Small Cell Clustering for Efficient Distributed Fog Computing: A Multi-user Case

Jessica Oueis[1], Emilio Calvanese Strinati[1], Stefania Sardellitti[2], and Sergio Barbarossa[2]

[1]CEA, LETI, Grenoble, France

[2] Dept. of Information Engineering, Electronics and Telecommunications, Sapienza Univ. of Rome, Rome, Italy

***Abstract*—Ultra-dense deployment of radio access points is a key enabler for future 5G networks. It allows the network to cope with the ever increasing mobile data traffic. In addition, these radio access points can serve as an infrastructure for a local mobile cloud computing platform referred to as fog computing. The fog is a capillary edge cloud that enables joint optimization of communication and computational resources for maintaining an efficient and scalable network design. In this paper, we address the problem of radio access points clustering for fog computing applications. We focus on the case where multiple users require fog computing services. We formulate the distributed clustering problem as a joint optimization of the computation and communication resources. We transform the non-convex original problem into an equivalent convex one. Our simulation results show that the clustering solution derived from this problem yields high users' satisfaction ratio while keeping low the communication power consumption of the computation cluster.**

I. Introduction

With the daily emergence of new complex mobile and IoT devices and applications, mobile data traffic has been spectacularly growing. From 2013 to 2014, there has been a 60% growth in mobile data traffic, and an 8x growth of this traffic is expected by 2020 [1]. To cope with this traffic evolution, ultra-dense deployment and fog computing are two key enablers. Indeed, offloading computation and data processing from mobile devices to cloud computing clusters is a powerful tool for offering a good Quality of Experience (QoE) to mobile users. Fog computing cluster, which operates by distributing computation load among neighbor radio access points endowed with computational capacities, is a cloud computing platform of proximity to mobile devices. Mobile computation offloading allows devices to run sophisticated applications while keeping a prolonged battery lifetime of their mobile devices. Different offloading strategies have been proposed [2]-[7] . Moving cloud capabilities closer to mobile devices has been the core idea of the European project TROPIC [8], in which a new architecture is proposed. It consists of a platform where small cell access points are endowed with computational and storage capacities, and thus, form a local cloud. Due to the proximity of the cloud to the users, communication power consumption and experienced latency can be greatly reduced. When mobile devices computation are to be offloaded, resources allocation is required. A joint allocation of transmission power of the mobile device and the computational capacity at the serving small cell is a must. In [9], an optimization method of resources allocation in the mobile cloud context is presented for the case of a single mobile user served by a single base station. In [10], the multi-user case has been studied. An optimization of the computational capacities and power consumptions allocated for each device is derived under latency constraints and power budget limits. The case of multi-users, multi-clouds MIMO networks has been studied in [11] where the optimal assignment of each mobile user to a base station and to a cloud is derived. To this end, the offloading problem is formulated as a joint optimization of the radio resources allocated for each user and the computational resources allocated at each small cell cloud. The small cell clouds are pre-established clusters with known computational capacities.

Small cells, even if considered endowed with computational capacities, cannot offer the same performance as the traditional cloud powerful servers. A small cell may not have a sufficient available computational capacity to satisfy a user's request, especially in the case where it is serving multiple users simultaneously. In this paper, a serving small cell with a computation offloading request, forms its own computation cluster for each request. This cluster incorporates small cells that contribute in the computation of the requested task. An important role in the cluster establishment is also given to the backhaul topology and technology that small cells use to communicate. A study on the impact of backhaul on the small cell clustering is presented in [12]. In this work, small cells communicate through point to point wireless channels. Therefore, the size of the computation cluster and which are its participants depend on both computational capacities of each small cell and on the link quality between small cells. In the case of a single user device asking for computation offloading, the authors derived several strategies that could be followed for establishing a small cells computing cluster [13]. The proposed strategies had different objectives. The first objective was to form a cluster that computes the task with lowest latency. Then, a sparsification method of this cluster was proposed in order to reduce its power consumption. Two additional strategies had the objective of forming the cluster with the lowest power consumption from the system and small cells point of views. In this paper, we consider the more complex multi-user case. Each of the devices is connected to a serving small cell. More than one device can share a serving small cell. Every computation request should be computed by a computation cluster. A first novel aspect

978-1-4799-8091-8/15/$31.00 ©2015 IEEE

of what we propose, is the cluster scalability according to the computation requests requirements. In fact, small cell cloud has always been considered in previous works as an established set with known characteristics. What we propose allows the cluster to have adaptive size, load distribution, and inside communication and computation resources allocation. Computation clusters should be built for all users having their computation requests satisfied without violating the imposed latency constraints. Hence, the second novel approach of what we propose is to jointly compute clusters for all active users requests simultaneously in order to better distribute computation and communication resources for a higher users' QoE. We formulate the clustering problem for multiple users as an optimization problem, to distribute the computation load of all requests among the active small cells in the network. Hence, we jointly allocate transmission powers for each small cell, and the computational capacity for each user at each small cell. The objective of the optimization problem is to minimize the clusters power consumption while respecting the imposed latency constraints of each user request.

II. Simulations System Model

We consider a multi-user scenario where a set $\mathcal{N}$ of N small cells are deployed. A set $\mathcal{K}$of K mobile users are served each by a small cell denoted by S_k. The set of serving small cells is denoted by $\mathcal{S}$. The set of devices associated to the small cell s is denoted by $\mathcal{K}_s$. Every device k in $\mathcal{K}$ sends a computational requests (W_k, Δ_k) to its serving small cell S_k. W_k and Δ_k denote for the number of instructions to be computed and the maximum latency imposed by the application, respectively. Computational requests are characterized by the number of input and output bits, which are the bits to be sent to the computing small cell and back to the user. Each small cell n in $\mathcal{N}$ is characterized by a computational capacity of F_n instructions per second. Each small cell can serve multiple devices simultaneously by according a part of its computational capacity, say f_{kn}, to each user k. We consider that the computation requests are already sent to the serving small cell. The serving small cell forms a computation cluster for each of the requests it received. The computation load of each request W_k is distributed among the small cells of the computation cluster. Each small cell n is accorded W_{kn} of user's k request. The serving small cell sends the necessary input bits to the cluster small cells. The number of input bits is equal to $\theta_{UL}W_{kn}$. The cluster small cells processes the bits and sends back the output bit to the serving small cell S_k. The number of output bits to be sent is equal to $\theta_{DL}W_{kn}$. The transmission power used to send input and output bits between the serving small cell s and the computing small cell n is p_{sn}. The information rate that can be achieved through the wireless channel link between small cells, taking into account packet retransmission is:

$$R_{sn} = B_{sn} log(1 + \frac{\sigma_c |h_{sn}|^2 p_{sn}}{(1-PER)\Gamma d^{\beta} N_0}) \tag{1}$$

where σ_c is the shadow fading coefficient of the adopted Rayleigh channel model. The channel fading is assumed constant for a whole transmission period. We assume perfect estimation of the coefficients h_{sn} of the channel between small cells $s \in \mathcal{S}$ and $n \in \mathcal{N}$. PER is the target packet error rate, Γ indicates the SNR margin to guarantee a minimum bit error rate BER ($\Gamma(BER) = -\frac{2log(5BER)}{3}$), d represents the distance between s and n, β indicates the path loss exponent which depends, in an indoor environment, on the number of walls separating the two communicating SCs [14], and N_0 is the noise power.

III. Joint optimization of computation offloading load distribution and in-cluster transmit powers

Denoting by $\mathbf{p} \triangleq (p^k_{sn})_{\forall n,s,\forall k \in \mathcal{K}}$, $\mathbf{f} \triangleq (f_{kn})_{\forall k,n}$, $\mathbf{w} \triangleq (w_{kn})_{\forall k,n}$ respectively, the transmit powers, computational rates and computational loads associated to each mobile user, the optimization problem can be formulated as follows:

$$\min_{\mathbf{p},\mathbf{f},\mathbf{w}} \quad \sum_{s\in\mathcal{S}} \sum_{k\in\mathcal{K}_s} \sum_{n\in N, n\neq s} p^k_{sn}$$

s.t.

$$\begin{aligned}
&(a) \quad w_{kn} \geq 0, f_{kn} \geq 0, && \forall k \in \mathcal{K}, n \in \mathcal{N} \\
&(b) \quad \sum_{n\in\mathcal{N}} w_{kn} = W_k, && \forall k \in \mathcal{K} \\
&(c) \quad 0 \leq p^k_{sn} \leq P_{max}, && \forall s \in \mathcal{S}, k \in \mathcal{K}_s, n \in \mathcal{N}, n \neq s \\
&(d) \quad \sum_{k\in\mathcal{K}} f_{kn} \leq F_n, && \forall n \in \mathcal{N} \\
&(e) \quad \Delta^k_{sn}(p^k_{sn}, f_{kn}, w_{kn}) \leq \Delta_k, && \forall n, s \in \mathcal{S}, k \in \mathcal{K}_s.
\end{aligned} \tag{$\mathcal{P}$}$$

where we define the delay function

$$\Delta^k_{sn}(p^k_{sn}, w_{kn}) \triangleq \begin{cases} \frac{w_{kn}}{f_{kn}} + \frac{w_{kn}\theta}{R^k_{sn}(p^k_{sn})} & \text{if } p^k_{sn}.f_{kn}.w_{kn} > 0, \forall n \neq s \\ 0 & \text{if } p^k_{sn}.f_{kn}.w_{kn} = 0, \forall n \neq s \\ \frac{w_{kn}}{f_{kn}} & \text{if } f_{kn} > 0, n = s \\ 0 & \text{if } w_{kn}.f_{kn} = 0, n = s \end{cases} \tag{2}$$

with $\theta = \theta_{UL} + \theta_{DL}$. Note that the delay defined as in (2) means that: i) for all the non serving small cells, the delay function is forced to be strictly positive only if the transmit power, the computational rate and load assigned to the mobile user are non-zero; ii) in case of computation at the serving small cell, i.e. for $n = s$, the transmit power p^k_{sn} is null then the delay constraint is reduced to a computation time constraint.

Unfortunately problem $\mathcal{P}$ is non-convex, due to the non-convexity of the delay constraints (e). Nevertheless, in the following we cast $\mathcal{P}$ into a convex equivalent problem. To this end, observe that the delay constraint can be equivalently rewritten for $p^k_{sn}.f_{kn}.w_{kn} > 0, n \neq s$ and under the feasibility condition $\Delta_k f_{kn} > w_{kn}$, as

$$g^k_{sn}(p^k_{sn}, f_{kn}, w_{kn}) \triangleq -B_{sn}\log_2(1 + a^k_{sn}p^k_{sn}) + \frac{w_{kn}f_{kn}\theta}{\Delta_k f_{kn} - w_{kn}} \leq 0 \tag{3}$$

where $a^k_{sn} \triangleq \frac{\sigma_c |h^k_{sn}|^2}{(1-PER)\Gamma d^{\beta} N_0}$. Note that the delay constraint in (3) is convex as can be easily verified by proving that the Hessian of g^k_{sn} is a semi-definite positive matrix. Additionally, the non-convex delay constraint in (e) can be reduced for $f_{kn} > 0, n = s$,

to the linear convex constraint $w_{ks} \leq \Delta_k f_{ks}$. Hence, problem $\mathcal{P}$ can be reformulated as:

$$\min_{\mathbf{p},\mathbf{f},\mathbf{w}} \quad \sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}_s}\sum_{n\in N, n\neq s} p_{sn}^k$$

s.t.

$$\begin{aligned}
&(a) \quad w_{kn} \geq 0, f_{kn} \geq 0, && \forall k \in \mathcal{K}, n \in \mathcal{N}\\
&(b) \quad \sum_{n\in\mathcal{N}} w_{kn} = W_k, && \forall k \in \mathcal{K}\\
&(c) \quad 0 \leq p_{sn}^k \leq P_{max}, && \forall s \in \mathcal{S}, k \in \mathcal{K}_s, n \in \mathcal{N}, n \neq s\\
&(d) \quad \sum_{k\in\mathcal{K}} f_{kn} \leq F_n, && \forall n \in \mathcal{N}\\
&(e) \quad g_{sn}^k(p_{sn}^k, f_{kn}, w_{kn}) \leq 0, && \forall s \in \mathcal{S}, k \in \mathcal{K}_s, n \in \mathcal{N}, n \neq s\\
&(f) \quad w_{kn} - \Delta_k f_{kn} < 0, && \forall s \in \mathcal{S}, k \in \mathcal{K}_s, n \in \mathcal{N}, n \neq s\\
&(g) \quad w_{ks} - \Delta_k f_{ks} < 0, && \forall s \in \mathcal{S}, k \in \mathcal{K}_s.
\end{aligned} \tag{$\mathcal{P}_c$}$$

Problem $\mathcal{P}_c$ enjoys some desirable properties as stated in the following theorem.

Theorem 1. *Given problem $\mathcal{P}$ and $\mathcal{P}_c$, the following hold:*
(i) Necessary conditions for $\mathcal{P}$ to be feasible are:

$$\frac{W_k}{\sum\limits_{n\in\mathcal{N}} F_n} \leq \Delta_k,\ W_k \leq \sum_{n=1}^{N} \frac{\Delta_k}{\frac{1}{F_n} + \frac{\theta}{R_{sn}^k(P_{max})}}; \tag{4}$$

(ii) $\mathcal{P}_c$ is a convex problem then any local optimal solution is a global optimal minimum;
(iii) $\mathcal{P}$ and $\mathcal{P}_c$ are equivalent.

Proof: To prove point (i) of Th. 1 observe that from the constraint (b) it exists $\forall k$ at least a server n for which $w_{kn} > 0$. Then from (e) in $\mathcal{P}$, we can write $w_{kn} < \Delta_k f_{kn}$ which leads to the first condition in (4). This implies that the maximum delay imposed by the application Δ_k cannot be less than the minimum execution time which can be achieved by a single equivalent server with computational capacity equal to that of the overall network, i.e. $\sum\limits_{n\in\mathcal{N}} F_n$. The second condition in (4) is a global condition which can be easily derived from (e) in $\mathcal{P}$. To prove point (ii) in Th. 1 it is sufficient to observe that problem $\mathcal{P}_c$ is convex since the objective function and all the constraints are convex. Then any stationary point is a *global* optimal solution of the problem. It remains to prove point (iii). Since for $\mathcal{P}_c$ the Slater's constraint qualification holds true, any optimal solution satisfies the KKT conditions of $\mathcal{P}_c$. The Lagrangian function associated to $\mathcal{P}_c$ is:

$$\begin{aligned}
\mathcal{L}(\mathbf{p},\mathbf{f},\mathbf{w}) \triangleq & \sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}_s}\sum_{n\in\mathcal{N}, n\neq s} p_{sn}^k + \sum_{k=1}^{K}\lambda_k\Big(\sum_{n=1}^{N} w_{kn} - W_k\Big)\\
& - \sum_{n=1}^{N}\sum_{k=1}^{K}\beta_{sn}^k w_{kn} + \sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}_s}\sum_{n\in\mathcal{N}, n\neq s}\Big[\alpha_{sn}^k(p_{sn}^k - P_{max})\\
& - \mu_{sn}^k p_{sn}^k + \tau_{sn}^k\Big(-R_{sn}^k(p_{sn}^k) + \frac{w_{kn} f_{kn}}{\Delta_k f_{kn} - w_{kn}}\Big)\\
& + \eta_{sn}^k(w_{kn} - \Delta_k f_{kn})\Big] + \sum_{n\in\mathcal{N}}\gamma_n\Big(\sum_{k=1}^{K} f_{kn} - F_n\Big)\\
& - \sum_{n\in\mathcal{N}}\sum_{k=1}^{K}\rho_{kn} f_{kn} + \sum_{s\in\mathcal{S}}\sum_{k\in\mathcal{K}_s}\kappa_{ks}(w_{ks} - \Delta_k f_{ks})
\end{aligned}$$

where the non negative variables $\alpha_{sn}^k, \mu_{sn}^k, \tau_{sn}^k, \eta_{sn}^k, \beta_{kn}, \rho_{kn}, \gamma_n, \kappa_{ks}$ and $\lambda_k \in \mathbb{R}$ are the Lagrangian multipliers. The KKT conditions are:

(a'): $\frac{d\mathcal{L}}{dp_{sn}^k} = 1 + \alpha_{sn}^k - \mu_{sn}^k - \tau_{sn}^k \frac{B_{sn} a_{sn}^k}{1 + a_{sn}^k p_{sn}^k} = 0,$

$\forall s \in \mathcal{S}, k \in \mathcal{K}_s, n \in \mathcal{N}, n \neq s,$

(b'): $\frac{d\mathcal{L}}{df_{kn}} = \gamma_n - \tau_{sn}^k \frac{w_{kn}^2 \theta}{(\Delta_k f_{kn} - wkn)^2} - \rho_{kn} - \eta_{sn}^k \Delta_k = 0,$

$\forall s \in \mathcal{S}, k \in \mathcal{K}_s, n \in \mathcal{N}, n \neq s,$

(c'): $\frac{d\mathcal{L}}{dw_{kn}} = \lambda_k - \beta_{kn} + \tau_{sn}^k \frac{f_{kn}^2 \Delta_k \theta}{(\Delta_k f_{kn} - w_{kn})^2} + \eta_{sn}^k = 0,$

$\forall s \in \mathcal{S}, k \in \mathcal{K}_s, n \in \mathcal{N}, n \neq s,$

(d'): $\frac{d\mathcal{L}}{df_{ks}} = \gamma_s - \kappa_{ks}\Delta_k - \rho_{ks} = 0, \forall s \in \mathcal{S}, k \in \mathcal{K}_s,$

(e'): $\frac{d\mathcal{L}}{dw_{ks}} = \lambda_k - \beta_{ks} + \kappa_{ks} = 0, \forall s \in \mathcal{S}, k \in \mathcal{K}_s,$

(f'): $\lambda_k \in \mathbb{R}, \forall k, \quad \sum_n w_{kn} - W_k = 0, \forall n,$

(g'): $0 \leq \beta_{kn} \perp w_{kn}, \forall k, n,$

(h'):$0 \leq \alpha_{sn}^k \perp (P_{max} - p_{sn}^k) \geq 0, \forall k, n \neq s,$

(i'):$0 \leq \mu_{sn}^k \perp p_{sn}^k \geq 0, \forall k, n \neq s$

(l'):$0 \leq \gamma_n \perp (F_n - \sum_{k=1}^{K} f_{kn}) \geq 0, \forall n,$

(m'):$0 \leq \rho_{kn} \perp f_{kn} \geq 0, \forall k, n,$

(n'):$0 \leq \tau_{sn}^k \perp (R_{sn}^k(p_{sn}^k) - \frac{w_{kn} f_{kn} \theta}{\Delta_k f_{kn} - \omega_{kn}}) \geq 0, \forall k, n \neq s,$

(o'):$0 \leq \eta_{sn}^k \perp (\Delta_k f_{kn} - w_{kn}) > 0, \forall k, n \neq s,$

(p'):$0 \leq \kappa_{ks} \perp (\Delta_k f_{ks} - w_{ks}) \geq 0, \forall k, s.$

where $a \perp b$ stands for $\langle a, b \rangle = 0$. Observe that from the complementary condition (o') we get $\eta_{sn}^k = 0, \forall k, n \neq s$. Let us first consider $n \neq s$ by studying separately the two cases i) $\tau_{sn}^k > 0$ and ii) $\tau_{sn}^k = 0$. Under assumption i), it follows from the complementarity condition (n') that the delay constraint is always active. Hence if $p_{sn}^k > 0$ then $w_{kn} f_{kn} > 0$ and conditions (g') and (m') lead to $\beta_{kn} = \rho_{kn} = 0$. Then from (b') we get $\gamma_n > 0, \forall k, n$ so that the computational rate constraint holds with equality and from (c') it results $\lambda_k < 0, \forall k, n$. Finally, assume $\tau_{sn}^k > 0$ and $p_{sn}^k = 0$. Then $w_{kn} > 0, f_{kn} > 0$ is not an admissible solution since it contradicts the constraint qualification (n') being the delay constraint always active. On the other hand, $w_{kn} > 0, f_{kn} = 0$ lead to an absurdum since from (c') one gets $\lambda_k = 0$ while it must always be $\lambda_k < 0$. Under assumption i), it remains to check if the solution $p_{sn}^k = 0, w_{kn} = 0$ is achievable. In this point conditions (b') and (c') reduce respectively to $\gamma_n = \rho_{kn}$ and $\lambda_k - \beta_{kn} + \tau_{sn}^k = 0$. The condition $\gamma_n = \rho_{kn}$ implies $f_{kn} = 0$ since γ_n is always positive. It is important to remark that albeit the feasible set of $\mathcal{P}_c$ does not include the all zeros solution, the admissible solution $p_{sn}^k = w_{kn} = 0$ leads to $f_{kn} = 0$.

This permits to reach along gradient directions for which the KKT conditions are not violated, the null delay point enclosed through (2) in the feasible set of $\mathcal{P}$.

Let us consider now the case ii), i.e. $n \neq s, \tau_{sn}^{k} = 0$. From (c') one gets $\lambda_k = \beta_{kn}$ and this contradicts the fact that $\lambda_k < 0$.

The only case left to study is $n = s$. Under this assumption, observe that $\kappa_{ks} = 0$ is not admitted since from (e') λ_k could not be strictly negative. Then let us consider $\kappa_{ks} > 0$. From the complementary condition (p') $w_{ks} = \Delta_k f_{ks}$ and w_{ks}, f_{ks} can assume non-negative values while meeting the KKT conditions. This implies that according to (2), the delay can assume the zero value for $w_{ks} = f_{ks} = 0$.

It is important to remark that $\mathcal{P}$ is a hard to be handled problem due to the discontinuous non-convex delay constraints. Nevertheless as stated in Th. 1, we can find out its optimal solution by solving the equivalent convex problem $\mathcal{P}_c$. ■

IV. Numerical Evaluation

In this section, we evaluate the effectiveness of the proposed small cell clustering solution in computation clusters. We consider the case of femtocell deployment for urban scenarios model of the 3GPP framework [14]. This scenario is represented by a single floor building of a 25 apartments grid. In each of these apartments a femtocell is deployed. Parameter ρ_a determines the ratio of active femtocells in the grid. Parameter ρ_s determines the ratio of active femtocells that are connected to mobile devices (serving femtocells). We adopt the same system model described in Section II with parameters values listed in Table I.

TABLE I: Simulation parameters values

Parameter	value	Parameter	value
ρ_a	0.5	ρ_s	0.32
B	20MHz	σ_c	10
N_0	-118.4 [dB/Hz]	BER	10^{-6}
W	$[2.10^6; 10.10^6]$	Δ_{app}	[0.5;3.5]
F_n	$[10.10^6; 15.10^6]$	P_{max}	1 [W]
θ_{DL}	0.2	θ_{UL}	1

We compare the results of the joint clusters optimization to the case where all requests are handled by the serving small cell ('No Clustering'), the case where a static clustering rule of equal load distribution between active neighbor small cells is imposed ('Static Clustering'), and the case where clusters are formed for each user successively ('Successive Clusters Optimization').

Figure 1 shows the percentage of satisfied users. In order to evaluate this percentage, we try to solve the optimization problem with the total number of active users in the network. In case of failure of reaching a solution, users that request higher computation load are eliminated one by one until all considered users are satisfied. The satisfaction ratio is evaluated for an increasing number of possible active users per small cell. In Figure 1 we show how the joint clustering strategy for all users greatly outperforms all other strategies. The fact of taking into account all the active devices in the system allows better distribution and allocation of both computation and communication resources, and thus, higher QoE.

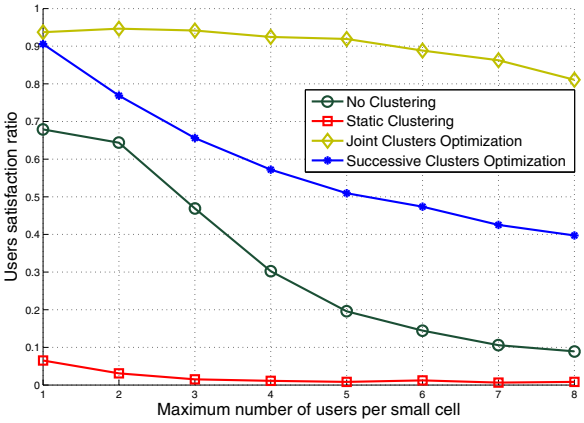

Fig. 1: Users satisfaction ratio in dependence on number of users per small cell

Results in Figures 2 and 3 should be read jointly with Figure 1. Figure 2 shows the average power consumption in the computation clusters. In the *no clustering* case, no data transmissions take place so there's no computation offloading, no communication power consumption but extremely low QoE. In the case of *static clustering*, the power consumption in the computation clusters is higher than in the proposed joint optimization. The distribution load does not take into account the available computational capacities at the cluster small cells. Therefore, the serving small cell may end up increasing its power consumption for a faster transmission of input data in order to assure the tasks computation without any latency violation. This leads to higher power consumption for lower users' satisfaction ratio. This figure also shows that the *joint clusters optimization* consumes more transmission power than *successive clustering* with the goal of increasing the satisfaction ratio through more adapted resources allocation.

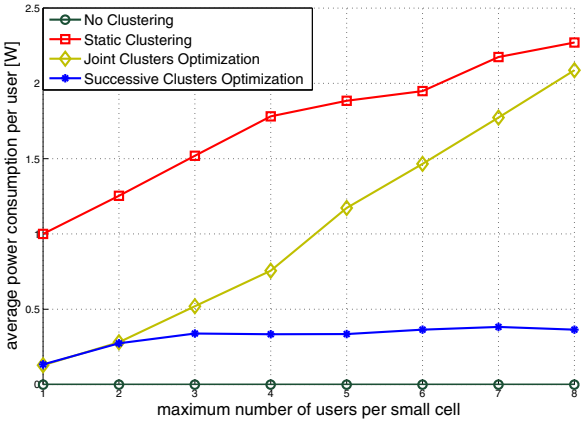

Fig. 2: Average power consumption per user in dependence on the number of users per small cell

Even though the objective function of the joint clustering optimization problem targets power consumption minimiza-

tion, we notice that it achieves some gain in the cluster latency. This is mostly due to the cases where local computation resources at the serving small cell are enough for computing its users requests. In this case, local computational capacity is accorded to the users power consumption free, and a latency gain can be achieved. It is clear that *joint clusters optimization*, compared to *successive clustering*, exploits the latency gain and trades it with higher satisfaction ratio.

Different trade-offs with latency can be exploited, where we degrade performance in latency to an acceptable level to achieve higher gains in other matters. For example, we could reduce the cluster size by sparsifying it and exclude some small cells in order to put them in an idle state for a lower equipment power consumption. The computation load of some small cells could be redistributed to others at the cost of increasing the experienced latency. The proposed joint

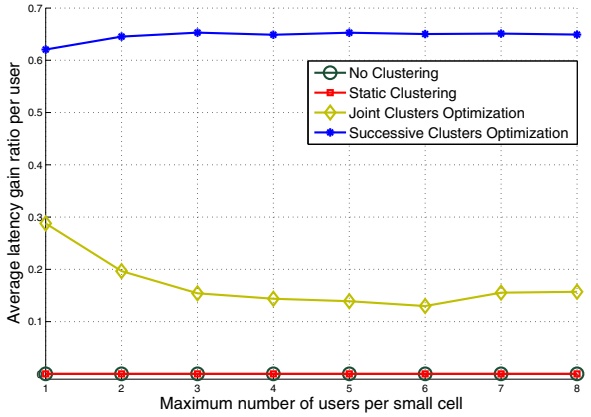

Fig. 3: Average user latency gain in dependence on the number of users per small cell

clustering optimization might not give the ultimate optimal solution for the multi-user computation clusters establishment problem. This is due to the fact that the computing small cells are in an idle state, i.e. not performing any computation, during the time they are receiving input data. The data transmission time depends itself on the cluster establishment and especially on the distribution load and transmission powers. This idle time is seen as a loss of computational resources that could eventually be used for serving local devices at each small cell. This issue is not addressed in the proposed optimization problem since it is a one shot optimization that has as a starting point the initial state and characteristics of the requests and small cells. A possible way to overcome this problem, which we are currently investigating, is to launch the optimization problem with a different starting point of the system state that allows better exploitation of computational resources. However, and despite its non-optimality, the proposed method achieves relevant gains comparing to static strategies where each computation cluster is predefined.

V. Conclusion

In this paper, we proposed an multi-user small cell clustering optimization policy for distributed fog computing. The interest of this approach is two-fold. First, it allows adaptive sizing and resources management of computation clusters. Second, it simultaneously establishes computation clusters for all active requests for better exploitation of available resources, targeting a higher QoE. The proposed method is a joint optimization of computational load distribution, computational rate allocation and transmission power control. Simulations results showed the effectiveness of the derived solution despite its non-optimality. A high QoE is achieved even for a large number of devices per small cell. Future work includes exploiting trade-offs in cluster formation in the multi-user case for reducing the number of active small cells. In addition, we are investigating an updated implementation of this method in order to cope for its non-optimality due to the unpredictable idle time of the computing small cells.

References

[1] Ercisson, "Ericsson Mobility Report", Nov 2014. http://www.ericsson.com/res/docs/2014/ericsson-mobility-report-november-2014.pdf

[2] Yonggang Wen; Weiwen Zhang; Haiyun Luo, "Energy-optimal mobile application execution: Taming resource-poor mobile devices with cloud clones," INFOCOM, 2012 Proceedings IEEE, pp.2716-2720, 25-30 March 2012.

[3] Lagerspetz, E.; Tarkoma, S., "Mobile search and the cloud: The benefits of offloading," Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on, pp.117-122, 21-25 March 2011.

[4] Oueis, J.; Calvanese Strinati, E.; Barbarossa, S., "Multi-parameter decision algorithm for mobile computation offloading," Wireless Communications and Networking Conference (WCNC), 2014 IEEE, pp.3005-3010, 6-9 April 2014

[5] Barbarossa, S.; Di Lorenzo, P.; Sardellitti, S., "Computation offloading strategies based on energy minimization under computational rate constraints," Networks and Communications (EuCNC), 2014 European Conference on, pp.1-5, 23-26 June 2014

[6] Xiaohui Gu; Nahrstedt, K.; Messer, A.; Greenberg, I.; Milojicic, D., "Adaptive offloading inference for delivering applications in pervasive computing environments," Pervasive Computing and Communications, 2003. (PerCom 2003). Proceedings of the First IEEE International Conference on, pp.107-114, 23-26 March 2003.

[7] Guangyu Chen; Kang, B.-T.; Kandemir, M.; Vijaykrishnan, N.; Irwin, M.J.; Chandramouli, R., "Studying energy trade offs in offloading computation/compilation in Java-enabled mobile devices," Parallel and Distributed Systems, IEEE Transactions on, vol.15, no.9, pp.795-809, Sept. 2004.

[8] FP7 project TROPIC: "Distributed computing storage and radio resource allocation over cooperative femtocells," ICT-318784, www.ict-tropic.eu

[9] O. Muñoz, A. Pascual-Iserte, and J. Vidal, "Optimization of Radio and Computational Resources for Energy Efficiency in Latency-Constrained Application Offloading," submitted to IEEE Transactions on wireless communications, June 2013.

[10] Barbarossa, S.; Sardellitti, S.; Di Lorenzo, P., "Joint allocation of computation and communication resources in multiuser mobile cloud computing," Signal Processing Advances in Wireless Communications (SPAWC), 2013 IEEE 14th Workshop on, pp.26-30, 16-19 June 2013

[11] Sardellitti, S.: Barbarossa, S.; Scutari, G., "Distributed Mobile Cloud Computing: Joint Optimization of Radio and Computational Resources," Global Telecommunications Conference, 2014. Globecom'14, December 2014.

[12] Oueis, J.; Calvanese Strinati, E.; De Domenico, A.; Barbarossa, S., "On the impact of backhaul network on distributed cloud computing," Wireless Communications and Networking Conference Workshops (WCNCW), 2014 IEEE, pp.12-17, 6-9 April 2014

[13] Oueis, J.; Calvanese Strinati, E.; Barbarossa, S., "Small Cell Clustering for Efficient Distributed Cloud Computing," Personal Indoor and Mobile Radio Communications (PIMRC), 2014 IEEE 25th International Symposium on, 9-12 Sept. 2012.

[14] 3GPP TSG-RAN4#51, Alcatel-Lucent, picoChip Designs, and Vodafone, "R4-092042, Simulation assumptions and parameters for FDD HENB RF requirements," May 2009.