs for capacity-aware, max-rate, and cache-aware UE association approaches with and without our proposed ARB scheme.

Variation in the occupied capacity of RRHs is visualized using a box-and-whisker diagram [33], shown in Fig. 11. According to sub-figures 11(a), 11(c), and 11(e), it is recognized that the occupied capacity of RRHs resulting from the capacity-aware approach are much more balanced than the others (represented by the small height of the box and the close whiskers), due to the available capacity consideration. On the contrary, the cache-aware approach leads to random dispersion among the occupied capacity of the RRHs (represented by the large height of the box and the far whiskers); this is because it focuses on cached contents. These three approaches are supplemented by the ARB scheme leading to the results shown in sub-figures 11(b), 11(d), and 11(e), respectively. The advantages that the ARB scheme provides can be described as two-fold:

- Decrease the diversity of the occupied capacity among the RRHs. Comparing each pair of sub-figures 11(a)-(b), 11(c)-(d), and 11(e)-(f), the shortened height of the box and whiskers represents the positive effects of the ARB scheme over the corresponding UE associations in terms of capacity utilization balancing.
- Reduce the probability of, and delay the time until reaching overcapacity. It is observed that almost all RRHs reach overcapacity around time slots 150 to 200 when using the existing UE association approaches. When applying the ARB scheme, this time till overcapacity is postponed to around time slots 250 to 300.

Although the diversities of the occupied capacities among RRHs between the capacity-aware and max-rate approaches are different, applying the ARB scheme results in approximately the same performance (sub-figures 11(b) and 11(d)). Moreover, the outliers help indicate the RRHs that have small neighboring relations. If the outliers exist above the whisker,

i.e., the occupied capacities are much higher than the average threshold, then we should locate mobile RRHs close to these RRHs to alleviate the burden. In contrast, if the outliers are below the whisker, i.e., the occupied capacities are much lower than the average threshold, then we should turn off some of the transceivers to reduce the power consumption. In the worst case scenario where the number of outliers is large, re-planning the radio resources among RRHs is needed to achieve a better spectrum efficiency.

VI. CONCLUDING REMARKS

In this paper, adaptive resource balancing (ARB) scheme has been proposed for serviceability maximization in Fog radio access networks (F-RANs). The backpressure algorithm is applied to the network model for service migration from higher-capacity-occupied RRHs to neighboring lower-capacity-occupied RRHs with respect to the spectral efficiency. The optimal UE selection for service migration between every two neighboring RRHs is determined by the Hungarian method to achieve the minimum resource blocks utilization. Simulation results show that the proposed ARB scheme provides significant improvement over the capacity-aware, max-rate, and cache-aware approaches in terms of serviceability, availability, and throughput. For future studies, the optimal balance factors will be used to determine the neighboring relations, and the amount of service migration should be identified according to the varying network environment conditions. Moreover, a relaxing schedule will be studied to reduce the complexity of the ARB in time and space domains while maintaining sufficient performance.

ACKNOWLEDGMENT

This work was supported by ETRI R&D Program funded by the Government of Korea (17ZH1510, Development of Technologies for Proximity, Real-time, and Smart Service Recommendation Platform).

REFERENCES

[1] S.-C. Hung, H. Hsu, S.-Y. Lien, and K.-C. Chen, "Architecture harmonization between cloud radio access networks and fog networks," *IEEE Access*, vol. 3, pp. 3019–3034, 2015.

[2] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46–53, Jul./Aug. 2016.

[3] Y.-Y. Shih, W.-H. Chung, A.-C. Pang, T.-C. Chiu, and H.-Y. Wei, "Enabling low-latency applications in fog-radio access networks," *IEEE Netw.*, vol. 31, no. 1, pp. 52–58, Jan./Feb. 2017.

[4] S.-H. Park, O. Simeone, and S. Shamai (Shitz), "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, Nov. 2016.

[5] C. Wang, Y. Li, D. Jin, and S. Chen, "On the serviceability of mobile vehicular cloudlets in a large-scale urban environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 10, pp. 2960–2970, Oct. 2016.

[6] P. Marsch *et al.*, "5G radio access network architecture: Design guidelines and key considerations," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 24–32, Nov. 2016.

[7] A. Gotsis, S. Stefanatos, and A. Alexiou, "UltraDense networks: The new wireless frontier for enabling 5g access," *IEEE Veh. Technol. Mag.*, vol. 11, no. 2, pp. 71–78, Jun. 2016.

[8] M. Peng and K. Zhang, "Recent advances in fog radio access networks: Performance analysis and radio resource allocation," *IEEE Access*, vol. 4, pp. 5003–5009, 2016.

[9] *Draft New Report ITU-R M.[IMT-2020.TECH PERF REQ]—Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)*, document 5/40-E, ITU-R, 2017.

[10] M. Peng, S. Yan, and H. V. Poor, "Ergodic capacity analysis of remote radio head associations in cloud radio access networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 4, pp. 365–368, Aug. 2014.

[11] S. Yan, M. Peng, M. A. Abana, and W. Wang, "An evolutionary game for user access mode selection in fog radio access networks," *IEEE Access*, vol. 5, pp. 2200–2210, 2017.

[12] D. Jungnickel, "Weighted matchings," in *Graphs, Networks and Algorithms* (Algorithms and Computation in Mathematics), 4th ed. Berlin, Germany: Springer, 2013, ch. 14.

[13] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.

[14] M. Erol-Kantarci, "Content caching in small cells with optimized uplink and caching power," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2015, pp. 2173–2178.

[15] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

[16] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[17] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 94–101, May 2014.

[18] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[19] D. Liu, Y. Chen, K. K. Chai, T. Zhang, and M. Elkashlan, "Opportunistic user association for multi-service hetnets using nash bargaining solution," *IEEE Commun. Lett.*, vol. 18, no. 3, pp. 463–466, Mar. 2014.

[20] J. Ghimire and C. Rosenberg, "Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1340–1351, Mar. 2013.

[21] D. Liu *et al.*, "User association in 5g networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2nd Quart., 2016.

[22] S. Yan, M. Peng, and W. Wang, "User access mode selection in fog computing based radio access networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[23] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[24] T. Zhou, Y. Huang, L. Fan, and L. Yang, "Load-aware user association with quality of service support in heterogeneous cellular networks," *IET Commun.*, vol. 9, no. 4, pp. 494–500, 2015.

[25] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. 18th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, vol. 1. Mar. 1999, pp. 126–134.

[26] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures*, Standard 36.213, 3GPP, 2016.

[27] *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation*, Standard 36.211, 3GPP, 2011.

[28] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, Dec. 1992.

[29] Z. Jiao, B. Zhang, C. Li, and H. T. Mouftah, "Backpressure-based routing and scheduling protocols for wireless multihop networks: A survey," *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 102–110, Feb. 2016.

[30] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures Commun. Netw.*, vol. 3, no. 1, pp. 1–211, 2010.

[31] J. W. Rupe, "Network nodal independence, hierarchical path search, and model reuse for network availability computation," *IEEE Trans. Rel.*, vol. 65, no. 4, pp. 1842–1851, Dec. 2016.

[32] A. B. Sediq, R. H. Gohary, R. Schoenen, and H. Yanikomeroglu, "Optimal tradeoff between sum-rate efficiency and jain's fairness index in resource allocation," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3496–3509, Jul. 2013.

[33] J. Kehrer and H. Hauser, "Visualization and visual analysis of multifaceted scientific data: A survey," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 3, pp. 495–513, Mar. 2013.

NHU-NGOC DAO received the B.S. degree in electronics and telecommunications from the Posts and Telecommunications Institute of Technology, Vietnam, in 2009, and the M.S. degree in computer science from Chung-Ang University, South Korea, in 2016, where he is currently pursuing the Ph.D. degree in computer science. His research interests include wireless communications, interference mitigation, MAC routing protocols, and networking architecture.

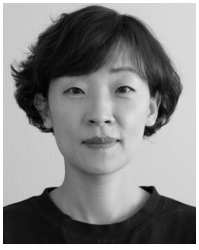
JUNWOOK LEE received the B.S. degree in computer science from the Korea Military Academy, South Korea, in 2011. He is currently pursuing the M.S. degree in computer science with Chung-Ang University. His research interests include fog computing, wireless communication, and sensor network.

DUC-NGHIA VU received the B.S. degree in electronics and telecommunications from the Hanoi University of Science and Technology, Vietnam, in 2015. He is currently pursuing the M.S. degree in computer science with Chung-Ang University. His research interests include wireless network and fog computing.

JEONGYEUP PAEK (M'15) received the B.S. degree in electrical engineering from Seoul National University in 2003, and the M.S. degree in electrical engineering and the Ph.D. degree in computer science from the University of Southern California, in 2005 and 2010, respectively. He was with Cisco Systems Inc., from 2011 to 2014, where he was a Technical Leader with the Internet of Things Group, Smart Grid Business Unit. He is an Assistant Professor with School of Computer Science and Engineering, Chung-Ang University, South Korea. His research interests are in wireless networked systems, including reliable communication and data delivery in low-power and lossy networks, challenges and new real world applications in embedded wireless sensor networks, and design of interesting mobile systems and services.

JOONGHEON KIM received the B.S. and M.S. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 2004 and 2006, respectively, and the Ph.D. degree in computer science from the University of Southern California (USC), Los Angeles, CA, USA, in 2014. Before joining USC, he was a Research Engineer with LG Electronics, Seoul, from 2006 to 2009. During his Ph.D. research at USC, he was an Intern with InterDigital, San Diego, CA, USA, in 2012. He was also a Systems Engineer with Intel Corporation, Santa Clara, CA, USA, from 2013 to 2016. He has been an Assistant Professor of Computer Science and Engineering with Chung-Ang University, Seoul, since 2016. His current research interests are in the theory, design, and implementation of distributed computing platforms, including radio platforms (especially, millimeter-wave radio and medium access technologies), advanced video streaming platforms, and multi-core embedded platforms. He is a member of the ACM and the IEEE Computer Society. He was awarded the USC Annenberg Graduate Fellowship with his Ph.D. admission from USC, in 2009.

SUNGRAE CHO received the B.S. and M.S. degrees in electronics engineering from Korea University, Seoul, in 1992 and 1994, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, in 2002. He is currently a Full Professor with the School of Computer Science and Engineering, Chung-Ang University. Prior to joining Chung-Ang University, he was an Assistant Professor with the Department of Computer Sciences, Georgia Southern University, Statesboro, from 2003 to 2006, and a Senior Member of Technical Staff with the Samsung Advanced Institute of Technology, Kiheung, South Korea, in 2003. From 1994 to 1996, he was a member of Research Staff with the Electronics and Telecommunications Research Institute, Daejon, South Korea. From 2012 to 2013, he held a visiting professorship with the National Institute of Standards and Technology, Gaithersburg, MD, USA. His research interests include wireless networking, ubiquitous computing, performance evaluation, and queuing theory. He is an Editor of the *Elsevier Ad Hoc Networks Journal* since 2012 and has served numerous international conferences as an Organizing Committee Member, such as the IEEE SECON, ICOIN, ICTC, ICUFN, TridentCom, and the IEEE MASS.



KI-SOOK CHUNG received the B.S. degree in computer engineering from the Pohang University of Science and Technology in 1995, and the M.S. degree in computer engineering with the Department of Computer Science, Korea Advanced Institute of Science and Technology, in 1997.

She joined ETRI, South Korea, as a member of Research Staff in 1997, where she is a Researcher. Her research interests include mobile edge computing, software architecture, and service platform.

CHANGSUP KEUM received the B.S. degree in computer science from the University of Seoul, South Korea, in 1992, and the M.E. degree in software engineering from Carnegie Mellon University in 2005, and the Ph.D. degree from the Korea Advanced Institute of Science and Technology in 2013. He joined ETRI, South Korea, as a member of Research Staff in 1994. From 2000 to 2003, he was engaged in research and development of soft switch systems. Since 2004, he has been engaged in research and development of various mobile service platforms. He is a Research Director with ETRI. His interests include 5G mobile service platform, mobile cloud architecture, and software engineering. He is a member of the Korea Information Processing Society and the Korean Institute of Information Scientists and Engineers.