# Resource Allocation in Heterogeneous Cloud Radio Access Networks: A Workload Balancing Perspective

Chen Ran * and Shaowei Wang*†
* School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China
† State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
E-mail: rchnju@gmail.com, wangsw@nju.edu.cn

***Abstract*—Heterogeneous cloud radio access networks (H-CRANs) have been proposed as a promising architecture for providing high energy efficiency, high spectral efficiency and high data rate at low cost. In this paper, we develop a novel resource allocation scheme among macro base stations (BSs) and remote radio heads (RRHs) for the H-CRANs. The key idea of our proposal is that we design the coverage regions of macro BSs in a seamless and balanced way. Secondly, we calculate the areas that can not rely on macro BSs for reliable high speed data transmission and allocate resources in these areas among RRHs in the balanced way to alleviate the high transmission pressure on fronthauls between RRHs and the centralized baseband unit (BBU) pool. The proposed resource allocation scheme is triggered when severe disbalance is detected. Numerical results show that our proposal can perform quite well for the data traffic distribution in a real city environment. It provides QoS guaranteed performance with lower capital expenditure (CAPEX) and operating expenditure (OPEX).**

***Index Terms*—H-CRANs, resource allocation, workload balancing**

## I. Introduction

Mobile data traffic has been increasing explosively throughout the world over the past 20 years, as more and more mobile devices together with high-speed data applications such as social networking, Internet of Things (IoTs), e-Banking, and high-definition wireless video streaming, have sprung up to make our life more convenient. Investigations have discovered over 100 percent annual growth in mobile data traffic starting from 2008, and predicted that the Internet traffic would increase by more than 1000 times by the year 2020 [1]. In order to handle the crazily increasing traffic demands, mobile network operators have to increase network capacity accordingly. As the spectral efficiency for the Long Term Evolution (LTE)-Advanced standard is approaching Shannon limit, the possible method for further improving the system spectral efficiency is increasing the density of access nodes. However, solid cell splitting gains can only be achieved in sparse deployment of access nodes, as the intercell interference will get so strong as cell splitting gains are not available. Besides, costs concerning site acquisition in urban areas can be a heavy burden. Furthermore, current cellular network architecture was primitively designed for coverage and mobility considerations, instead of achieving high energy efficiency (EE) and spectral efficiency (SE). To deal with such challenging difficulties, revolutions concerning novel wireless network architectures supported by advanced signal processing and networking technologies are urgently needed [2].

Challenges associated with the dense deployment of traditional macro BSs can be partly avoided by the utilization of low power nodes (LPNs). A network that consists of a combination of macro BSs and LPNs is referred to as a heterogeneous network (HetNet) [3]. LPNs are usually deployed in traffic hot zones underlaying macro BSs. However, HetNets still face the problem of severe interference with densely deployed LPNs. As a result, intercell interference coordination (ICIC) and coordinated multiple point (CoMP) transmission and reception are rising as promising solutions to enhance the performance of HetNets [4]. However, due to the drawbacks that these technologies rely heavily on the backhaul links and the computation resources of cellular networks, the application of the ICIC and the CoMP is still limited.

To enhance the SE and EE performance and decrease energy consumption, novel architectures for improving both SE and EE through suppressing the inter-tier interference and enhancing the cooperative processing capabilities are in urgent need. Cloud computing has been deemed as a promising technology for providing high data rates at the cost of lower energy across software defined wireless communication networks. As a consequence, heterogeneous cloud radio access networks (H-CRANs) are put forwarded as a cost-effective architecture to restrain interference and enhance cooperative processing gains in HetNets accompanied by cloud computing technology [2]. H-CRANs improve the capabilities of macro BSs with massive multiple-input multiple-output (MIMO) technology and simplify LPNs by building connections between LPNs and a signal processing cloud (baseband unit pool, BBU pool) through high-speed optical fibers. Inter-cell Interference coordination (ICIC) can also be implemented in H-CRANs scenario without much effort to improve the performance of H-CRANs. The baseband data processing and radio resources control in traditional LPNs are transferred to the BBU pool to take advantage of cloud computing technologies. Fig.1 gives an overview of the H-CRANs.

Despite the potential advantages of H-CRANs, they still require long-term development before H-CRANs coming into practice. Macro BSs play a significant role in H-CRANs as

978-1-4799-5952-5/15/$31.00 ©2015 IEEE

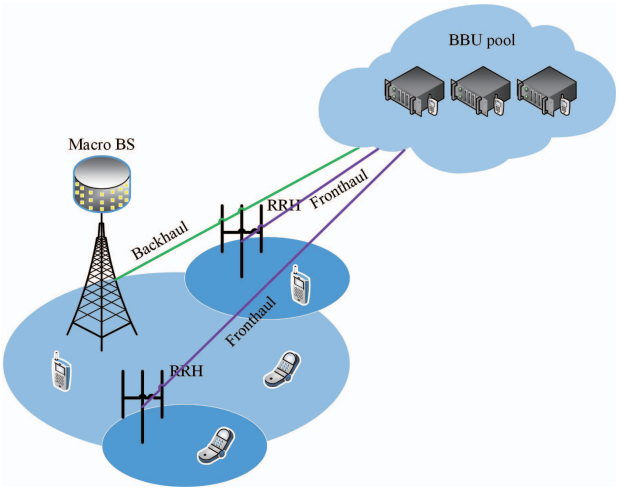


Fig. 1. Heterogeneous cloud radio access networks.

they connect with the centralized BBU pool to cooperate on addressing the cross-tier interference between RRHs and the macro cells with centralized cooperative processing techniques based on cloud computing. They are also responsible for supporting seamless coverage and delivering system control signals. RRHs compress and forward the received signals from user equipments (UEs) to the centralized BBU pool through wired or wireless fronthaul links and provide high capacity in hot zones. Under the consideration of improving EE performance of H-CRANs, the activated RRHs are adaptive to the traffic demand of the network. When the traffic demand is low, some potential RRHs fall into sleep mode under the control of the BBU pool together with macro BSs. However, when the traffic demand becomes huge in some area, both the macro BSs equipped with massive MIMO and dense RRHs work together to meet the traffic demands. But how these designs apply to the network in practice has not been investigated fully as far as the authors have known.

In this paper, we are motivated to make an effort to design a cost-saving scheme for allocating resources among macro BSs and RRHs. In particular, our scheme is designed based on workload balancing, which not only decreases the number of macro BSs required, but also alleviates the high pressure expected in the fronthauls between RRHs and the BBU pool [5]. Moreover, it is the method to cope with the situation that the traffic demands vary greatly in the mobile communication environment, which has been investigated in [6]. Meanwhile, details and experimental results of the scheme are presented. The challenging issues together with future works concerning techniques are discussed as well.

The remainder of this paper is outlined as follows. In Section II, we discuss about the relationship between our work and prior ones. In Section III, we give a brief introduction of our resource allocation strategy. In Section IV, we develop efficient algorithms to implement the resource allocation scheme. Numerical results are reported with discussions in Section V. Conclusions are drawn in Section VI.

## II. Related Work

The key idea of our resource allocation scheme is to balance the workload among macro BSs and RRHs in traffic demand aspects. Workload balancing has been researched into widely in the literature. It is a way to balance the workload among various servers [7] and machines in order to optimize factors like resource utilization, fairness, waiting/processing delays, or throughput [8]. The problem of balancing load among access points can be seen as the equitable location problem in a broad sense. By assigning the density of workload to each spot, the weighted equitable location problem is exactly the workload balancing one. Equitable location problem on a plane has been studied in the Operational Research field, which is generally designed to locate $M$ facilities on a unit square so as to minimize the maximum demand faced by any facility subject to the closest assignments and coverage constraints. The proposed strategies are usually based on local or global adjustments (depending on which strategy designers practically adopt) and Voronoi diagram [7]. It can be observed that most of the strategies work quite effectively in continuous cases where facilities can move continuously in any direction or where the density of candidate spots for facilities is quite large. However, when the distribution of candidate spots cannot be approximated continuously, as is always the case of BS or RRH locations in communication networks, they will not be so satisfactory or even lose effect. In [9], a given region $R$ is divided into $n$ subregions so as to balance the overall utilities on the subregions, and it inspires us that we can balance the workload among macro BSs and RRHs with similar algorithms.

## III. Notational Conventions and Proposed Strategy

Consider a given area $R$ made up of $n$ districts $\bigcup_i R_i = R$ is served by $n$ macro BSs $P = \{p_1, p_2, \ldots, p_n\}$ to deliver control signals and provide seamless coverage. It is reasonable to assume that $R$ is a connected, polygonal region with non-empty interior for practical communication networks. Without consideration of the offloading effect, we assume that each macro BS $p_i$ serves the district $R_i$, and each district should not overlap with others. As it is our natural desire that all cells should be connected, a penalty function denoted as $u_i(\cdot)$ is introduced to punish the objective function by preventing it from getting to its optimality when the macro cells are far from connected. In order to measure the connectivity of a macro cell, we may define $u_i(\cdot)$ to be the Euclidean distance between the user node and the macro BS $i$, which means $u_i(x) = \|x - p_i\|$. As we are considering a practical case where the distribution of traffic demands is far from uniform, the density of traffic demands across $R$ is statistically formulated as $f(x)$, where $x$ is a bi-vector representing coordinates. Thus the integral $\iint_{R_i} f(x)u_i(x)dA$ would denote the overall penalty of BS $i$. At last, we assume that $n_r$ RRHs are required to facilitate macro BSs to provide robust high speed data transmission.

Our suggested resource allocation strategy can be summarized as follows:

- Allocate resources among macro BSs. We first estimate the number of macro BSs, denoted as $n$, that is required to serve the objective region $R$ at the minimum data rate, as macro BSs are mainly deployed to deliver control signals and serve users requiring low data rates. Candidate sites for deploying macro BSs should be determined under the consideration of height, terrain, and the density of population. Then we design the service area of $R_i$ of each macro BS $i$ in a balanced fashion.
- Calculate the effective service area that can rely on macro BSs alone for robust data transmission. Since the step above has not taken into account factors such as propagation model, spectrum allocation, etc, some areas such as those which are far away from their corresponding macro BSs may not get reliable service. It is necessary that we calculate the effective service area in each subregion to make sure that RRHs will be deployed at other places for robust data transmission.
- Allocate resources among RRHs. Firstly, we estimate the number of RRHs, denoted as $n_r$, that is required to satisfy the traffic demands in areas that can not rely on macro BSs for robust data transmission. Then we collect the candidate sites for deploying RRHs. Finally, we design the service region of each RRH in these areas in a balanced fashion.

When the balance of the cellular system is broken or the total traffic demands vary drastically, this resource allocation scheme is triggered to adjust current network to a more balanced state.

## IV. Allocating Resources in H-CRANs Based on Workload Balancing

### A. Allocating Resources among Macro BSs

*1) Initialization:* Before allocating resources among macro BSs, we should determine the number of macro BSs required according to the traffic demands at the moment. Firstly, as macro BSs are mainly deployed to deliver control signals which requires relatively low transmission data rate, and serve users with low data rates, we set the number of macro BSs required under the condition that macro BSs serve the area $R$ at the minimum data rate. Reserving capacity margin is important for a practical communication network as it is impossible to guarantee that all macro BSs can provide the same capacity with the same transmission parameters. Besides, it can also help absorb the constant variations of traffic demands that are not violent. Secondly, we collect all the candidate sites for macro BSs all over the region $R$. If the number of macro BSs required at the moment is the same as before, the initial sites for macro BSs for starting the resource allocation scheme remain the same. When the number required is larger, we select extra macro BSs randomly from all candidate sites, making the starting condition for the scheme. Otherwise, we abandon the macro BSs facing the least traffic demands from formal macro BSs.

*2) Designing the Service Areas:* The most critical step in allocating resources among macro BSs is designing the service areas of them in a balanced fashion, and one way to realize it is to minimize the maximum traffic demand of all macro BSs with penalty phase while introducing constraints on the amount of $f(\cdot)$ that are served by them [9]. That is,

$$\begin{aligned} \min_{R} \quad & t+\mu\sum_{i=1}^{n}\iint_{R_i} f(x)u_i(x)dA \\ s.t. \quad C_1: \quad & t \geq (1-\mu)\iint_{R_i} f(x)dA, \forall i, \\ C_2: \quad & \iint_{R_i} dA \geq \Omega, \forall i, \\ C_3: \quad & R_i \cap R_j = O, \forall i \neq j, \\ C_4: \quad & \bigcup_i R_i = R. \end{aligned} \tag{1}$$

Here, we introduce a variable $\mu$ to represent the penalty phase, and a variable $t$ to represent the maximum value of the traffic demands all over macro BSs with the penalty factor, which is denoted as $C_1$. $C_2$ indicates that all cells should have an area larger than a constant denoted as $\Omega$, which helps guarantee that service areas of macro BSs do not differ too much from each other to avoid the case where some areas are too large to be covered by only one macro BS per region. Furthermore, it can also help avoid yielding ill-shaped regions, which guarantees the practicability of our proposal. For the purpose of simplification, we normalize the total area to constant one. So, the constant $\Omega$ in $C_2$ should be set to $1/n$. Without considering the offloading effect, $C_3$ presents the assumption that all service areas should never overlap with each other. $C_4$ denotes that there shouldn't be any coverage holes, which guarantees seamless coverage, and it is critical for macro BSs. Finally, the objective function here is designed to minimize the maximum traffic demands faced by all the macro BSs, with the value of the penalty function increasing when service areas are not as connected as we expect, to add obstructions to reaching the optimality.

In order to solve (1), we start by transforming the problem into the form of an infinite-dimensional integer program. By introducing a $\{0,1\}$-valued function $I_i(x)$ to indicate whether the demand point $x$ is served by macro BS $i$, our problem could be put as the equivalent formulation

$$\begin{aligned} \min_{I_1(\cdot),\ldots,I_n(\cdot)} \quad & t+\mu\sum_{i=1}^{n}\iint_{R} f(x)I_i(x)u_i(x)dA \\ s.t. \quad C_1: \quad & t \geq (1-\mu)\iint_{R} f(x)I_i(x)dA, \forall i, \\ C_2: \quad & \iint_{R} I_i(x)dA \geq \Omega, \forall i, \\ C_3: \quad & \sum_{i=1}^{n} I_i(x) = 1, \forall x, \\ C_4: \quad & I_i(x) \in \{0,1\}, \forall i, x. \end{aligned} \tag{2}$$

The most difficult part in solving (2) lies in the integer constraints. An intuitive way to cope with them is to relax the integer variables into continuous ones [10]. The linear

programming relaxation of (2) is given by

$$\begin{aligned} \min_{I_1(\cdot),\ldots,I_n(\cdot)} \quad & t+\mu\sum_{i=1}^{n}\iint_R f(x)I_i(x)u_i(x)dA \\ s.t. \quad & C_1 \sim C_3 \text{ in (2)}, \\ & I_i(x)\geq 0, \forall i,x. \end{aligned} \tag{3}$$

By introducing Lagrange multiplier vectors $\lambda \in R^n$ ,and $\gamma \in R^n$ , we can obtain the dual problem to (3) as follows,

$$\begin{aligned} \max_{\lambda,\gamma} \quad & \iint_R \min_i(\mu f(x)u_i(x)+(1-\mu)\lambda_i f(x)-\gamma_i)dA \\ & +\Omega\sum_{i=1}^{n}\gamma_i \\ s.t. \quad C_1: \quad & \lambda_i \geq 0, \forall i, \\ C_2: \quad & \sum_{i=1}^{n}\lambda_i = 1, \\ C_3: \quad & \gamma_i \geq 0, \forall i. \end{aligned} \tag{4}$$

Up to now, a convex, $2n$-dimensional dual problem is obtained [11]. It can be proven that (4) can be efficiently solved with many mature convex optimization techniques and tools. In this paper, we adopt a well-developed tool CVX, which is a modeling system for constructing and solving disciplined convex programs, to solve (4) [12].

After solving the dual problem (4), the dual variables $\lambda$ and $\gamma$ corresponding to the optimal solution to the original problem (1) are obtained. For any place, it will be served by the macro BS $p_i$ which minimizes $\mu f(x)u_i(x)+(1-\mu)\lambda_i f(x)-\gamma_i$ among $i \in [1,n]$.

Last but not the least, it still remains to be shown that the solution to (1) can be recovered from the optimal solution to (4). Consider any point $x \in R$ and the optimal solution to (4). Suppose $\bar{i}$ is the index such that $\mu f(x)u_i(x)+(1-\mu)\lambda_i f(x)-\gamma_i$ is minimal (assuming such an index is unique). From basic linear programming theory, we know that the complementary slackness conditions of problem (4) stipulate that $I_i^*(x)=0$ for all indices $i$ other than $\bar{i}$ [13], and consequently that $I_{\bar{i}}^*(x)=1$. In a conclusion, despite relaxation, the optimal solution to (3) remains valid as proved in [9].

*3) Decreasing Power Consumption:* After designing the service areas of macro BSs, the traffic demands faced by macro BSs have been balanced. As the selection of initial sites of macro BSs does not pay enough attention to the power consumption, even though the traffic demands are balanced after the step above, the power consumption has not been minimized. Since there may be multiple candidate sites to place macro BS in each service area, it is logical to select the one with the minimum power consumption. Considering the facts that not all places are suitable for placing macro BSs and that the number of candidate sites is limited in practical communication networks, we can adopt a straightforward method, exhaustive search, to obtain the best coordinates from all candidate sites.

Till now, the power consumption of every macro BS has been minimized based on the service areas we obtain before that. However, the design of service area is based on the initial coordinates of macro BSs that have nothing to do with the distribution of traffic demands or power consumption, which means that we can decrease power consumption further, by doing service area design and power decrease based on it in a loop until the total power consumption can not be decreased any more. By allocating resources among macro BSs in this balanced way, the case where some macro BSs suffer from heavy load while their adjacent ones are vacant can be avoided, which improves the utilization of macro BSs. Furthermore, by reducing the number of macro BS actually employed and deploying macro BSs in the place with the minimum power consumption, both the CAPEX and the OPEX can be reduced, and the EE performance is enhanced.

### *B. Calculating the Effective Service Area of Macro BSs*

After the step listed above, resources have been allocated to macro BSs in a balanced way. It can be easily figured out that as the step above has not taken into account factors such as propagation model, spectrum allocation, etc, some areas such as those which are far away from their corresponding macro BSs may only able to get control signals and may not get reliable data transmission. To guarantee that all areas are provided with quality of service (QoS) guaranteed service, RRHs should be deployed at places where traffic demands can not be satisfied by macro BSs. Before that, it is necessary that we calculate the effective service area of macro BSs.

Calculating the effective service area of macro BSs can be transferred into the traffic demand points (TDP) assignment problem, which is discussed about in detail in [14]. For lack of space, we adopt its result without much introduction.

### *C. Allocating Resources among RRHs*

The method for allocating resources among RRHs is almost the same with the one adopted for macro RRHs, except that we focus on the area $R$ that can not rely on macro BSs for robust data transmission. The area where we allocate resources among RRHs can be obtained in the step above. Since the resource allocation scheme is well-documented in the first step, we will not explore it here. It is worth mentioning that allocating resources among RRHs in a balanced way can not only guarantee that the minimum number of RRHs are activated, which enhances the EE performance of the whole system, but it also helps alleviate the high pressure expected in the fronthauls. The centralization of the processing units is accompanied by large numbers of wired or wireless fronthauls. On the other hand, the peak transmission bandwidth required in fronthauls is 10 times higher than during off-the-peak hours, and fronthauls are designed for the peak transmission bandwidth required to guarantee robust data transmission, which result in the large waste of transmission capability. Through balancing the resources allocated to RRHs, the peak transmission bandwidth required is reduced, thus alleviating the high pressure on the fronthauls.

Finally, as is expected, the resource allocation of last hour may not be suitable for this moment, as the distribution of traffic demand varies with time. On the other hand, it is not

TABLE I
RESOURCE ALLOCATION SCHEME PROPOSED

| | |
|---|---|
| 1: | *Initialize*: Monitor the fairness index of RRHs $\phi$, and collect all the candidate sites for macro BSs and RRHs; |
| 2: | Determine the number of macro BSs required and set initial sites for them; |
| 3: | Allocate resources in a balanced fashion among macro BSs; |
| 4: | Decrease power consumption of macro BSs through relocation; |
| 5: | **while**($\phi$ is not acceptable) |
| 6: | Calculate the effective service area of macro BSs; |
| 7: | Determine the number of RRHs required at the moment and set initial sites for RRHs; |
| 8: | Allocate resources in a balanced fashion among RRHs in the area that can not get robust data transmission; |
| 9: | Decrease power consumption of RRHs through relocation. |
| 10: | **end while** |

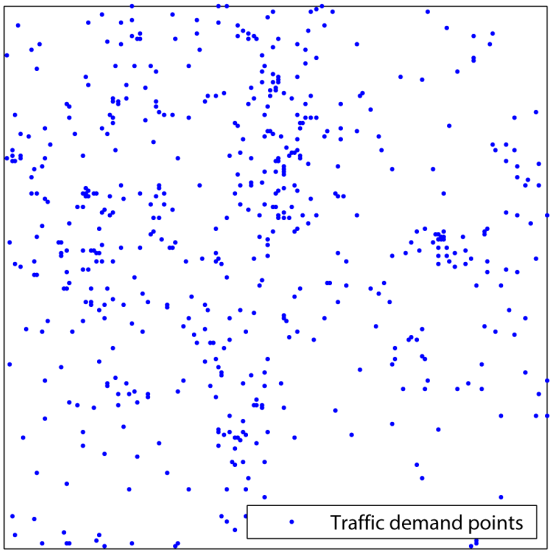


Fig. 2. Distribution of traffic demands in a real city.

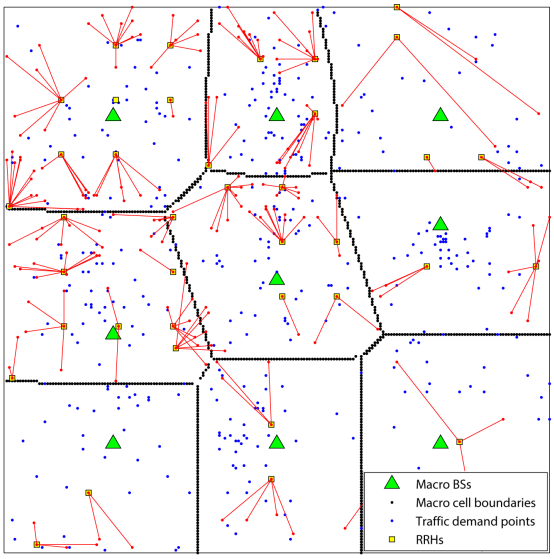


Fig. 3. Resource allocation among macro BSs and RRHs in a real city. Macro BSs are marked with green triangles with their service boundaries marked in black. RRHs are marked with yellow squares, connected to corresponding traffic demand points with red lines.

practical that we adjust the resource allocation frequently, as it can create vast handoff which results in unbearable signaling load in core network. Considering all these factors, we propose to introduce a parameter called fairness index, which is designed to quantificationally describe the balance degree of traffic demand among RRHs. As long as the fairness index is within the acceptable range, it is not necessary to adjust the network. Whenever it is paranormal, the resource allocation scheme is triggered to adjust current network to a more balanced state. To sum up, our resource allocation scheme can be presented in Table I.

## V. EXPERIMENTAL RESULTS AND DISCUSSIONS

We test our resource allocation scheme based on the distribution of traffic demand in a real city. The distribution of traffic demands are shown in blue dots, as illustrated in Fig. 2. It can be observed that there are three main traffic hot zones near the center, and the difference of density between hot zones and suburban areas (the border of the city) is quite large. The resource allocation results by using our scheme are presented in Fig. 3. It can be seen that RRHs are mostly deployed in traffic hot zones and other areas that can not rely on macro BSs for reliable communication.

The traffic demands faced by macro BSs are further reported in Table II. The standard deviation of traffic demands among macro BSs is 0.0455, with the average of them to be 0.1111. Considering the traffic demand gap among different access nodes in practical communication networks at the present, such tiny imbalance among macro BSs is totally acceptable. As we allocate resources among RRHs in areas that can not rely on macro BSs for reliable communication in each macro cell, we also list the standard deviation of traffic demands among RRHs in each macro cell in Table III. Comparing Table II and Table III, we can figure out that RRHs are mostly deployed in traffic hot zones. Even though tiny gap between traffic demands faced by RRHs is observed, such resource allocation scheme can reduce the peak transmission bandwidth required, which alleviates the high pressure on fronthauls, and reduces the number of RRHs activated to enhance the

TABLE II
NORMALIZED CAPACITY OF EACH MACRO BS

| Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Traffic demands | 0.0585 | 0.0556 | 0.1555 | 0.1083 | 0.0609 | 0.1556 | 0.1555 | 0.1555 | 0.0947 |

TABLE III
TRAFFIC DEMANDS SERVED BY RRHS

| Macro BS index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Number of RRHs required | 2 | 1 | 12 | 4 | 2 | 8 | 9 | 10 | 3 |
| Average normalized traffic (M) | 0.500 | 1.000 | 0.0833 | 0.2500 | 0.5000 | 0.1250 | 0.1111 | 0.1000 | 0.3333 |
| Standard deviation of normalized traffic (STD) | 0.0886 | 1.0000 | 0.0782 | 0.1153 | 0.0000 | 0.1152 | 0.0998 | 0.0773 | 0.2077 |
| STD/M | 0.1773 | 1.0000 | 0.9390 | 0.4613 | 0.0000 | 0.9000 | 0.8982 | 0.7735 | 0.6230 |

EE of the whole system. Last but not the least, it is worth mentioning that when the number of macro BSs or RRHs required changes, we select extra BSs/RRHs or abandon some BSs/RRHs. The "select" and "abandon" here do not mean constructing or deleting some sites for macro BSs/RRHs, but mean activating or deactivating macro BSs/RRHs. Last but not the least, the computational complexity of our resource allocation scheme is the combination of the Designing the Service Area part and the Decreasing Power Consumption part. The computational complexity of the Designing the Service Area part is $O(n^3)$, where $n$ is the number of access points required. Since we adopt a straightforward method, exhaustive search, for the Decreasing Power Consumption part, the computational complexity of this part is exponential with $n$, where $n$ is the number of candidate sites for macro BSs and RRHs. Despite the high computational complexity, the algorithm is still effective since the number of candidate sites for access points is always small and limited.

## VI. CONCLUSION

In this paper, we presented a novel method for allocating resources in the H-CRANs. We show that the radio resources in the H-CRANs can be distributed in a balanced fashion in order that the least number of macro BSs and RRHs is required. Moreover, the high pressure on fronthauls between RRHs and the BBU pool can also be alleviated. Our resource allocation scheme is divided into three parts. The first step is to allocate resources among macro BSs under the condition that the traffic demands served by macro BSs should be as balanced as possible and that macro BSs should provide seamless coverage. The second step is to calculate the effective service area that can rely on macro BSs for robust data transmission. The third step is to allocate resources among RRHs in the opposite area in a balanced way. Through balancing the traffic demands faced by RRHs, the peak data transmission is decreased, which minimizes the transmission bandwidth of fronthauls required. Besides, balancing the traffic demands faced by macro BSs and RRHs can reduce the number of macro BSs and RRHs actually required, which improves the utilization of access points and decrease energy consumption. Experimental results verify the effectiveness and the efficiency of our proposal. For future work, extensive experiments to evaluate our scheme should be conducted, especially on the much handover generated when the service areas of RRHs changes. Moreover, the complexity of the scheme should be decreased in the future under the consideration of real-time performance.

## ACKNOWLEDGEMENT

This work was partially supported by JiangsuSF (BK20151389) and the Fundamental Research Funds for the Central Universities (021014380013).

## REFERENCES

[1] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. Thomas, J. Andrews, P. Xia, H. Jo, H. Dhillon, and T. Novlan, "Heterogeneous cellular networks: From theory to practice," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 54–64, June 2012.

[2] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[3] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, June 2011.

[4] X. Tao, X. Xu, and Q. Cui, "An overview of cooperative communications," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 65–71, June 2012.

[5] C. Ran, S. Wang, and C. Wang, "Balancing backhaul load in heterogeneous cloud radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 42–48, June 2015.

[6] ——, "Cellular networks planning: A workload balancing perspective," *Comput. Netw.*, vol. 84, no. 19, pp. 64–75, June 2015.

[7] O. Baron, O. Berman, D. Krass, and Q. Wang, "The equitable location problem on the plane," *Eur. J. Oper. Res.*, vol. 183, no. 2, pp. 578–590, Dec. 2007.

[8] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in HetNets: Old myths and open problems," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 18–25, Apr. 2014.

[9] J. G. Carlsson and R. Devulapalli, "Shadow prices in territory division," *University of Minnesota, Available at: http://menet. umn. edu/˜ jgc/shadow-prices-rev2. pdf*, 2013.

[10] M. Ge and S. Wang, "Fast optimal resource allocation is possible for multiuser OFDM-based cognitive radio networks with heterogeneous services," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1500–1509, Apr. 2012.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press: New York, 2004.

[12] M. Grant, S. Boyd, and Y. Ye, "cvx users' guide," Technical Report Build 711, Citeseer. Available at: http://citeseerx. ist. psu. edu/viewdoc/download, Tech. Rep., 2009.

[13] D. G. Luenberger and Y. Ye, *Linear and nonlinear programming*. Springer, 2008, vol. 116.

[14] S. Wang, W. Zhao, and C. Wang, "Budgeted cell planning for cellular networks with small cells," *IEEE Trans. Veh. Technol.*, Nov. 2014.