Evaluating Energy-Efficient Cloud Radio Access Networks for 5G

Tshiamo Sigwele, Atm Shafiul Alam, Prashant Pillai, and Y. Fun Hu
Faculty of Engineering and Informatics
University of Bradford, Bradford BD7 1DP
West Yorkshire, United Kingdom
Email: {T.Sigwele, A.S.Alam5, P.Pillai, and Y.F.Hu}@bradford.ac.uk

Abstract—Next-generation cellular networks such as fifth-generation (5G) will experience tremendous growth in traffic. To accommodate such traffic demand, there is a necessity to increase the network capacity that eventually requires the deployment of more base stations (BSs). Nevertheless, BSs are very expensive and consume a significant amount of energy. Meanwhile, *cloud radio access networks* (C-RAN) has been proposed as an energy-efficient architecture that leverages cloud computing technology where baseband processing is performed in the cloud, i.e., the computing servers or baseband processing units (BBUs) are located in the cloud. With such an arrangement, more energy saving gains can be achieved by reducing the number of BBUs used. This paper proposes a bin packing scheme with three variants such as First-fit (FT), First-fit decreasing (FFD) and Next-fit (NF) for minimizing energy consumption in 5G C-RAN. The number of BBUs are reduced by matching the right amount of baseband computing load with traffic load. In the proposed scheme, BS traffic items that are mapped into processing requirements, are to be packed into computing servers, called bins, such that the number of bins used are minimized and idle servers can then be switched off to save energy. Simulation results demonstrate that the proposed bin packing scheme achieves an enhanced energy performance compared to the existing distributed BS architecture.

Index Terms—BBU reduction, Cloud Computing, C-RAN, Energy-Efficiency, Virtualization.

I. Introduction

Recently, the number of connected devices has grown into billions and today's mobile operators are facing a serious challenge. For example, according to Huawei Technologies, 100 billions of devices will be connected by 2020 [1]. This will cause a surge in traffic and this traffic growth will be from smart phones and other high-end devices like the Android, iPad, Kindle and gaming consoles spawning a raft of data intensive applications, Internet of Things (IoT) and machine-to-machine connections. As a result, next-generation networks such as fifth-generation (5G) have received unprecedented expectations with targeting to increase 1000 times capacity, 100 times data rate, and millisecond-level delay [2]. To fulfil these capacity demands, more base stations (BSs) with a mixer of macro and small cells have to be deployed by operators. Nevertheless, BSs are very expensive and consume over a significant amount of energy [3]. This contributes to the mobile networks Operating Expenditure (OPEX) and causes a greater impact to the environment by emitting large amounts of CO_2.

Within each BS, a large amount of power is consumed by the power amplifier and baseband unit (BBU) or computing servers. The energy consumption of BBU implementation is getting more and more dominant in small cells [4] due to gradual shrinking of cell size and the growing complexity of signal processing. Hence, it is crucial to optimize energy-efficiency in the BBU servers. Many energy-efficient schemes for wireless systems have been implemented such as BS sleeping where offloading traffic to neighbouring BSs and then completely turning off the BS during low traffic [5], discontinuous transmission (DTX) where a BS is temporally switched off without offloading [6], cell zooming [7], utilizing renewable energy sources [8]. However, research on energy-efficient schemes for cloud BSs in centralised RANs is mostly ignored.

Traditional distributed BSs consume a significant amount of energy and waste a lot of computing power as the BBU processing servers are not shared but serve each individual cell [3], [9]. Traditionally, BSs have been preconfigured to provide peak capacities to reduce outages. Nevertheless, the mobile traffic varies significantly, irrespective of the either the time of day or traffic profile and is rarely at its peak in practical scenarios [10]. This means traffic load in a cellular network changes gradually in a time-geometry pattern called the Tidal Effect, which is the fluctuation of traffic load in the BS due to the dramatic subscriber density increase in both business and residential areas. However, current BS processing capacity is only being used for its own coverage rather than being shared in a large geographical area. Therefore, during the evening BSs in residentinitiallyial areas are over-subscribed while BSs in business areas stay under-subcribed. However, these under-subcribed BSs still consume a significant amount of energy even when they are not necessarily required to be kept active. Therefore, it is imperative to solve this problem and free up the processing capacity and save the corresponding energy.

Recently, cloud radio access networks (C-RAN) have been proposed as a promising solution for conserving energy within the cellular networks that leverages cloud computing technology [3]. C-RAN comprise of three parts: (i) remote radio head (RRH), which performs lower layer analogue radio frequency (RF) functions, (ii) BBU for digital signal processing, and (iii) fronthaul connection between the BBU and RRH. In C-RAN, digital baseband processing of multiple distributed
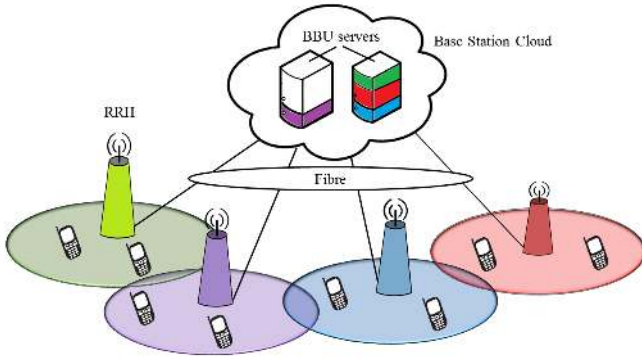
Fig. 1. An illustration of a C-RAN Architecture.

RRH is performed in the cloud and dynamically provisioned according to traffic demands. Furthermore, more energy savings can be gained from reduced air conditioning cost and reduced equipment room size. This paper proposes a scheme for evaluating the enrgy-efficiency of the C-RAN in the 5G networks. The number of computing servers or BBUs are reduced by matching the right amount of baseband processing load with respect to the traffic load. This is formulated as a bin packing problem with three variants such as First Fit (FT), First Fit Decreasing (FFD) and Next Fit (NF) where BS traffic items are to be packed into compute servers, called bins, such that the number of bins used are minimized and idle servers can then be switched off to conserve energy consumption. The BS traffic items are mapped into the processing requirements. The simulations results validates the energy-efficiency improvement of the proposed bin packing algorithms and is compared the distributed BS architecture.

This paper is structured as follows: Section II discusses the related works while the C-RAN architecture is described in Section III. The proposed scheme with the system and computational resource models are formally introduced in Section IV. The problem formulation and heuristic algorithms are also proposed. Section V provides the simulation results and discussion, while providing some concluding comments.

II. RELATED WORKS

Recently, energy-efficiency of cellular networks has attracted research communities both in academia and industry due to the social, economical and environmental pressures. The advantages of energy-efficient RANs are mainly due to the reduction OPEX and the amount of CO_2 emission. There are a plethora of solutions towards energy-efficient BSs ranging from energy-efficient hardware design, BS sleeping, to the optimal deployment of BSs [5], [11].

Authors in [12] proposed a BBU-RRH switching scheme for C-RAN that dynamically allocates BBUs to RRHs based on the imbalance of subscribers in business and/or residential areas. An upper limit threshold of BBU utilization is defined to avoid overloading of the BBU. Even though the scheme in [12] reduces the number of BBUs required, the model performs poorly during high-traffic periods and thus still consumes a lot of energy because more BBUs are allocated to meet traffic demands. Authors in [13] developed a BBU pool testbed using virtualization technology and GNU radio platform on general purpose processors. The BBUs are dynamically provisioned according to traffic load. However, the paper fails to show how the number of BBUs are reduced while traffic load varies. In addition, Linux Operating System (OS) assisted virtualization is used, which adds more delays and jitter when performing baseband processing on virtual BSs.

L. Jingchu et al. [14] presented a mathematical model to quantify the statistical multiplexing gain of pooling virtual BSs. The authors use a multi-dimensional Markov model to evaluate pooling gain considering both compute and radio resources. Nevertheless, the authors have not considered energy consumption in the BS-Cloud. In [15], the authors proposed an analytical energy model of a computational resource-aware virtual BSs in a cloud-based cellular network architecture. The authors consider the energy-delay trade-offs of a virtual BSs considering the BS sleeping approach in general IT platforms. The paper does not show how the energy savings of the virtual BSs model scales with traffic load. In [16], authors developed an OFDMA-based C-RAN test-bed with a reconfigurable backhaul that allows four BBUs to connect flexibly with four RHHs using radio-over-fiber technology. The backhaul architecture allows the mapping between BBUs and RRHs to be flexible and changed dynamically to reduce energy consumption in the BBU pool. However, the paper assumes static user traffic whereas in reality BS traffic is dynamic. S. Namba et al. [17] proposed a network architecture, called colony-RAN, which has the ability to flexibly change cell layout by changing the connections of BBUs and RRHs in respect to the traffic demand. However, the proposed method has frequent reselections of RRH to BBU, i.e., ping-pong effects.

III. C-RAN: ARCHITECTURE

The C-RAN comprises of the 4 Cs which stand for centralized, collaborative, cooperative and clean/green [3], [9]. The C-RAN architecture adopted in this paper is shown in Fig. 1. The BBUs are separated from the cell areas and centralized in the BS-Cloud or BS Pool leaving only the RF RRH in the cell sites. High bandwidth fiber links are then used to link the distributed RRHs to the BS-Cloud. The BBU performs digital baseband processing functions such as Physical (PHY) and Media Access Control (MAC) layer. The BS-Cloud is deployed on an open IT architecture using software-defined radio (SDR) technology. The SDR technology involves the implementation of all wireless baseband processing in software. The main advantages of C-RAN architecture are:

- Reduction in air conditioning and other onsite power-consuming equipment.
- A single BBU can process multiple RRH baseband signals simultaneously due to SDR technology.
- Any RRH traffic load can be processed on any BBU.

The main drawback for C-RAN is that the fronthaul links require high bandwidth and low latency due to the transmission

of digital Inphase/Quadrature (IQ) signals. Fibre optic cables is the most suitable fronthaul link but are costly.

IV. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

The proposed system model is shown in Fig. 2. Assume a set of RRHs $\mathcal{R} = [RRH_j : j = 1, 2, ..., N]$ where N is the maximum number of RRHs serving N cell areas. Moreover, assume a set of computing servers in the pool $\mathcal{M} = [S_i : i = 1, 2, ..., M]$ where M is the number of physical computing servers for processing baseband signals of N cells. The baseband processing procedure of each RRH is divided into L tasks with a set $\mathcal{L} = [T_k : k = 1, 2, ..., L]$ where T_k is the kth baseband task for RRH_j. The computing processing power is measured in *Giga Operations Per Second* (GOPS). Each server has maximum capacity C GOPS. The total computing resources required by RRH_j is denoted ρ_j^{req} GOPS such that:

$$\rho_j^{req} = \sum_{k=1}^{L} \rho_{j,k}^{req}; \qquad \rho_{j,k}^{req} \epsilon (0, 1] \tag{1}$$

where $\rho_{j,k}^{req}$ is the computing resource requirement for T_k from RRH_j.

Therefore, the computing resource at server S_i used by RRH_j can be calculated as:

$$\rho_{i,j}^{server} = \sum_{k=1}^{L} \xi_{i,j,k} \rho_{j,k}^{req}; \qquad \xi_{i,j,k} \epsilon \{0, 1\} \tag{2}$$

where $\xi_{i,j,k} = 1$ when T_k from RRH_j is processed by server S_i and $\xi_{i,j,k} = 0$ otherwise. Tasks from RRH_j can be processed by a single server or distributed among different serves such that the constraint below hold:

$$\sum_{i=1}^{M} \sum_{k=1}^{L} \xi_{i,j,k} = L \tag{3}$$

And the BBU server processing is limited by server capacity C as:

$$\sum_{j=1}^{N} \rho_{i,j}^{server} \leq C \tag{4}$$

Even though there will be communication costs incurred between servers' processing tasks from the same RRH, the ideal server inter-communication is assumed. The energy minimization in the cloud for M BBU servers can be formulated from two components [10]: dynamic and static power consumption. The dynamic energy consumption is dependent on the amount of processing resources on the server while the static part comprises the energy consumption irrespective of traffic load, but other purposes such as coolings, etc. Now, the energy minimization problem can be formulated as:

$$\min_{\xi_{i,j,k}} \quad \sum_{i=1}^{M} \left(\delta \sum_{j=1}^{N} \rho_{i,j}^{server} + \varepsilon_i P_{static} \right) \tag{5}$$
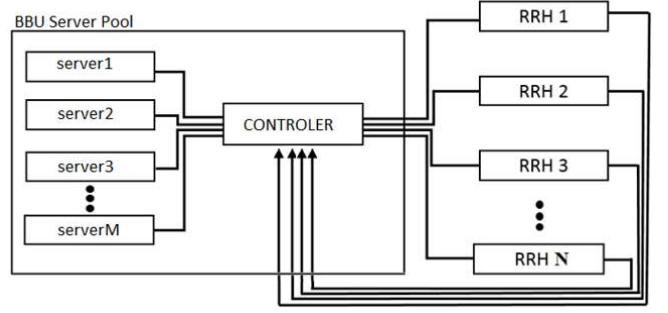

Fig. 2. System Model.

$$\varepsilon_i = \begin{cases} 0 & \sum_{j=1}^{N} \sum_{k=1}^{L} \xi_{i,j,k} = 0 \\ 1 & \text{Otherwise} \end{cases} \tag{6}$$

where δ is the power factor in GOPS/watts. ε_i shows the status factor of server S_i whether S_i is ON or OFF. P_{static} denotes the static power that is constant for every BBU server. Constraints are from (3) and (4).

B. Computational Resource Model

The baseband tasks from cells need to be quantified, i.e., they need to be mapped into computing processing in GOPS. The computing resource requirement per user per task is calculated based on the energy consumption model in [18]. The model provides energy modelling for different types of BSs such as macro, micro, pico and femto BSs. The micro BSs will be selected to represent an RRH. The authors in [18] give the power budget for BS components in the uplink and downlink. The complexity for BS functionalities in GOPS are defined per function block and how they scale with transmission mode and load. The complexity values given are used as reference cases for the baseband in downlink. The reference case assumes 20 MHz bandwidth, which is equivalent to 100 physical resource blocks (PRBs) in LTE systems, single-antenna, 64-QAM, coding rate 1 and 100% load. Note that, the load is defined as a fractional use of time and frequency resources, i.e., load = dtxdf). Scaling tables in [18] contain specific scaling factors telling how each power figure evolves with each specific parameter.

In this paper, the power equation in [18] for calculating the computing resources required for baseband tasks is adopted. Defining $\mathcal{L}$ as the set of baseband subcomponents' tasks and $\mathcal{X} = \{BW, Ant, M, R\}$ (see Table I) is the list of parameters affecting the scaling of baseband processing tasks. Now, the following expression can be written as:

$$P_{u,t} = \sum_{i \epsilon L} P_{i,ref} \prod_{x \epsilon X} \left(\frac{x_{act}}{x_{ref}} \right)_t^{s_{i,x}} \tag{7}$$

where $P_{u,t}$ is processing power required by user u at time t. $P_{i,ref}$ denote the processing power of reference system. Users that generate traffic are randomly distributed in the cell area and the generated traffic are mapped into processing

Algorithm 1: Next Fit Algorithm

Input: a set of N RRH cells $\mathcal{R}$, a set of tasks $\mathcal{L}$ within RRH_j, their resource requirements $\rho_{j,k}^{req}$, and the single server capacity C.
Output: Number of BBUs M
for *each $\rho_{j,k}^{req}$ that arrives* **do**
    if *there is a server where $\rho_{j,k}^{req}$ will fit* **then**
        Place $\rho_{j,k}^{req}$ into the left most server;
    else
        Close the server and launch a new server;
        Place $\rho_{j,k}^{req}$ into that server
    end
end
Return M

Algorithm 2: First-Fit Algorithm

Input: a set of N RRH cells $\mathcal{R}$, their resource requirements $\rho_{j,k}^{req}$, and the single server capacity C.
Output: Number of BBUs M
for *each $\rho_{j,k}^{req}$ that arrives* **do**
    if *there is a server where $\rho_{j,k}^{req}$ will fit* **then**
        Place $\rho_{j,k}^{req}$ into the left most server;
    else
        Launch a new server;
        Place $\rho_{j,k}^{req}$ into that server
    end
end
Return M

Algorithm 3: First-Fit Decreasing Algorithm

Input: a set of N RRH cells $\mathcal{R}$, a set of tasks $\mathcal{L}$ within RRH_j, their resource requirements $\rho_{j,k}^{req}$, and the single server capacity C.
Output: Number of BBUs M
Sort all RRH tasks in decreasing order of $\rho_{j,k}^{req}$.
Launch one server of capacity .
for *each $\rho_{j,k}^{req}$ that arrives* **do**
    if *there is a server where $\rho_{j,k}^{req}$ will fit* **then**
        Place $\rho_{j,k}^{req}$ into the left most server;
    else
        Launch a new server;
        Place $\rho_{j,k}^{req}$ into that server
    end
end
Return M

resources as per user per task. Even though there are many baseband tasks processed by a BS, this paper considers two baseband tasks for simplicity, i.e., $k = 2$: (i) Frequency-Domain (FD) processing for mapping/demapping and MIMO equalization, and (ii) Forward Error Correction (FEC) denoted by the following equations, respectively, in GOPS:

$$P_{u,t}^{FD} = \left(30Ant + 10Ant^2\right)\frac{R}{100} \tag{8}$$

$$P_{u,t}^{FEC} = 20\frac{M}{6}C * Ant * \frac{R}{100} \tag{9}$$

where $P_{u,t}^{FD}$ and $P_{u,t}^{FEC}$ are FD and FEC processing requirements, respectively, per user u per task k at time t in GOPS. Ant is the number of antennas used per user, M is the modulation bits, C is the coding rate used and R is the number of PRBs used by u at time t. In the bin packing algorithm, the tasks per cell are packed on servers hence the processing requirements per task per cell for the two tasks is calculated as follows:

$$\begin{cases} \rho_{j,1}^{req} = \sum_{u\epsilon\mathbf{U}} P_{u,t}^{FD}, & \text{when } k = 1 \\ \rho_{j,2}^{req} = \sum_{u\epsilon\mathbf{U}} P_{u,t}^{FEC}, & \text{when } k = 2 \end{cases} \tag{10}$$

where $\mathbf{U}$ is the set of users within a cell.

C. Bin Packing Scheme

The classical bin packing problem consists of packing a series of items with different sizes into a minimum number of bins with capacity C. The resource allocation can be modelled as the bin packing problem where the aim is to pack items, called baseband tasks $\mathcal{L}$, from cell areas $\mathcal{R}$ into a set of servers $\mathcal{M}$ such that the number of servers used are minimized and hence the energy consumption reduction. This problem has a variety of applications ranging from traditional stock-cutting problems, computer-disks allocation problem, to the packing of television commercials into station breaks. Since the problem of finding optimal packings is NP-hard, i.e., there is no way of being guaranteed the best solution

without checking every possible solutions. Amongst many other solutions, the approximation algorithm is the mostly used because of fast heuristics that generate good but not necessarily optimal packings. Three approximation algorithms are considered in this paper: Next-fit (NF), First-fit (FF) and First-fit decreasing (FFD).

The NF algorithm as illustrated in Algoirthm 1 works as follows: Initially all bins are empty and we start with bin $j = 1$ and item $i = 1$. If bin j has residual capacity for item i, assign item i to bin j, i.e., $a(i) = j$, and repeat the process for the next item, $i + 1$. Otherwise close bin j and assign bin $j + 1$ for item i. Repeat until all items are assigned. The NF algorithm never considers bins again that have been left behind. However, the wasted capacity therein leaves room for improvement. Therefore, the FF algorithm illustrated in Algorithm 2 has been proposed, which is a natural way of finding the approximation. In this algorithm, all bins are initially empty and it starts with the current number of bins K and item $i = 1$. Consider all bins $j = 1, ..., K$ and place item i in the first bin that has sufficient residual capacity.

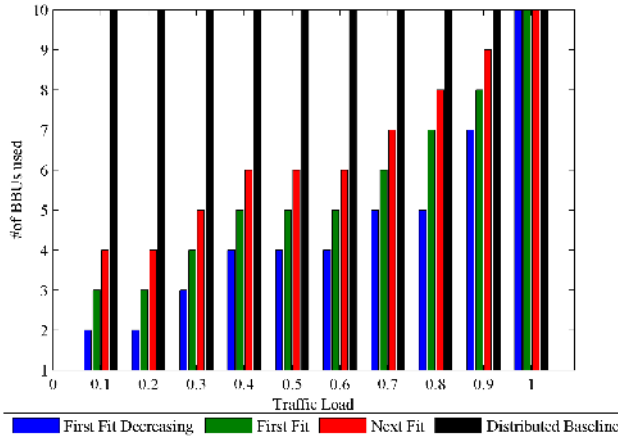| Notations	Descriptions	Range	Reference
BW	Bandwidth	10 (50 PRBs)	20 (100 PRBs)
Ant	Number of antennas	1, 2, 3	1
M	Modulation	1, 2, 4, 6	6
C	Coding rate	1/3 - 1	1

Fig. 3. Number of BBUs used in the cloud.

If there is no such bin, increment K and repeat until all items is assigned. The FFD approximation algorithm as shown in Algorithm 3 is the same as the FF algorithm but all items are first sorted in descending order.

V. SIMULATION RESULTS AND DISCUSSION

A. Parameter Settings

To analyse the performance of the proposed bin packing approximation algorithms, a simulation layout of 10 cells comprising of a maximum of 10 BBUs was considered. Bandwidth of 10 MHz was considered and 100 users in total are generated within the cells. Each cell is divided into three concentric circles, where users located at circle closer to the cell center use high modulation order and users at the edge of the cell use lower order modulation scheme. Other simulation parameters are shown in Table I. All results using the three bin packing algorithms are compared with the baseline system which comprises of distributed BSs with 10 individual BBU processing servers for 10 cells.

The users are allocated PRB per transmission time interval (TTI), whcih is 1ms, and the scheduling scheme used is proportional-fairness such that users are allocated equal number of PRBs per TTI. Adaptive Modulation and Coding scheme is used to adapt to the changing channel conditions. As the simulation runs, the values of Ant, M, C, R were captured and mapped into processing requirements and loaded into the bin packing scheme to reduce the number of servers M. For computing the power consumption, the power factor used is $\delta = 40$ GOPS/watt and $P_{static} = 200$ GOPS as in [18].
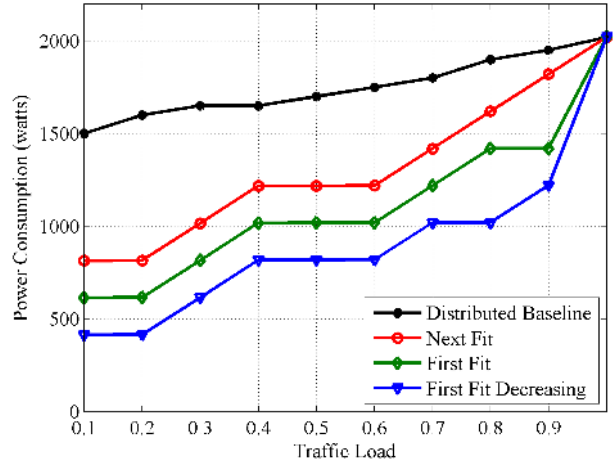
Fig. 4. Power Consumption in the cloud.

B. Results Evaluation

BBU server reduction: Fig. 3 shows the number of servers used in the cellular network for different traffic loads. The results show that as the traffic load increases, more BBU servers are required. When traffic load is at its peak, all BBUs are kept in active mode to process all traffic loads in all cases. It can be seen from the graph that the FFD approximation algorithm outperforms compared to other schemes in minimizing the number of servers used. This is because it first sorts items and then packs the bigger (harder) traffic item first. The smaller traffic items are usually packed at the last. On the other hand, the NF algorithm performs poorly since items are not ordered and also certain servers have closed before items are fully packed in it. The baseline scheme is expected to use more number of BBU servers and equals to the number of cells or RRHs, in this case 10 servers.

Energy minimization: Fig. 4 illustrates the energy consumption for each algorithm versus traffic load. The power consumption by employing all algorihms was increased with the traffic load. At the peak traffic load, maximum power was consumed for all algorithms because all BBUs were utilized. The FFD approximation algorithm consumes less power than others due to the utilization of the fewer number of servers. During the low-traffic periods, the FFD algorithm was able to save power consumption up to 73%. The baseline system consumes more power as expected because all 10 BBU servers were always utilized.

VI. CONCLUSION

This paper presented an energy-efficient cloud radio access networks (C-RAN) for 5G with employing three bin packing approximation algorithms, namely Next-fit (NF), First-fit (FF) and First-fit decreasing (FFD). The number of computing servers are reduced by matching the right amount of baseband processing with traffic load. For the NF algorithm, some of bins (BBUs) are not fully utilized because those bins are closed before they are filled up, whereas the FF algorithm

places items (processing tasks) where they fit without closing any underutilized bins. For the FFD algorithm, items are first sorted in descending order and then placed according to the FF approximation algorithm. The FFD algorithm provided the maximum energy savings among others, which is up to 73% during low-traffic periods. In future, this energy-efficient scheme will be extended for further conserving energy consumption by integrating switching on/off cells.

REFERENCES

[1] Huawei, "5g: a technology vision," *White Paper of Huawei Tech*, 2013.

[2] I. Chih-Lin, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: A 5G perspective," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 66–73, 2014.

[3] C. Mobile, "C-ran road towards green radio access network," *C-RAN International Workshop*, 2010.

[4] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume *et al.*, "How much energy is needed to run a wireless network?" *Wireless Communications, IEEE*, vol. 18, no. 5, pp. 40–49, 2011.

[5] E. Yaacoub, "A practical approach for base station on/off switching in green lte-a hetnets," in *Wireless and Mobile Computing, Networking and Communications (WiMob), 2014 IEEE 10th International Conference on*. IEEE, 2014, pp. 159–164.

[6] J.-F. Cheng, H. Koorapaty, P. Frenger, D. Larsson, and S. Falahati, "Energy efficiency performance of lte dynamic base station downlink dtx operation," in *Vehicular Technology Conference (VTC Spring), 2014 IEEE 79th*. IEEE, 2014, pp. 1–5.

[7] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 74–79, 2010.

[8] C. Murthy and C. Kavitha, "A survey of green base stations in cellular networks," *International Journal of Computer Networks and Wireless Communications (IJCNWC)*, vol. 2, no. 2, pp. 232–236, 2012.

[9] T. Sigwele, P. Pillai, and Y. F. Hu, "Call admission control in cloud radio access networks," in *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*. IEEE, 2014, pp. 31–36.

[10] A. S. Alam, L. S. Dooley, and A. S. Poulton, "Energy efficient relay-assisted cellular network model using base station switching," in *Global Telecommunications Workshop (GLOBECOM 2012) on Multicell Cooperation (MuCo), 2012 IEEE*, Anaheim, CA, USA, 2012.

[11] A. S. Alam and L. S. Dooley, "A Scalable Multimode Base Station Switching Model for Green Cellular Networks," in *IEEE Wireless Communication and Networking Conference (WCNC)*, 2015, pp. 1–6.

[12] S. Namba, T. Warabino, and S. Kaneko, "Bbu-rrh switching schemes for centralized ran," in *Communications and Networking in China (CHINACOM), 2012 7th International ICST Conference on*. IEEE, 2012, pp. 762–766.

[13] Z. Kong, J. Gong, C.-Z. Xu, K. Wang, and J. Rao, "ebase: A base-band unit cluster testbed to improve energy-efficiency for cloud radio access network," in *Communications (ICC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 4222–4227.

[14] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, "On the statistical multi-plexing gain of virtual base station pools," in *Global Communications Conference (GLOBECOM), 2014 IEEE*. IEEE, 2014, pp. 2283–2288.

[15] T. Zhao, J. Wu, S. Zhou, and Z. Niu, "Energy-delay tradeoffs of virtual base stations with a computational-resource-aware energy consumption model," in *Communication Systems (ICCS), 2014 IEEE International Conference on*. IEEE, 2014, pp. 26–30.

[16] C. Liu, K. Sundaresan, M. Jiang, S. Rangarajan, and G.-K. Chang, "The case for re-configurable backhaul in cloud-ran based small cell networks," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 1124–1132.

[17] S. Namba, T. Matsunaka, T. Warabino, S. Kaneko, and Y. Kishi, "Colony-ran architecture for future cellular network," in *Future Network & Mobile Summit (FutureNetw), 2012*. IEEE, 2012, pp. 1–8.

[18] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. J. Gonzalez *et al.*, "Flexible power modeling of lte base stations," in *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*. IEEE, 2012, pp. 2858–2862.