Offloading decision algorithm for 5G/HetNets cloud RAN

Olfa Chabbouh, Sonia Ben Rejeb, Zièd Choukair, Nazim Agoulmine

Offloading Decision Algorithm for 5G/HetNets Cloud RAN

Olfa CHABBOUH
MEDIATRON Laboratory
Higher School of Communication of Tunis, SupCom, Tunisia
olfa.chabbouh@supcom.tn

Sonia BEN REJEB
MEDIATRON Laboratory
Higher School of Communication of Tunis, SupCom, Tunisia
Sonia.benrejeb@supcom.tn

Zied CHOUKAIR
MEDIATRON Laboratory
Higher School of Communication of Tunis, SupCom, Tunisia
z.choukair@supcom.tn

Nazim AGOULMINE
IBISC – IBGBI Laboratory
University of Evry-Val - d'Essonne, France
Nazim.Agoulmine@ibisc.univ-evry.fr

***Abstract*— The proliferation of mobile handsets has undoubtedly spurred on developers to build a large variety of applications that allow users to better exploit their powerful devices. However, running multiple sophisticated applications could result in poor performances and shortened battery lifetime. Mobile data offloading to edge cloud offers a promising solution to enhance both QoS and battery life of mobile terminals. In this paper, a novel multi parameters offloading decision algorithm based on cloud radio access network (C-RAN) is proposed. It takes decision about offloading mobile computation to Cloud-RRH in a 5G heterogeneous C-RAN. Unlike previous works, the proposed scheme incorporates a multitude of parameters in the offloading decision process while diminishing the mobile devise energy consumption and keeping a good user quality of experience by reducing application's response time. Simulation results show that the proposed algorithm is able to assure the computation of all the applications while respecting latency constraints and to extend the mobile battery lifetime.**

Keywords— Cloud RAN, Cloud RRH, 5G/HetNet, mobile data offloading, energy-saving, QoS.

I. INTRODUCTION

With the exponential growth of data traffic, it becomes more and more difficult for telecom operators to upgrade their Radio Access Networks (RAN) because the revenue is not growing at the same rate. To maintain profitability and growth, they must find solutions to reduce cost as well as to provide better services to the customers. In this context the Cloud RAN architecture was proposed. The concept was first introduced in [1] and described in detail in [2]. It consists of decoupling the Base Band Units (BBUs) from radio remote heads (RRHs) and move it to the cloud enabling a centralized processing and management. Traditional complicated base stations can be simplified to cost-effective and power-efficient radio units (RRHs). In addition, the centralized processing power enables more advanced and efficient network coordination and management.

A recent survey on computation offloading for mobile systems was described in [3]. Offloading may be performed at the levels of methods, tasks, applications, or virtual machines. Cloud computing is used to allow offloading at the virtual machine level [4][5][6][7]. It grants elastic resources and offloading to multiple servers. Several architectures and solutions have been proposed to improve offloading: they address different issues such as transparency to users, privacy, security, mobility, etc. For example, the femtocloud was proposed in the European TROPIC project [19] for LTE networks. This new paradigm consist of a distributed set of femtocell access points (FAPs) which combines the radio resource management of femtocell networks with cloud storage and computation capabilities in the same framework [8]. Thus, the traffic offloading is reduced thanks to these additional cloud resources closer to mobile user. In order to optimize the Cloud RAN architecture and following the same intuition, we have proposed the Cloud-RRH. It consist of upgrading High-RRH capabilities by adding extra resources of computing and storage located in the cloud. Bringing resources closer to the user improves not only power consumption at the terminal side but also the other major issue, latency. Having a hierarchical cloud will allow offloading of applications, partially or totally, from user equipment to Cloud-RRH.

Traditional mobile computation offloading algorithms incorporates only energy consumption or latency in offloading decision. However, many other parameters should be considered in order to have the solution that best fit system conditions. In fact, several offloading decision schemes were proposed in the literature, but the originality of our algorithm is that it takes into consideration several parameters related to network state and mobile terminal mobility speed and capabilities while reducing the offloading decision process complexity.

The remainder of this paper is organized as follows. In the next section, previous works are discussed. Section III describes the system model and basic idea of our offloading decision algorithm for C-RANs. Then, simulation results are presented in section IV. Finally, Section V concludes this paper.

II. RELATED WORK

With the explosion of Internet data traffic, especially the growing portion of traffic going through mobile networks, mobile data offloading has become an important issue in cellular networks. A set of existing works concerning offloading are discussed in this section.

Various cloud offloading systems were proposed in the literature [8][10]. MAUI [8] and ThinkAir [9] profile hardware components and make offloading to optimize energy usage of mobile devices. They only consider the

energy consumption and ignore the other aspects of offloading. CloneCloud [10], was proposed in order to optimize applications partitioning between local execution and offloading with the purpose of minimizing execution time or energy consumption.

Besides mobile cloud frameworks, some other research works focus on offloading decision making issues. Authors in [11] have proposed an energy-aware data offloading scheme for C-RAN where the BBU make offloading decision considering the UEs' transmission rate and energy consumption of the cellular and WiFi networks. It uses the centralized characteristic of C-RAN to schedule Ues' offloading from RRH to Wi-Fi access point. The proposed scheme reduces the energy consumption and improves the throughput in the network. However, performance improvement are visible only if the data size is big enough.

Another offloading decision algorithm in the context of femtocloud paradigm is proposed in [12]. It consists in performing a series of classifications that consider several parameters such as latency, battery level, UE memory, etc. The proposed algorithm divided the application tasks into "urgent tasks" and "not urgent task" and it sends offloadable tasks that are assigned as urgent to the femtocloud regardless the channel conditions. This affects the energy consumption of the mobile handset and violates latency constraints if total offloading is up 3%.

In [13], authors proposed a context-aware decision algorithm (CADA) in order to offload to the cloud servers. The decision engine is composed of four components: context-aware decision algorithm, context profiler, energy model, and context database. CADA uses the location and time-of-day to make the mobile offloading decisions of individual methods. That's require a profiler. However, it is a source of overhead and it consumes memory in order to store users' profiles.

Different factors have been involved in offloading decision for mobile cloud computing. In [14], an offloading decision model that takes network unavailability into consideration is proposed. Based on network connection states and durations that are recorded in a history buffer, the application partition that is aimed to give benefit in network with low availability is calculated. Then the decision about offloading is validated. The proposed algorithm enhances network performances in terms of execution time and energy consumption but the complexity of the system increase with the increase of the number of mobile users.

In summary, mobile cloud offloading issue was studied for different networks and several decision algorithm were proposed. They exploit multiple objective optimization techniques. However, regarding the network instability, decision needs to be refreshed whenever system conditions change.

In our work, we tried to avoid different limits that were highlighted in previous introduced works by considering a multi-parameter decision algorithm for offloading decision in a Cloud RAN in order to adapt the offloading decision to the current state of the system. Besides, we will consider application requirements and mobile available resources and mobility spped in offloading decision.

III. APPROACH : THE PROPOSED OFFLOADING DECISION ALGORITHM IN C-RAN

With the increasing of mobile data traffic, offloading to the cloud has become an important and popular issue in cellular network. In our approach, we aim at optimizing the QoS in terms of latency, throughput, application response time, etc. in Cloud Radio Access Networks. In this paper we will deal with energy efficiency and offloading.

In our architecture, we consider a Cloud RAN heterogeneous architecture composed of H-RRHs (High RRHs) which acts as macro cells and L-RRHs (Low RRHs) which acts as small cells. In our scenario, we introduced the Cloud-RRH which represents the edge cloud. While in a traditional C-RAN architecture all the RAN functionalities are centralized in BBU pools, we propose to flexibly split these functionalities between edge and central cloud. We also introduce additional computation and storage resources in the Cloud-RRH for computation offloading. These resources are represented by cloud containers. The infrastructure is represented in Figure 1.

Based on our proposed 5G/HetNets C-RAN architecture, we developed an offloading decision algorithm. The main objective of this algorithm is to ameliorate the user experience by optimizing the latency and reducing the energy consumption of mobile devices in order to extend battery lifetime and also to enhance the offloading data traffic for better wireless transmission quality.

A. Offloading interactions

The mechanism of offloading is depicted in Figure 2. When a mobile terminal receives a task, it starts by generating an Offloading Request packet and sends the request to RRH. The offloading request packet is shown as bellow:

- *Offloading Req (service ID*, CAP_{MT}, CAP, E_{loc}, B), where CAP_{MT} is the Mobile Terminal (MT) capacity, CAP represents the capacity required by the received task, E_{loc} is the energy spent for local execution and B is the bandwidth between RRH and UE.

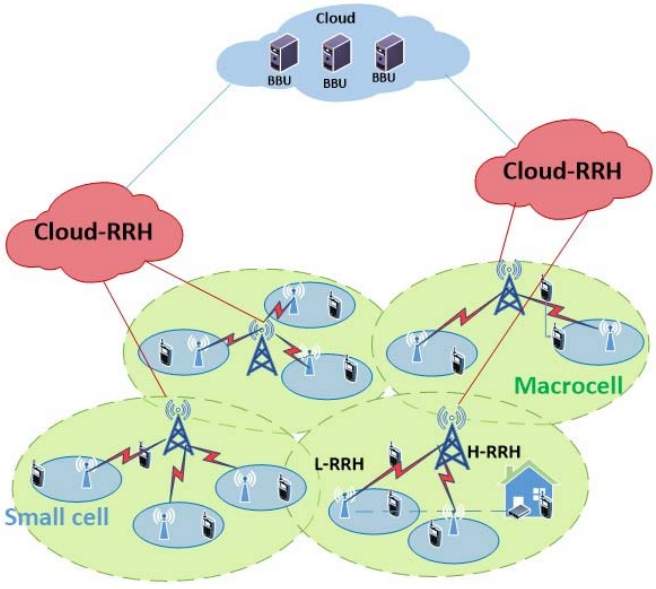

Fig. 1. Proposed C-RAN architecture

If the application will be offloaded to the Cloud-RRH, the cloudlet manager sends a *Resource Allocation Request* packet to the serving virtual machine where it indicates the capacity required. After processing, the Offloading Response will be routed up to the mobile terminal.

B. System model and QoS parameters

The basic idea of Cloud RAN is to separate the BBUs from cell sites and group them into BBU pools in real-time cloud for a centralized processing and management. Thanks to network flexibility, C-RAN can support multi- standard operations and can be easily deployed with heterogeneous networks.

In this paper, as represented in Figure 3, we consider a 5G system in a C-RAN context with K users served by either a High RRH (H-RRH) or a Low RRH (L-RRH). We consider uplink connection between mobile terminal and the serving RRH with a bandwidth B. According to [15], t_{up}, which represents the time employed to send S_{up} bits in the UL, only depends on the UP data rate r_{up} and the number of bits to be transmitted, i.e., $t_{up} = S_{up}/r_{up}$. Similarly, for the remote processing and the DL transmission from Cloud-RRH to the serving RRH, the time required is $t_{dl} = S_{dl}/r_{dl}$.

We adopted the following models for the power consumption at the mobile terminal in both UL and DL:

$$p_{up} = k_{(tx,1)} + k_{(tx,2)} * ptx \qquad (1)$$

$$p_{dl} = k_{(rx,1)} + k_{(rx,2)} * r_{dl} \qquad (2)$$

Where $k_{(tx,1)}$, $k_{(tx,2)}$, $k_{(rx,1)}$ and $k_{(rx,2)}$ are constants. These expressions are based on the measurements provided in [16] which indicates that the UL power consumed by the user increases with the radiation power and a baseline power is consumed just for having the transmission chain switched on. While in the DL, the power consumed by the user terminal increases with the downlink data rate and a baseline power is consumed just for having the reception chain switched on.

The maximum rate supported by the channel with K users and that depends on the transmission power and the quality of the channel is given by Shannon's theorem:

$$r_{(up,k)} = B\ log\ (1 + G_{up} * p_{(tx,k)}) \qquad (3)$$

$$r_{(dl,k)} = B\ log\ (1 + G_{dl} * p_{(tx,RRHce)}) \qquad (4)$$

Where G_{up} and G_{dl} represent the channel gain normalized by the noise power in UP and DL and $p_{(tx,k)}$ and $p_{(tx,RRHce)}$ are transmission power of the user and RRHce, respectively.

According to (1) and (2), the energy spent by MT (Mobile Terminal) in UP and DL is given by:

$$E_{up} = k_{(tx,1)} * t_{up} + k_{(tx,2)} * t_{up} * p_{tx} \qquad (5)$$

$$E_{dl} = k_{(rx,1)} * t_{dl} + k_{(rx,2)} * t_{dl} * r_{dl} \qquad (6)$$

Using equation (3), $p_{tx} = \frac{2^{\frac{r_{up}}{B}} - 1}{G_{up}}$. So the energy consumed by the MT for offloading is given by:

$$E_{off} = E_{up} + E_{dl}$$

$$= k_{(tx,1)} * t_{up} + k_{(tx,2)} * t_{up} * \frac{2^{\frac{S_{up}}{t_{up}\ B}} - 1}{G_{up}} + k_{(rx,1)} * t_{dl} + k_{(rx,2)} * S_{dl} \qquad (7)$$

We considered that the energy spent by MT in local processing is proportional to the number of processed bits. The expression is given by the following equation:

$$E_{loc} = \varepsilon_0 * S \qquad (8)$$

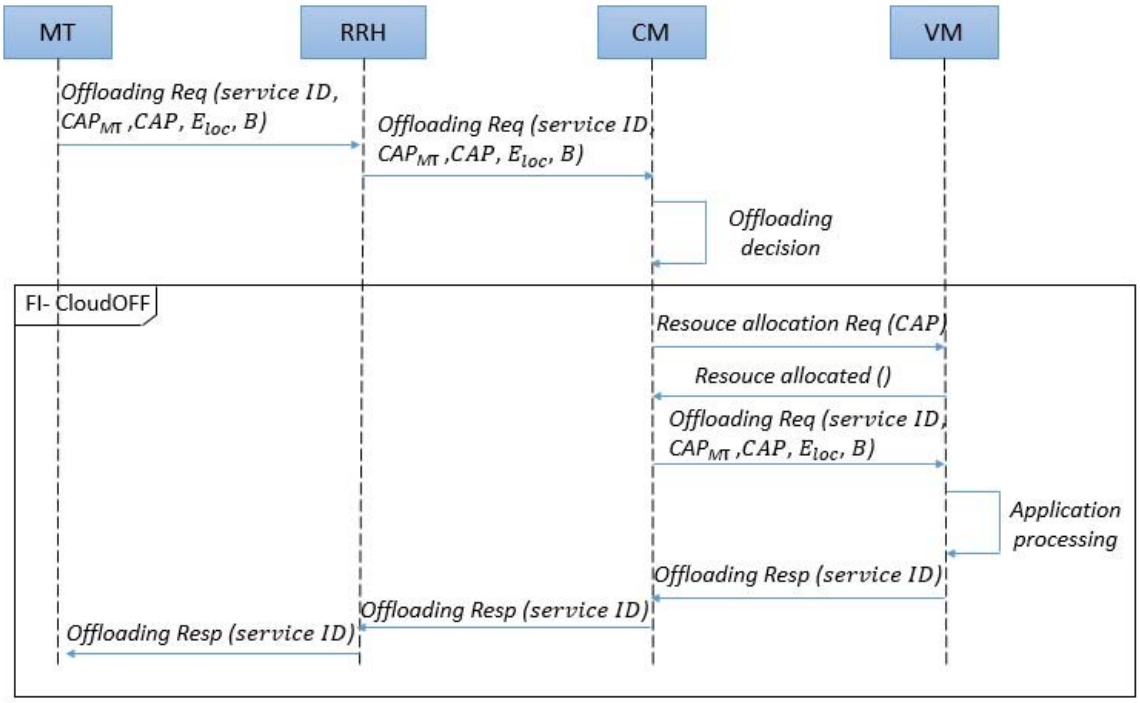

Fig. 2.Cloud RAN: Offloading mechanism

Where ε_0 is a constant that accounts jointly for the Joules/cycle and cycles/bit for the processor at the MT and S is the number of bits.

Concerning the latency, we considered t_0 and t_1 as the time needed to process one bit at the MT and at the RRHce respectively. The time required for offloading process to be completed is given by the sum of the time required to send bits from the mobile terminal to the serving RRHce through the UL, the time for the remote processor to execute the offloaded computation, and the time to send all the output bits through the DL. Latency expression for local processing and offloading are given by:

$$L_{loc} = t_0 * S \quad (9)$$

$$L_{off} = t_{up} + t_1 * S + t_{dl} \quad (10)$$

C. Algorithm description

In this section, we will discuss the workflow of the proposed algorithm. The main objective of this algorithm is to decide if the application have to be processed locally or need to be offloaded to the Cloud-RRH in order enhance the QoE (Quality of Experience) while optimizing the network and MT resource utilization.

We propose to introduce a multitude of parameters in the offloading decision process without including them in a complex optimization problem. The algorithm of mobile application offloading decision is run, at each time slot, on the set of tasks generated by the launched applications. Figure 3 shows the general workflow of the proposed algorithm.

When a task is received and as a first step, we will consider the UE velocity and compare it to a velocity threshold: if the UE velocity is high the task will be executed locally in order to minimize network overhead which can lead to user experience deterioration. Otherwise, we will compare the latency using equations (9) and (10) because of the importance of latency regarding QoS. If the latency generated when the task is offloaded to Cloud-RRH is greater than the latency generated by local execution, the task should be computed locally at the mobile handset. However, mobile terminals resources are limited in terms of computational capacity. So, if the computational capacity required is greater than the predefined percentage of the total locally available capacity, the task have to be offloaded to the Cloud-RRH.

Then, if the latency generated by offloading is lower, we have to compare the consumed energy at the mobile handset in case of offloading, E_{off},using equation (7) and E_{loc} , the consumed energy in case of local computation of this task using equation (8). If E_{off}>E_{loc}, then it compares the latency generated by local computation and the maximum latency authorized by the application. If L_{loc}<L_{max} the task is computed locally, otherwise, the task is offloaded.

Finally, if E_{off}< E_{loc} it examines the channel conditions. The channel capacity can be calculated using Shannon theorem. Then, the channel coefficient is compared to the average channel coefficient calculated and updated over time. If the current channel realization is above this average, it is considered that the channel is in a relatively "good" state and the task is offloaded. Otherwise, it is calculated locally in order to prevent the system from applying costly offloading when channel conditions are not favorable.

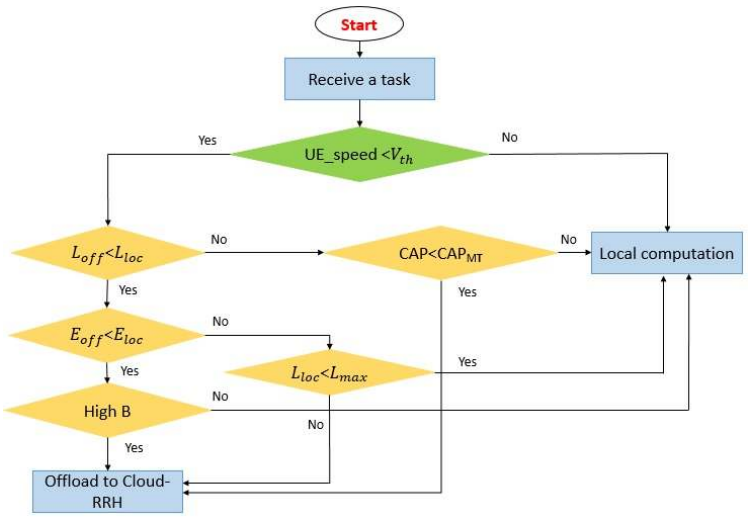

Fig. 3. Proposed offloading decision algorithm for C-RANs

IV. SIMULATION AND RESULTS

This section provides simulation results for the proposed method, when the goal is the minimization of the energy consumed by the MT and the application response time while considering a multitude of parameters concerning the network state and the handset characteristics.

In order to verify the performance of our approach, we consider an urban environment simulation scenario. We consider a heterogeneous C-RAN with 7 H-RRH and 4 L-RRH per cell. H-RRHs have a coverage of 500m and L-RRHs are 30m-radius [15]. System simulation parameters are listed in Table I.

TABLE I. SIMULATION PARAMETERS

Parameter	Value
K(tx,1)	0.4W
K(tx,2)	18
K(rx,1)	0.4W
K(rx,2)	2.86 10^{-3} W/Mbps
ε_0	8.6 10^{-8} J/bit
t_0	10^{-7} s/bit
t_1	t_0/2
Bandwidth (*B*)	10 MHz
Maximum latency authorized by the application (L_{max})	4s
Users mobility speed	3 Km/h < V ≤ 120 Km/h
V_{th}	5m/s

The values for ε_0 and t_0 have been aligned with the measurements given in [18] for energy and frequency characteristics of local computing in commercial handsets, as well as computation to data ratios in practical applications. The simulation aim to evaluate our proposed offloading decision algorithm compared to total offloading.

Figure 4 illustrates the variation of the response time over the data size (ranging from 1 to 100 Kbits). We can observe that the proposed offloading scheme can ameliorate the user experience by reducing the response time. When the data size is small the results of the two schemes are close. However, when the data size becomes larger, the difference between the two schemes becomes lager. On average there is an improvement of 14% in response time. Thus, the proposed offloading decision algorithm can be useful for high resource-demand applications.

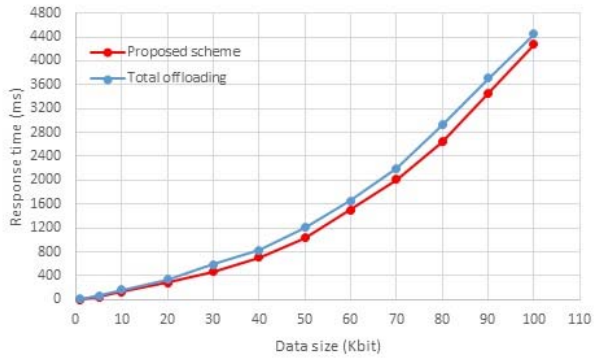

Fig. 4. Application response time

Figure 5 shows the simulation results of the energy spent by the MT under the data size (ranging from 1 to 100 Kbits). We can see that the proposed offloading decision algorithm can make the mobile handset consumes less energy. The difference is more important when the data size is big. Therefore, the proposed algorithm is able to augment the mobile handset battery lifetime while executing complex program applications compared to local execution and total offloading.

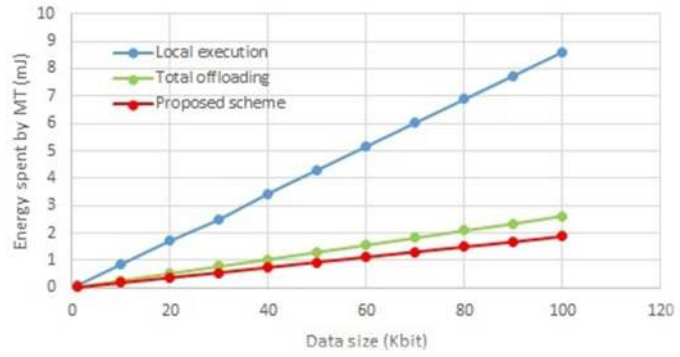

Fig. 5. Total energy consumed by MT

V. CONCLUSION AND PERSPECTIVES

Mobile data offloading has become an important issue for mobile cellular network in recent years. In this paper we propose an offloading decision algorithm which consider a multitude of parameters in order to make the offloading more efficient in heterogeneous C-RANs with Cloud-RRH. The proposed scheme helps to make the best usage of network and mobile terminal resources. We show the efficiency of our algorithm through numerical simulations. Results show that this algorithm can response time and the mobile terminal energy consumption and thus improve the user experience.

In future work, we will try to evaluate the cost depending upon energy consumption optimization. Furthermore, we will try to better investigate and evaluate the network performances by handling the interference and mobility management in C-RAN.

REFERENCES

[1] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," IBM Journal of Research and Development, january-february 2010.

[2] "C-RAN The Road Towards Green RAN," China Mobile Research Institute, Tech. Rep., October 2011.

[3] Kumar, Karthik; Liu, Jibang; Lu, Yung-Hsiang; Bhargava, Bharat, "A Survey of Computation Offloading for Mobile Systems", Mobile Networks and Applications, vol. 18, no. 1, 2013.

[4] Rim H, Kim S, Kim Y, and Han H, " Transparent method offloading for slim execution", International symposiumon wireless pervasive computing, pp 1–6, 2006.

[5] Yang K, Ou S, and Chen H-H, "On effective offloading services for resource-constrained mobile devices running heavier mobile internet applications, IEEE communications magazine 46(1), pp 56–63, 2008.

[6] Xian C, Lu Y-H, and Li Z, "Adaptive computation offloading for energy conservation on battery-powered systems", International conference on parallel and distributed systems, pp 1–8, 2007.

[7] Chun BG and Maniatis P , "Augmented smartphone applications through clone cloud execution. In: Conference on hot topics in operating systems", USENIX Association, pp 8–12, 2009.

[8] E. Cuervo, A. Balasubramanian, D. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "Maui: making smartphones last longer with code offload", in Proceedings of the 8th international conference on Mobile systems, applications, and services (MobiSys '10). ACM, New York, NY, USA, 49-62.

[9] S. Kosta, A. Aucinas, Pan Hui, R. Mortier, and Xinwen Zhang, "ThinkAir: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading," INFOCOM, 2012 Proceedings IEEE , pp.945,953, 25-30, March 2012.

[10] B. G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "Clonecloud: elastic execution between mobile device and cloud," in Proceedings of the sixth conference on Computer systems (EuroSys '11). ACM, New York, NY, USA, 301-314.

[11] C. Yuh-Shyan, H. Chih-Shun, J. Tong-Ying and L. Hsin-Han, "An Energy-Aware Data Offloading Scheme in Cloud Radio Access Networks", IEEE Wireless Communications and Networking Conference (WCNC), 2015.

[12] J. Oueis, E. Calvanese Strinati, and S. Barbarossa, "Multi-parameter Decision Algorithm for Mobile Computation Offloading", IEEE Wireless Communications and Networking Conference (WCNC), 2014.

[13] L. Ting-Yi, L. Ting-An, H. Cheng-Hsin, and K. Chung-Ta , "Context-Aware Decision Engine for Mobile Cloud Offloading", IEEE WCNC Workshop on Mobile Cloud Computing and Networking, 2013.

[14] W. Huijun, H. Dijiang, and S. Bouzefrane, "Making Offloading Decisions Resistant to Network Unavailability for Mobile Cloud Collaboration", 9th IEEE International Conference on Collaborative Computing: Applications and Worksharing, 2013.

[15] O. Muñoz, A. Pascual-Iserte, and J. Vidal, "Joint Allocation of Radio and Computational Resources in Wireless Application Offloading", Future Network & MobileSummit Conference Proceedings, 2013.

[16] A.R. Jensen, M. Lauridsen, P. Mogensen, T.B. Sørensen, and P. Jensen, "LTE UE Power Consumption Model: For System Level Energy and Performance Optimization," Proceedings IEEE VTC2012-Fall Vehicular Technology Conference Fall, Québec city, Canada, September 2012.

[17] A.P. Miettinen and J.K. Nurminen, "Energy Efficiency of Mobile Clients in Cloud Computing," Proc. 2nd USENIX Conference on Hot Topics in Clod Computing 2010 (HotClout'10), Boston (USA), June 2010.

[18] "TROPIC: Distributed computing, storage and radio resource allocation over cooperative femtocells", Seventh Framework Program for Research of the European Commission, 2013.