Computer Networks 108 (2016) 357–370

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Path selection enabling user mobility and efficient distribution of data for computation at the edge of mobile network

Jan Plachy*, Zdenek Becvar, Pavel Mach

Department of Telecommunication Engineering, Faculty of Electrical Engineering, Czech Technical University in Prague, Technicka 2, 166 27 Prague, Czech Republic

ARTICLE INFO

Article history:
Received 29 April 2015
Revised 30 June 2016
Accepted 6 September 2016
Available online 9 September 2016

Keywords:
Mobile edge computing
Mobility management
Path selection
Handover
Small cell cloud
Energy efficiency

ABSTRACT

Convergence of mobile networks and cloud computing enables to offload heavy computation from a user equipment (UE) to the cloud. The offloading can reduce energy consumption of the UEs. Nevertheless, delivery of data to a centralized cloud leads to high latency and to overloading backhaul network. To overcome these constrains, computing capabilities can be brought closer to the user and integrated into small cell base stations deployed in mobile networks. This concept of cloud-enabled small cells is known as small cell cloud (SCC). In the SCC, the UEs benefit from proximity to the computing stations resulting in both lower latency and alleviating load of backhaul. In this paper, we propose a path selection algorithm finding the most suitable way for data delivery between the mobile UE and the cells performing computation for this particular UE. The path selection algorithm estimates transmission delay and energy consumed by the transmission of offloaded data and selects the most suitable base station for radio communication accordingly. The path selection problem is formulated as Markov Decision Process (MDP). The algorithm is suitable for parallel computation in dynamic scenarios with mobile users and handles mobility for users exploiting computing services in the SCC. Comparing to conventional approach for delivery of data to computing cells, the proposed algorithm reduces the delay up to 54.3% and UE's energy consumption is decreased by up to 7.5%. Moreover, users' satisfaction with data transmission delay is increased by up to 28% and load of small cell's backhaul is lowered by up to 29%.

1. Introduction

As demands of mobile users are being shifted from hardware to software [1], an opportunity to offload computation from a user equipment (UE) into the cloud is becoming an interesting possibility. This option provides enough computing power even for computationally demanding applications while energy consumption at the side of the UEs is lowered. However, conventional mobile cloud computing (MCC) approaches [2,3] are characterized by a significant delay in delivery of offloaded data from the UE to a computing machine and back [4]. Therefore, exploitation of MCC for delay sensitive or real-time applications is limited. Moreover, offloading of computation to the conventional centralized cloud can overload backhaul network because of additional traffic generated by the offloaded applications. Last but not the least, the offloaded applications also result in higher energy consumption of whole network as each device has to transfer and handle more traffic.

To reduce network energy consumption and delay because of data transmissions, network and cloud infrastructure should be managed jointly [5]. Nevertheless, this does not solve the problem of high network load caused by application offloading and, in addition, the minimum achievable transmission delay is still high for delay sensitive applications. To overcome the problem of delay and high backhaul load, cloud resources should be deployed closer to the users. In common cellular networks, the closest place for deployment of computing resources is a base station (denoted as eNB in LTE-A). With emerging trend of dense deployment of small cell base stations (SCeNBs), these are seen as a convenient mean to distribute cloud computing resources to proximity of users. This concept, known as Small Cell Cloud (SCC) [6–8], enables computing at the edge of mobile network. In order to facilitate the SCC concept, the SCeNBs are empowered by additional computing and storage resources [9]. Bringing computing power closer to the UEs decreases the delay caused by offloaded data transmission. At the same time, load of the backhaul network is lowered as a part of offloaded data is processed directly at the edge of mobile network. Lowering data delivery delay enables to exploit the SCC also for delay-sensitive services and applications, such as gaming, augmented reality, or image processing [10].

* Corresponding author.
E-mail addresses: jan.plachy@fel.cvut.cz (J. Plachy), zdenek.becvar@fel.cvut.cz (Z. Becvar), machp2@fel.cvut.cz (P. Mach).

http://dx.doi.org/10.1016/j.comnet.2016.09.005
1389-1286/

To satisfy even high demands of the UEs on computation, the computing power distributed over nearby cloud-enhanced SCeNBs can be virtually merged together to form computing clusters. At each cluster, Virtual Machines (VM) [11] can be run. The VMs are deployed at the SCeNBs with respect to their communication and computation capabilities. Selection of the SCeNBs forming the computing cluster and management of the computation according to the overall state of the network (i.e., current radio, backhaul and VMs state) is done in Small cell Cloud Manager (SCM) [6].

In the SCC, the application is offloaded from the UE to the SCeNBs if it is profitable from energy consumption and/or delay perspective [11,12]. After selection of the SCeNBs, which take care of computation, data must be delivered to these cells. Typically, the SCeNBs are connected to network through a low quality backhaul comparing to common backhaul of the macrocell eNBs. Hence, distribution of data for computation from the cell providing radio access (denoted as serving cell) to all computing cells through the backhaul of limited capacity (e.g., DSL) can lead to significant delay. To that end, it is efficient to deliver data to selected computing cells not only through the serving cell but also via neighboring cells provided that those are in the user's radio communication range. In mobile networks, switching radio communication from the serving cell to another cell (labeled as target cell) in UE's neighborhood is known as handover. The purpose of handover in mobile networks is to provide seamless connection to moving users. The handover is usually initiated according to radio channel quality offered by the serving and target cells [13,14], available capacity of backhaul [15], or energy consumption of the UE [16].

In this paper, we propose an algorithm exploiting handover in order to avoid distribution of offloaded data via a backhaul of limited capacity. Our motivation is to shorten the time necessary for transferring the offloaded data to individual computing SCeNBs. To prevent high energy consumption at the UE side, the energy spent by the UE for data transmission as well as energy spent by handover itself is also considered for selection of the most suitable way of data delivery. The problem is formulated as a Markov Decision Process (MDP). In the MDP, any change of the serving SCeNB (i.e., each handover) motivated by an improvement of data delivery is rewarded depending on its impact on the UE's energy consumption and transmission delay caused by both data transmission and handover. The algorithm selects also the path for delivery of computation results back to the UE. Independent selection of the communication paths for data offloading (uplink) and results delivery to the UE (downlink) solves problem of mobility management for users exploiting the SCC (i.e., the problem of users moving from one cell to another when the offloaded application is currently computed). Consequently, the SCC can be efficiently utilized also by moving users. In addition, the algorithm is suitable for parallel computing, so, parallelized parts of code (offloaded data) can be delivered to multiple computing cells via multiple routes to minimize the transmission delay.

This paper is an extension of our previous work presented in [17], where we have proposed general framework for the path selection and we have provided basic performance analysis. With respect to [17], we extend our work in the following aspects: 1) we consider path selection not only for uplink data offloading but also for downlink reception of computation results in order to address a problem of user mobility management; 2) we present more detailed description of the proposed algorithm including implementation aspects related to derivation of required parameters; 3) we enhance simulations by consideration of multi-user multi-cell scenario and user's mobility; 4) we evaluate also impact of the proposed algorithm on the load of backhaul network.

The rest of this paper is organized as follows. In the next section, we summarize related work. In Section 3, the proposed algorithm for path selection is described along with implementation aspects. Simulation methodology and scenario are presented in Section 4. Section 5 provides performance evaluation and discussion of simulation results. The last section summarizes major conclusions and outlines potential future research work.

2. Related work

Conventional mobile network can be represented by "tree" topology with a common centralized node (core network). We consider possibility to perform handover during transmission of offloaded data for computation if it is beneficial from transmission delay or UE's energy consumption perspective. This makes the routing of data from the UE to the computing cells more flexible and it changes the "tree" topology with data routed through the core network to more "ad-hoc-like" topology. Thus, selection of the most appropriate way for data delivery to the computing cells (or back to the UE) becomes problem analogical to routing in Wireless Sensor Networks (WSN). Routing protocols designed for WSN [18] provide an inspiration how to treat the path selection in the SCC. However, in the SCC, the energy is key limiting factor only for radio communication part (i.e., between the UE and the SCeNBs). Also, dynamicity of the system is inherent feature of mobile networks. Therefore, energy consumed by the source node (in our case, the UE) as well as dynamic path selection must be taken into account.

The dynamicity of scenario and multipath communications are investigated in [19], where issues and challenges of multipath routing are described. Nevertheless, this approach assumes only hop count for the selection of routing path (i.e., the number of hops between source and destination nodes). Dynamicity together with path selection based on Received Strength Signal Indicator (RSSI) are addressed by Ad-hoc On-demand Multipath Distance Vector with Dynamic Path Update (AOMDV-DPU) [20]. However, even selection of paths with good RSSI to avoid weak radio links does not guarantee low delay for the SCC due to communication over backhaul, which is usually of a lower quality. Furthermore, the AOMDV-DPU does not consider transmission energy, which is essential in our case. Similarly, the algorithm presented in [21] proposes to route data based on RSSI, latency, and node occupancy, but it does not consider energy consumption. In order to combine transmission delay and energy, Power and Delay-aware Multi-path Routing Protocol (PDMPRP) is proposed in [22]. The PDMPRP chooses multi-paths in order to minimize energy consumption without increasing delay. With respect to [18–22] where whole network is wireless, backhaul from the serving cell to the operators core network is typically wired. In addition, if the serving cell selection is based on RSSI, the same path to the core network would be selected all the time disregarding the SCeNBs selected for computation and backhaul status.

Mobile network topology, even with consideration of handover, does not enable such freedom in routing as the conventional WSN but it rather corresponds to a hierarchical network structure in WSN where some nodes are selected as gateways (cluster heads) relaying data to a destination [18]. In our scenario, the SCeNBs can be seen as gateway nodes. Each gateway has a fix number of options how to distribute offloaded data to computing cells through fixed network infrastructure. This infrastructure is represented typically by a wired backhaul and core network of the operator. Therefore, the problem consists in selection of proper gateway (serving cell) for individual parts of offloaded data. The selected gateway must minimize data transmission delay and energy consumed by the UEs for the transmission. The same problem applies also for delivery of computation results back to the UE. The reason is that the conditions in uplink and downlink are usually different. Moreover, the UE can change position between the time instance of offloading and the time instance of reception of the computing re-

sults. The UE's movement is reflected by change of radio channel or even change of the serving base station. Also, the energy spent for transmission of computation results by the SCeNB is not such a limiting factor, since the SCeNBs are not powered by short life-time batteries.

Problem of path selection for scenario considering common mobile cloud computing is addressed in [23] where the authors propose to select the path using fuzzy logic. This idea covers selection of target cloud offloading system based on the path parameters such as delay, packet loss, and benefits of offloading. However, this solution focuses only on centralized cloud services while radio aspects or mobility of users and possibility of handover, are not reflected.

Authors in [24] propose three clustering strategies, which select a set of computing SCeNBs together with wired path (excluding radio) to computing cluster. The objective of these clustering strategies is to minimize either cluster latency, cluster power consumption, or SCeNB power consumption. Contrary to [24], we focus on minimization of energy consumed at the UE and possibility to change radio path (between UE and SCeNB) for distribution of parallelized computation at several SCeNBs. In addition, our approach considers jointly energy consumed by the UE and transmission delay.

In [25] the authors investigate whether it is efficient to migrate VMs from one node to another during user's movement. The authors use MDP along with a threshold policy-based mechanism to optimize the VM migration. The proposed algorithm is designed for 1-D mobility model without consideration of energy consumption and actual path selection. Further enhancement of the VM migration by exploitation of prediction is investigated in [26]. The enhancement consists in prediction, with a given accuracy, of future cost of VM placement and migration. Comparing to [25,26], our proposed path selection algorithm assumes that computation for one offloading task is done by the same SCeNBs (i.e., no migration of VM is considered) as the VM migration leads to a significant delay [27] due to the transfer of the VM from one SCeNB to another one. Furthermore, our algorithm selects the proper SCeNBs through which the UE should transmit/receive offloaded task in order to minimize offloaded task delay and/or UE's energy consumption (based on weighting) during its movement.

In existing approaches focusing on task offloading into the SCC, the data to the computing cells is always delivered through the static serving cell [6,11]. It means the UE is still attached to the same cell during delivery of whole offloaded data. Then, the serving cell distributes data through operator's core network to the computing cells. This approach can be efficient if both radio channel between the UE and its serving cell as well as backhaul connection of the serving and all computing cells are of sufficient throughput. Otherwise, a limitation at any part of the communication chain leads to a prolongation of the overall delay due to computation offloading.

3. Path selection algorithm

In this section, we present system model and proposed algorithm for data delivery from the UE to individual cells performing computation and delivery of computing results back to the UE. Designed path selection algorithm takes into account UEs energy consumption (both energy spent by data transmission and handover), handover delay, radio channel quality, and backhaul conditions. Furthermore, we discuss implementation aspects and possible reduction of computation complexity. For easy following the algorithm description, notation of all key parameters used in this paper is summarized in Table 1.

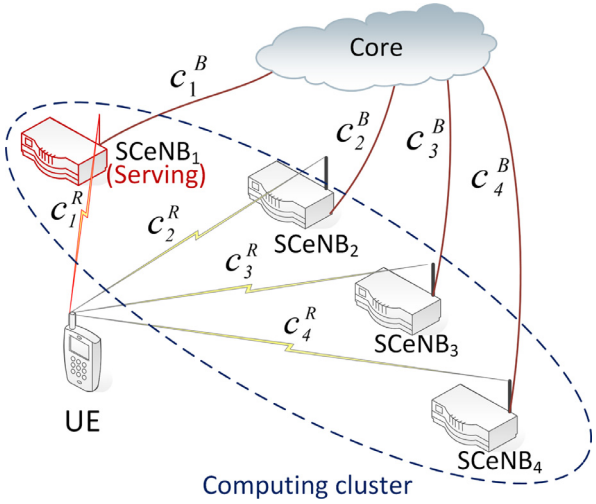

Fig. 1. Network topology and definition of parameters required for path selection.

3.1. System model

We assume the system composed of S SCeNBs and U UEs. Furthermore, for each UE, we define set X consisting of n computing SCeNBs and set I consisting of m SCeNBs that are in the neighborhood of the UE and can communicate with the UE directly through the radio link. Note that sets X and I may be fully or partially overlapping if computing cells are in radio communication range of the UE (i.e., if the UE can connect directly to the computing cells via radio link). In our model, the SCeNB providing the highest RSSI to the UE is selected to be the serving cell for each UE [13,29]. In case of the UE's movement, the serving cell is updated if the RSSI from the target SCeNB ($RSSI_{TC}$) becomes higher than the RSSI of the serving cell ($RSSI_{SC}$) plus handover hysteresis ($\triangle_{HM}$), i.e., if $RSSI_{TC} > RSSI_{SC} + \bigtriangleup_{HM}$.

An example of the network model is shown in Fig. 1. In this figure, c represents capacity of the link and upper indexes B and R stand for backhaul and radio links, respectively. In the given example, the cluster of cells performing computation is formed of four SCeNBs. Out of those SCeNBs, the $SCeNB_1$ is selected as the serving cell.

As depicted in Fig. 1, data from the UE can be transferred to the $SCeNB_i$ over the radio link with capacity c_i^R. The $SCeNB_i$ is connected to the operator's core via the backhaul with capacity c_i^B. The offloaded data is processed by the $SCeNB_i$ or forwarded to another computing $SCeNB_x$ through backhaul of the $SCeNB_i$ (with capacity c_i^B in uplink) and backhaul of the computing cell $SCeNB_x$ (with capacity c_x^B in downlink). Note that index x stands for any SCeNB out of X except the $SCeNB_i$ (i.e., $x = 2, 3, 4$ in Fig. 1). Selection of the computing cells can be done according to complexity of the offloaded data processing and available computing power of the SCeNBs as suggested, e.g., in [30–31]. After the SCeNBs finish data computation, the results are delivered back to the UE. New path for backward delivery of computation results (from each $SCeNB_x$ to the UE) must be derived if radio and backhaul links are not symmetric in uplink and downlink, if the UE moves during computation, or if the channel/link load or quality changes. Therefore, computation results can be delivered to the UE through a new cell(s), which again minimize delay and/or energy consumption according to the user's preference.

In case when the offloaded task requires simultaneous uploading (task offloading) and downloading (results reception), both are done via the same serving cell selected by the proposed path selection algorithm. It means, the UE communicates over the same

Table 1
Definition of parameters and sets used for the explanation of the proposed path selection algorithm.

Sets and Parameters	
D_q	Delay of the q-th path consisting of delay caused by radio and backhaul transmission and duration of handover (known also as handover interruption [28]).
$\overline{D_q}$	Normalized delay of the q-th path scaled to a range of (0,1].
E_q	Energy consumption caused by transmission/reception of data using the q-th path; the energy consists of energy required for transmission/reception over LTE-radio interface and energy spent by handover (transmission/reception of related signaling).
$\overline{E_q}$	Normalized energy of the q-th path scaled to a range of (0,1].
s/s'	Current/future state in Markov Decision Process.
a	Action in Markov Decision Process.
k	Duration of transmission of data represented as the number of steps required for transmission. The step is understood as a discrete time value considered for computation of transmission *delay* and *energy*, e.g., one frame in LTE-A with duration of 10 ms.
π	Optimal policy.
D_R	Duration of data transmission (delay) over radio link.
D_B	Duration of data transmission (delay) over backhaul.
D_H	Delay due to handover interruption.
$v_{bits}^{B,i}$	Amount of bits to be transferred over backhaul of i-th SCeNB.
$v_{bits}^{R,i}$	Amount of bits to be transferred over radio of i-th SCeNB.
c_i^R	Capacity of radio link between the UE and the i-th SCeNB.
c_i^B	Capacity of backhaul of the i-th SCeNB (between i-th SCeNB and core network).
c_x^B	Capacity of backhaul of the x-th computing SCeNB (between core network and x-th SCeNB).
d_x^i	Delay due to transmission of data through the i-th SCeNB's radio and backhaul, used for transmission of an offloaded task to x-th computing SCeNB.
e_x^i	Energy consumed by the UE spent for radio transmission of an offloaded task to x-th computing SCeNB via the i-th SCeNB.
L_x	Size of the subtask for computation handled by the x-th SCeNB.
λ_x	Ratio of the whole task computed by the x-th SCeNB.
I	Set of cells within radio access of the UE.
m	Number of SCeNBs within radio communication range of the UE. $m = \|I\|$.
X	Set of all cells performing computation of a single task (cells participating on parallel computation of the task).
n	Number of computing SCeNBs (i.e., number of parallelized subtasks). $n = \|X\|$.
Q	Set of all possible paths from the UE to the computing cells in X.
p	Number of possible paths from the UE to the computing cells in X. $p = \|Q\|$.
M_q	Total metric of the q-th path, consisting of delay and energy.

serving cell for uplink and downlink until all data is transmitted and received. Then, if handover is beneficial, new serving cell is selected (still just one serving cell for both uplink and downlink). Changing the UE's serving cell for offloading and results reception can lead to a change in assigned IP address and, thus, it may lead to problem with routing of data to destination. Nevertheless, this problem is solved by a method for addressing devices in MEC as outlined in [9].

Each computing task offloaded to the SCC (in this paper denoted as "offloaded task") is divided between the computing cells. Each $SCeNB_x$ (where $x \in X$) is expected to compute a part $\lambda_x \in (0, 1]$ of the whole offloaded task, which is of the overall size of L_{UE}. The individual part L_x computed by the $SCeNB_x$ is then expressed as $L_x = \lambda_x \cdot L_{UE}$ where $\sum \lambda_x = 1$. In this paper, we assume to split the offloaded task into parts with the same size among all computing cells, i.e., $\lambda_1 = \lambda_2 = \ldots = \lambda_x$. In general, the size of each offloaded part should correspond to the computing power of individual SCeNB involved in computation. The optimal distribution of offloaded task to individual computing SCeNBs is out of scope of this paper and this topic is left for future research.

The common approaches for delivery of the offloaded task to the computing cells and back assume that data from the UE to the computing cells is always delivered through the same serving cell ($SCeNB_1$ in Fig. 1) [6,11]. This serving cell is selected only according to the rules applied in common mobile networks, i.e., with respect to radio channel quality or available capacity of radio channel [14,29]. To overcome potential delay due to distribution of data among all computing $SCeNB_x$ over the backhaul with limited throughput, we exploit an opportunity to transfer data also via neighboring cells. In this case, individual parts of the data for computation are delivered to individual computing cells through specific neighbors, which offer the lowest transmission delay over both radio and backhaul links.

Note that for each computing cell, data can be delivered through different neighboring cell. This implies a need for performing handover during communication. In our proposed algorithm, handover is not enforced during transmission of the offloaded task to each of assigned computing cells. Instead, handover is performed when no data is being transmitted at the moment. For example, following Fig. 1, data designated to be computed at the $SCeNB_1$ is transmitted to this SCeNB. After successful transmission, handover is performed before data to the next SCeNB (e.g., $SCeNB_2$) is transmitted. Therefore, offloading is not interrupted by handover as each part of the task is offloaded/received at its destination before handover.

Comparing to the conventional handover in mobile networks, the handover is not conditioned only by radio quality but also by backhaul [15] and selection of computing cells. The proposed algorithm is labeled as a Path Selection with Handover (PSwH). The scheme using single serving cell selected in conventional way according to radio quality for delivery of all offloaded data is denoted as a Serving Only (SO) in the rest of the paper.

3.2. Path selection exploiting handover

The proposed path selection algorithm suitable for the SCC combines the time required for transmission of offloaded task over radio and backhaul links (in the rest of paper denoted as *transmission delay*) and energy consumed by the UE by both transmission of data over radio link and handover (in the rest of paper denoted as *energy*).

Each q-th path between the UE and the computing cell is described by the transmission delay (D_q) and the energy consumed by the UE's transmission (E_q). In our path selection algorithm, both D_q and E_q are combined into a single metric of the q-th path, M_q:

$$M_q = \gamma \overline{E_q} + (1 - \gamma)\overline{D_q} \quad (1)$$

where γ is the weighting factor showing preference for low delay ($\gamma \to 0$) or for high energy efficiency ($\gamma \to 1$), $\overline{D_q}$ ($\overline{E_q}$) represents normalized delay (energy) of the q-th path. In order to enable combination of both metrics, both are normalized (i.e., scaled into range from 0 to 1) with respect to the maximum observed

value as follows:

$$\overline{D_q} = \frac{D_q}{\max\{D_1, D_2, \ldots D_p\}} \tag{2}$$

$$\overline{E_q} = \frac{E_q}{\max\{E_1, E_2, \ldots E_p\}} \tag{3}$$

where p is the number of possible paths from the UE to the computing cells. Parameter p is calculated as the cardinality of Q, i.e., $p = |Q|$, where Q is the set of all possible paths including all combinations of computing SCeNBs (set X) and all SCeNBs within radio communication range of the UE (set I). The path selection algorithm is defined as the MDP, which determines reward (penalty) of transition from the current state s to one of possible future states s' [33]:

$$V_\pi^k = Est\left(\sum_k R^t | \pi, s\right) = R(s) + \sum_k T\left(s, \pi(s,k), s'\right) V_\pi^{k-1}(s') \quad \pi : s \to a \tag{4}$$

where the state s represents currently selected path q_s (using the serving cell selected in conventional way according to radio link quality) and the future state s' is another possible path $q'_{s'}$ (composed of radio and backhaul connections) out of Q. Note that Q includes also paths obtained by performing handover to neighboring cells. Total reward for transition from the state s to the s' consists of two parts (see (4)). The first part, $R(s)$, denotes immediate reward for transition from the state s. The second part, summation of reward per time step ($T(s, \pi(s,k), s')V_\pi^{k-1}(s')$), represents expected future payoff as a sum over k steps. The estimate (Est) represents possible reward by utilizing a new path instead of the currently used one. As the radio and backhaul parameters fluctuate over time, calculated time of transmission is only an estimation of the expected transmission time. This estimation introduces an error in derivation of the reward. The estimation error can be avoided by reservation of radio and backhaul resources solely for the purposes of data offloading to the SCC, i.e., using Guaranteed Bit Rate (GBR) [34] in LTE-A networks. Nevertheless, this would lead to QoS degradation for other UEs. Thus, Est is computed as a sum over k steps, representing estimated duration of the data transmission. Minimization of the estimation error is left for future research. Parameter π in (4) stands for the policy, which defines what action (a) should be taken in the state s to maximize the total reward. We define two actions that can be done to maximize the total reward: 1) transit to another state s' (change current path) if it improves M_q or 2) stay in the current state s (use the same path) if M_q cannot be improved by selection of another path. However, with dynamicity of the mobile networks, transitions from one state to another (option 1) cannot be stationary mapped to states and need to reflect changes of the network topology and transmission parameters of each link. Thus, we calculate table of transitions among states every time when the task is offloaded. Optimum policy π is obtained at the end of the algorithm and it gives desired policy maximizing the reward. As the delay and the energy are used as metrics, the optimal policies can be calculated in order to minimize delay, energy or a trade-off between both metrics. The reward depends on the delay due to handover (D_H) if the handover is performed, delay by the transmission over radio (D_R) and transmission delay on backhaul (D_B).

Lets demonstrate our MDP approach on an example of a chain of two consecutive time steps as shown in Fig. 2. Note that the presented MDP problem is derived from Fig. 1, where each SCeNB represents one state (i.e., 1, 2, 3, and 4). At the beginning of the process (at time $t = 0$), current serving SCeNB is $SCeNB_1$. Thus, the current state is $s = 1$. In the next time step (i.e., $t = 1$), a transition

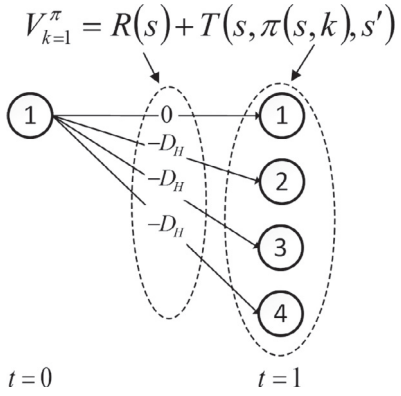

Fig. 2. Reward on a chain of a single time step (i.e., $k = 1$).

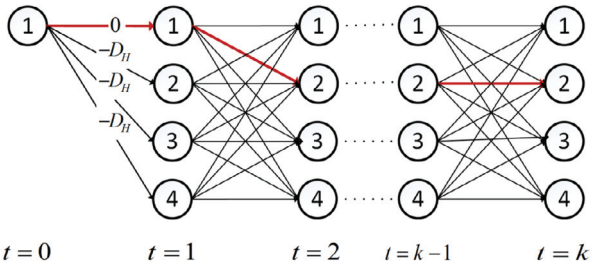

Fig. 3. Chain for calculation of the total reward (for the sake of figure clarity, negative rewards $-D_H$ due to handover between states is not depicted for $t > 1$).

to four different states is possible. It means the SCeNB may stay in the same state s (i.e., $s = 1$) or may change to a different one (i.e., $s' \in \{2, 3, 4\}$). These states are connected with the current state via edges denoting the immediate transition reward $R(s)$. The immediate transition reward of staying in the current state (s) is 0 as the serving SCeNB remains the same (there is no handover) and thus no gain (loss) is obtained by this transition. In case of the transition to another state (i.e., s'), there is a negative immediate reward due to handover (in Fig. 2 denoted as $-D_H$). The handover introduces negative immediate reward as it leads to an overhead and no useful data are being transmitted during the handover. Nevertheless, in case of the transition to another state (2, 3, 4), there is also expected future reward (in Fig. 2 denoted as $T(s, \pi(s, k), s')$), which is calculated as a reward introduced by connection to a new SCeNB ($SCeNB_2$, $SCeNB_3$, or $SCeNB_4$) and staying there for a duration of a one time step.

To obtain the total reward V_k^π, the procedure from Fig. 2 has to be repeated multiple times (for all time steps) until all required data is transmitted as shown in Fig. 3. Then, the selected path is represented by a chain of serving SCeNBs (e.g., red line starting in state 1 in time $t = 0$ and ending in state 2 in time $t = k$). Thus, the reward for transition from the state s to the s' defined in (4) can be rewritten for purposes of the path selection as follows. In our proposal, the immediate reward ($R(s)$) is represented by handover delay $D_H(q_s, q'_{s'})$ and energy consumed during handover $E_H(q_s, q'_{s'})$. The reward in terms of delay per step (in (4) denoted as $T(s, \pi(s,k), s')V_\pi^{k-1}(s')$) is calculated as a difference in communication delay if path $q'_{s'}$ is selected instead of q_s for both radio (i.e., $D_R(q_s) - D_R(q'_{s'})$) and backhaul (i.e., $D_B(q_s) - D_B(q'_{s'})$). The reward in terms of UE's energy consumption per step is calculated as a difference in energy consumed if the communication over radio takes place via path $q'_{s'}$ instead of q_s(i.e., $E_R(q_s) - E_R(q'_{s'})$).

All rewards (delay and energy) are added up to obtain the total reward. We also reflect the possibility of weighting energy and delay (as defined in (1)) by parameter γ. Thus, the total reward for

our proposal is defined as:

$$V_{\pi}^{k}(q_s, q'_{s'}) = \gamma\left[-E_H(q_s, q'_{s'}) + \sum_k \left(E_R(q_s) - E_R(q'_{s'})\right)\right] + (1-\gamma)\left[-D_H(q_s, q'_{s'}) + \sum_k \left(D_R(q_s) - D_R(q'_{s'})\right) + \sum_k \left(D_B(q_s) - D_B(q'_{s'})\right)\right] \quad (5)$$

where $E_R(q_s)$ and $E_R(q'_{s'})$ denote the energy consumed by the UE's radio communication using current path q_s and the new path $q'_{s'}$, respectively, $D_H(q_s, q'_{s'})$ and $E_H(q_s, q'_{s'})$ stands for the delay and the energy consumed by handover from the serving cell to the neighboring cell (transition from the path q_s to the path $q'_{s'}$), respectively. Delay due to the handover ($D_H(q_s, q'_{s'})$) and energy consumed by the UE during handover procedure ($E_H(q_s, q'_{s'})$) reflect an overhead in terms of additional delay and energy consumption caused by the handover procedure, respectively.

The transmission delays D_R and D_B are computed knowing amount of data to be transferred over radio ($v_{bits}^{R,i}$) and backhaul ($v_{bits}^{B,i}$) and knowing capacity of the radio link (c_i^R), capacity of backhaul of the serving (c_i^B) and the computing (c_x^B) cells:

$$D_R = \frac{v_{bits}^{R,i}}{c_i^R} \quad (6)$$

$$D_B = \frac{v_{bits}^{B,i}}{c_i^B} + \frac{v_{bits}^{B,i}}{c_x^B} \quad (7)$$

The energy consumption at the UE due to radio transmission/reception (E_R) is computed, according to the 3GPP [35–39], based on received signal, required throughput, used MCS, and signal propagation (path loss). The energy E_R represents the energy spent for either the task offloading in uplink (E_R^{UL}) or reception of computation results in downlink (E_R^{DL}).

The energy consumption in uplink, E_R^{UL}, depends on Modulation and Coding scheme (MCS) and available bandwidth represented by RBs in LTE-A system. The MCS is a function of Signal to Interference plus Noise Ratio (SINR) observed at the receiver. The SINR at the eNB is proportional to the UE transmission power (P_{Tx}), path loss and interference from other UEs (in the FDD case). In LTE-A, the P_{Tx} required for selected MCS for a given number of allocated RBs is defined as follows (see [36] and [39]):

$$P_{Tx} = min(P_{MAX}, P_0 + \alpha \cdot PL + 10log_{10}(M) + \Delta_{TF} + f) \quad (8)$$

where P_{MAX} is the maximum available transmission power (23 dBm for the UE class 3 [39]); $\alpha \in \{0, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ corresponds to the path loss compensation factor, PL is the downlink path loss estimate, M stands for the number of assigned RBs, Δ_{TF} represents a closed loop UE specific parameter based on the applied MCS, f is a correction value (also referred to as a TPC command [36,39]), and parameter P_0 represents the power offset computed as:

$$P_0 = \alpha \cdot (SINR_0 + P_N) + (1-\alpha) \cdot (P_{MAX} - 10log_{10}(M_0)) \quad (9)$$

where P_N is the noise power per RB, and M_0 defines the number of allocated RBs for the case when the UE would be transmitting with its maximal transmission power).

Parameters Δ_{TF} and f are used for dynamic adjustment of the transmission power to keep required SINR at the receiver. As we assume open loop power control, we can omit these parameters as indicated in [39]. The parameter α is set to 1 so the UE fully compensates the path loss. Under these assumptions (commonly considered in real mobile networks), we can simplify power offset to $P_0 = \alpha \cdot (SINR_0 + P_N)$ and then, (8) can be rewritten as:

$$P_{Tx} = min\{P_{MAX}, \alpha \cdot (SINR + P_N + PL) + 10log_{10}(M)\} \quad (10)$$

Finally, the energy consumed by transmission of data over a radio channel is derived, according to [39], as:

$$E_R^{UL} = P_{Tx} \cdot D_R \quad (11)$$

For the reception of computation results in downlink, the energy E_R^{DL} is derived based on knowledge of the power required for data reception (P_{Rx}). This power depends on the level of signal received by the UE (S_{Rx}) and data rate as specified in [40]. In this paper, the power consumption model defined in [38] is used to match the power consumption to data transmission as follows:

$$P_{Rx} = P_{RxRF} + P_{RxBB}[mW] \quad (12)$$

$$P_{RxRF} = \begin{cases} -0.04S_{Rx} + 24.8[mW] & S_{Rx} \le -52.5dBm \\ -0.11S_{Rx} + 7.86[mW] & S_{Rx} > -52.5dBm \end{cases} \quad (13)$$

$$P_{RxBB} = 0.97R_{Rx} + 8.16[mW] \quad (14)$$

where P_{RxRF} denotes the radio component depending on the S_{Rx}, and P_{RxBB} is the radio component depending on bit rate of received transmission.

The energy consumed for downlink reception of computing results is calculated in the similar way as for uplink, i.e.:

$$E_R^{DL} = P_{Rx} \cdot D_R \quad (15)$$

3.3. Path selection algorithm

The pseudo-code for the proposed algorithm, which selects the path between the UE and the computing SCeNBs for given γ is shown in Algorithm 1. The algorithm calculates delay D_q and energy E_q spent by the UE for delivery of the offloaded task to each $SCeNB_x$ (Step 3) using available radio links of neighboring cells (Step 4). Delay and energy due to transmission using radio of the $SCeNB_i$ to deliver data to the $SCeNB_x$ are derived using (6), (7) and (11), (15), respectively (Steps 6 and 7). If data is sent over backhaul link (Step 8), its delay is added to the path delay (Steps 9, 10). Afterwards, the delay and energy of each combination of radio and backhaul links is calculated (Steps 15 and 16). Impact of handover on the path selection is included by adding delay of handover (D_H) and energy consumed by the UE during handover (E_H) to the delay and energy derived for the $q-th$ path (Steps 18 and 19). Energy consumed by the UE during handover is calculated using (11) by substituting D_H for D_R. Subsequently, D_q and E_q are normalized in order to be weighted (Steps 23 and 24). Then, the path metric M_q is calculated by weighting $\overline{D_q}$ and $\overline{E_q}$ (Step 25). Finally, the new path $q'_{s'}$ with the lowest M_q for given γ is returned (Steps 27, 28).

3.4. Implementation aspects

To enable implementation of the proposed path selection algorithm, capacity of radio link, capacity of backhaul link, and transmission/reception power level at the UEs side must be obtained. The capacity of uplink radio link is derived from the number of allocated resource blocks (RB). This can be obtained through uplink grant reception [41]. Similarly, the capacity of downlink radio link depends on the number of allocated RBs. This information is derived by means of downlink assignment as described in [41]. Apart from the number of allocated RBs, the capacity depends on MCS used for transmission based on SINR. From knowledge of the amount of bits to be transmitted to each computing $SCeNB_x$, and radio capacity of each SCeNB, we calculate the delay of radio

Algorithm 1 Selection of path for data delivery

$c_i^B, c_x^B, c_i^R, v_{bits}^{R,i}, v_{bits}^{B,i}, \gamma$

1: $D_q \leftarrow null$
2: $E_q \leftarrow null$
3: **for** $x \in X$ **do**
4: **for** $i \in I$ **do**
5: $D_R \leftarrow v_{bits}^{R,i}/c_i^R$
6: $d_x^i \leftarrow D_R$
7: $e_x^i \leftarrow E[D_R]$
8: **if** $v_{bits}^{B,i} > 0$ **then**
9: $D_B \leftarrow v_{bits}^{B,i}/c_i^B + v_{bits}^{B,i}/c_x^B$
10: $d_x^i \leftarrow d_x^i + D_B$
11: **end if**
12: **end for**
13: **end for**
14: **for** $q \in Q$ **do**
15: $D_q \leftarrow \sum_q d_x^q$
16: $E_q \leftarrow \sum_q e_x^q$
17: **if** handover **then**
18: $D_q \leftarrow D_q + D_H$
19: $E_q \leftarrow E_q + E_H$
20: **end if**
21: **end for**
22: **for** $q \in Q$ **do**
23: $\overline{D_q} \leftarrow D_q/max\{D_q\}$
24: $\overline{E_q} \leftarrow E_q/max\{E_q\}$
25: $M_q \leftarrow \gamma\overline{E_q} + (1-\gamma)\overline{D_q}$
26: **end for**
27: $q'_{s'} \leftarrow argmin\{M_q\}$
28: **return** $q'_{s'}$

transmission using (6). The capacity of backhaul connection is calculated based on known maximum backhaul capacity and link utilization. Required parameters for calculation of energy consumption are measured directly by the UE or obtained via control channels from the SCeNBs.

To determine the most suitable paths, it is necessary to identify cells, which are in communication range of the UE. This can be done according to the SINR. In common cellular networks, such as LTE-A, the UE can monitor SINR from the SCeNBs included in Neighbor Cell List (NCL) (for more details about NCL, refer to [42]). The NCL contains all potential neighbors of the UE's serving cell. Thus, this corresponds to the list of the SCeNBs, which might be available for data transmission.

Each SCeNB can be switched off at any time since the SCeNBs can be deployed also by users (e.g., femtocells) [43]. Thus, a secondary path should be defined for a case when the primary path is no longer available due to its failure. To keep routing overhead low in case of the link failure, data to be sent over this link will be rerouted through the original serving SCeNB selected according to signal quality, if possible. In case of the serving SCeNB failure, the UE will reinitiate path selection as there is a major change in state of links and there is no other backup route. Note that this problem requires also selection of a new serving cell for communication purposes. However, this is a common problem, for which existing mobile networks are able to find a solution by selection of new serving cell according to RSSI (for more details, see [44]).

3.5. Complexity of the path selection algorithm

Complexity of the proposed path selection algorithm is proportional to the number of computing SCeNBs (n) and the number of SCeNBs in radio communication range of the UE (m). The number of possible paths can be computed as partial permutation. Thus, the complexity of algorithm is $O(m^n)$. This complexity might be redundant as many SCeNBs in communication range of the UE provide radio channel of quality not suitable to satisfy requirements on delay imposed by a service (D_{req}). Hence, we narrow-down former set of all SCeNBs in communication range of the UE (I) to the set Y with a size of m_y, consisting of the SCeNBs with SINR above a threshold ρ_{SINR}. The set Y is created to cut off unusable SCeNBs with very low channel quality. The cut, defining the set Y, has no negative impact on performance of the proposed algorithm, as the cut removes only the SCeNBs, which cannot be used for communication due to low channel quality. This means that we set the SINR threshold ρ_{SINR} equal to a minimum SINR when devices can communicate and thus, SCeNBs with $SINR_{SCeNB} < \rho_{SINR}$ are not considered for the path selection. note that the size of set Y can be controlled by setting the threshold ρ_{SINR}. However, high ρ_{SINR} could lead to performance degradation as some base stations in proximity of UE, which potentially available for data transmission, might be excluded from the set Y disregarding the amount of available radio resources and backhaul. Anyway, considering low radius of small cells, wall attenuation and interference, the number of small cells in radio communication range (i.e., $SINR_{SCeNB} > \rho_{SINR}$) is very low (in our simulations, typically between two and four). Thus, complexity of the proposed solution after removing unsuitable base stations form set Y is also kept at low level.

Consequently, the set Y includes only SCeNBs, which can provide SINR high enough to satisfy D_{req}, i.e., the set Y is defined as:

$$Y = \{y \mid y \in I, y > \rho_{SINR}\} \tag{16}$$

By this approach, the list of SCeNBs in UE's proximity is reduced from a size of m to m_y. Consequently, the complexity of the proposed path selection algorithm is reduced from $O(m^n)$ to $O({m_y}^n)$, where $m > m_y$. Note that this leads to replacement of the set I by the set Y in Algorithm 1 in Step 4.

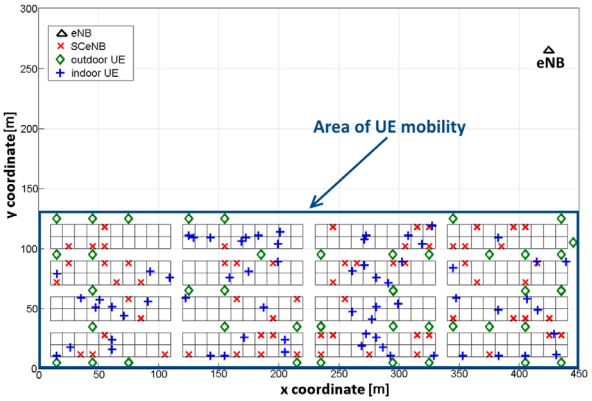

Fig. 4. Simulation scenario with example of deployment of buildings, users, SCeNBs and eNB for simulations.

4. Evaluation methodology and scenario

In this section, scenarios, deployment and simulation models used in MATLAB simulations for performance evaluations are presented. The simulation area is composed of two-stripes of buildings (as shown in Fig. 4) as suggested by 3GPP in [45]. The size of each building's block is 20 x 100 m and blocks are separated by streets with a width of 10 m. The overall simulation area is composed of 4 x 4 blocks of offices or apartments. The size of the whole simulated area is 430 x 270 m. Fifty outdoor UEs are randomly deployed at the beginning of the simulation and they move along the streets according to Manhattan Mobility model [46], with

Table 2
Simulation parameters.

Parameter	Value
Simulation area	430 × 270 m
Carrier frequency	2 000 MHz
Bandwidth for downlink/uplink	20/10 MHz
Tx power of eNB/SCeNB	43/23 dB
Attenuation of external/internal/separating walls	20/3/7 dB
SCeNB deployment ratio	0.2
Shadowing factor	6 dB
Handover interruption duration	30 ms
Number of Indoor UEs/Outdoor UEs/SCeNBs	64/50/64
Speed of outdoor users	1 m/s
Traffic generated by one request	300 kB/30 MB
Time between two requests for 300 kB/30 MB tasks	64/512 s
Simulation time	20 000 s
Number of simulation drops	4

a movement speed of 1 m/s. In addition, also indoor UEs are randomly deployed in offices with 20% offices occupied with one UE, i.e., there are 64 indoor UEs. Movement of the indoor UEs is modeled so that the UEs move within the apartments at discrete positions with a specific time distributions as defined in [47]. Inside the buildings, also the SCeNBs are randomly dropped to the offices with equal probability in a way that 20% of offices are equipped with a SCeNB. Therefore, 64 SCeNBs are deployed indoor. Besides the SCeNBs, also a macrocell (eNB) is placed outside the block of buildings at coordinates of [425 m, 265 m] (see Fig. 4).

The offloaded tasks is computed at 1, 2, 3 or 4 SCeNBs, with equal probability of each option. One of the computing SCeNBs is the serving one if this one can offer enough computing resources as suggested in [6]. In simulations, the computing SCeNBs are selected as a random set of n closest available SCeNBs.

We assume the size of offloaded task is either 300 kB or 30 MB to represent two different loads corresponding to different types of applications [48]. Interval between two offloaded tasks is set to 64 s and 512 s for a size of tasks of 300 kB and 30 MB, respectively. This interval is selected to generate enough traffic so that one request transfer affects selection of path for another.

If two or more offloaded tasks are generated at the same moment, the path selection is done sequentially for each task. Therefore, the impact of the path selection for the first offloaded task is subsequently considered for the second offloaded task and so on.

Major parameters of the simulation, summarized in Table 2, are in line with the recommendations for networks with small cells as defined by 3GPP in [45]. We also follow parameters of the physical layer frame structure for LTE-A mobile networks and signal propagation as defined in the same document. Based on [13] and [49], we set handover delay to be 30 ms. In LTE-A, Orthogonal Frequency-Division Multiple Access (OFDMA) is used for communication at the physical layer in downlink whereas Single-Carrier Frequency-Division Multiple Access (SC-FDMA) is used in uplink. The smallest unit to be allocated to the UE is a resource block (RB), which consists of 12 subcarriers and 7 symbols. Downlink and uplink are separated by means of Frequency Division Duplex (FDD).

Radio and backhaul resources are shared among the UEs in such manner that newly incoming request can be assigned with up to half of available resources to guarantee resource availability also for other potential UEs and services. A part of radio link capacity is assumed to be consumed by the background traffic (common voice and data services exploited by other users). Thus, the maximum number of available RBs per subframe for uplink and downlink in our simulations is 40 and 80, respectively.

We consider SCeNBs connected to the operator's network through either DSL or optical fiber. Maximum throughput of both is generated by a normal distribution with mean value μ and

Table 3
Parameters of backhaul models.

Optical fiber μ (uplink/downlink)	100/100 Mbit/s
Optical fiber σ (uplink/downlink)	11.5/11.5
DSL μ (uplink/downlink)	1/5.5 Mbit/s
DSL σ (uplink/downlink)	100 Mbit/s
eNB uplink/downlink	1000/1000 Mbit/s

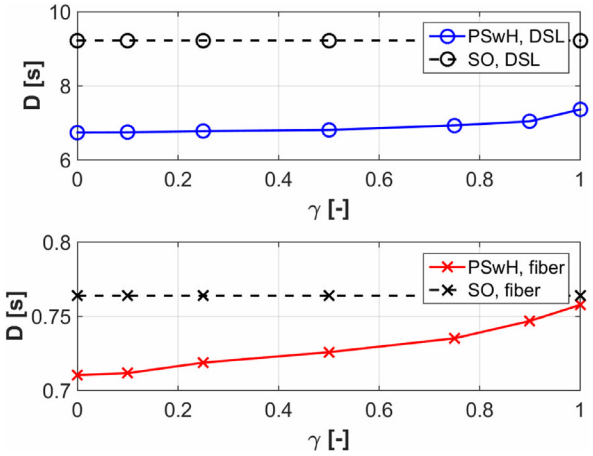

Fig. 5. Average time (D) required for transmission of offloaded task with size of 300 kB for DSL backhaul (top subplot) and fiber optic (bottom subplot) backhauls.

standard deviation σ as specified in Table 3. The optical fiber is used solely for the corporate scenario which assumes several cells within one building [47] sharing the same backhaul. All cells belonging to the same corporate building are interconnected with Local Area Network (LAN) offering throughput of 100 Mbit/s among the SCeNBs. If two or more SCeNBs within the same corporate building communicate with each other, we assume direct communication between the SCeNBs via LAN. Consequently, no part of the offloaded task is distributed to the core network over optical fiber. The DSL backhaul connection corresponds to the residential scenario where the SCeNBs are deployed in private flats and connected to the core network [47]. For both scenarios, the UE can communicate with the computing SCeNBs also via eNB. The eNB is connected to operator's network through a link with a throughput of 1000 Mbit/s.

5. Simulation results

In this section, simulation results are presented and discussed. The performance is evaluated for the proposed PSwH algorithm and also for commonly adopted SO approach [6,8,11,24] (only the serving SCeNB selection based on RSSI level is assumed). Simulation results are divided into subsections analyzing : 1) transmission delay, 2) energy consumed by the UE, 3) satisfaction of users with experienced delay, 4) load of the SCeNB's backhaul, and 5) number of additional handovers generated by the PSwH. We also discuss selection of proper values of weighting parameter γ and we summarize major findings from simulations in this section.

5.1. Delay of UE data transmission/reception

Impact of the PSwH algorithm on the average delay caused by transmission of the offloaded task between the UE and the computing SCeNBs is depicted in Fig. 5 and Fig. 6. From both figures, we can observe that delay increases with γ. This is because high γ indicates priority for low energy consumption while delay becomes less important (see (1)). The proposed PSwH reaches lower delay comparing to the SO for all values of γ. Low delay achieved by the

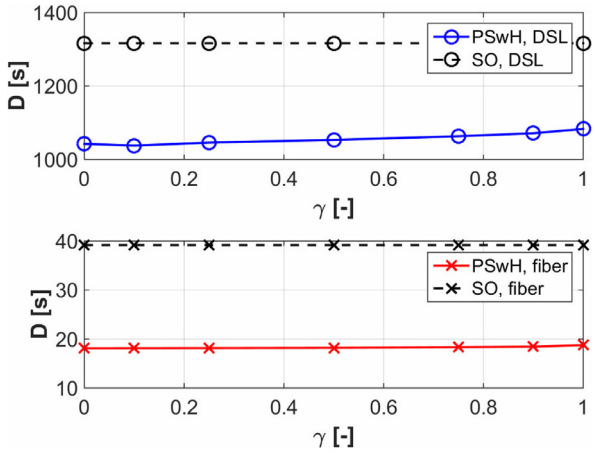

Fig. 6. Average time (D) required for transmission of offloaded task with size of 30 MB for DSL backhaul (top subplot) and fiber optic (bottom subplot) backhauls.

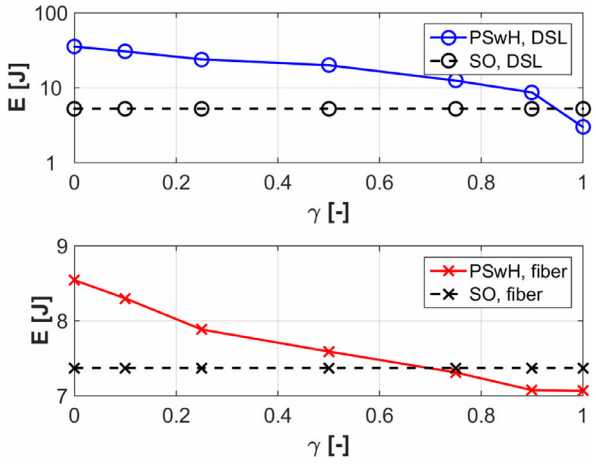

Fig. 7. Average energy (E) required for transmission of offloaded task with size of 300 kB for DSL (top subplot) and fiber optic (bottom subplot) backhauls.

PSwH results from avoiding low quality backhaul if it is possible to transmit data directly to the computing SCeNBs or through different SCeNBs with less loaded backhaul. For the offloaded task with a size of 300 kB, the average transmission delay D is reduced by up to 26.9% for DSL backhaul and up to 7% for optical fiber backhaul, as shown in Fig. 5. For the offloaded task's size of 30 MB, the PSwH shortens the delay by up to 21.5% comparing to the SO in case of DSL backhaul and by up to 53.7% for the optical fiber backhaul as shown in Fig. 6.

5.2. Energy consumed by UE for data transmission/reception

The proposed PSwH should avoid draining of the UE's battery caused by data transmission/reception and handover. By increasing γ, radio paths with lower energy consumption are used more often and the energy consumption is decreasing (see Figs. 7 and 8). Energy required for the UE's transmission depends on transmission power level and transmission duration as specified in (11). Lower energy consumption comparing to the SO is achieved by shortening the transmission time resulting from performing handover to less loaded SCeNBs (offering more available RBs for transmission). The reason for lower energy consumption for transmission via less loaded SCeNB is that a linear increase in the number of consumed RBs leads to a linear decrease in the transmission delay while increase in the energy consumption is logarithmic (see (10)). Another reason for lowering the energy consumption by the PSwH is the

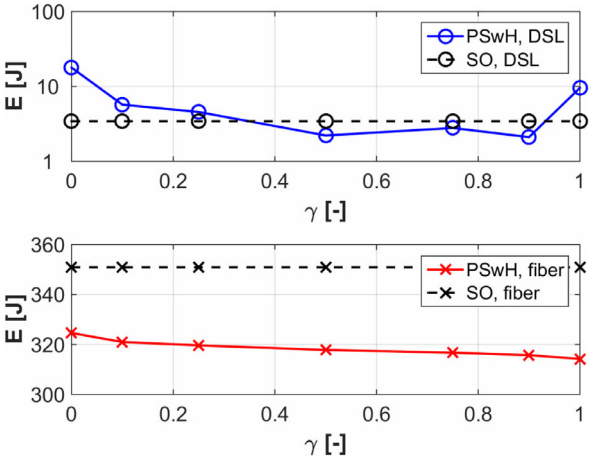

Fig. 8. Average energy (E) required for transmission of offloaded task with size of 30 MB for DSL (top subplot) and fiber optic (bottom subplot) backhauls.

usage of the connection with a more robust MCS, which requires lower transmission power (see [17] for more details).

From Figs. 7 and 8, we can see that the PSwH reduces energy consumption by up to 3.2% in case of the DSL backhaul and by up to 4.1% for the optical fiber backhaul if the offloaded task is of 300 kB as shown in Fig. 7. For the offloaded task of 30 MB (shown in Fig. 8), the PSwH lowers energy consumption by up to 4.1% for DSL backhaul and by up to 10.4% for the optical fiber. The energy consumption can be increased comparing to the SO if the users do not care about energy (low γ). However, in this case, the users indicate their preference for the delay so they are not unhappy with increased energy consumption.

There is one singular point when impact of γ on the energy is unexpected (energy rises with γ). This situation is shown in Fig. 8 for large offloaded tasks (30 MB) and low quality backhaul (DSL). In this scenario, the algorithm is trying to minimize energy consumption by selection of the most appropriate radio path disregarding delay (see (1)). Hence, the algorithm tends to associate all UEs to the SCeNBs with radio links requiring the lowest energy consumption. However, for the UEs trying to offload data later when other transmissions are already in progress, not enough radio resources are available. Consequently, those UEs are associated to the SCeNBs, which may lead to even higher energy consumption than in case of the SO.

5.3. Satisfaction of users with experienced transmission delay

The satisfaction of UEs with experienced transmission delay D with respect to their required delay D_{req} is shown in Fig. 9 and Fig. 10 for DSL and optical fiber backhauls, respectively. The satisfaction is understood as a ratio of users (R_s), who experience delay lower than the requested one (i.e., $D \leq D_{req}$). The satisfaction increases as γ decreases since lowering delay becomes of higher priority than energy consumption. As can be seen from Figs. 9 and 10, the UEs' satisfaction is increasing with D_{req} for both compared algorithms. This fact is expected as more time is available for delivery of data for higher D_{req}. Comparing the PSwH with the SO for DSL backhaul, the proposed algorithm increases the satisfaction up to 10% for the offloaded task with a size of 300 kB (Fig. 9a) and up to 15% for the offloaded task with a size of 30 MB (Fig. 9b).

For the optical fiber backhaul, the satisfaction of UEs with experienced transmission delay is shown in Fig. 10. The PSwH improves the satisfaction by up to 7% for the offloaded task of 300 kB (Fig. 10a), and up to 29% for the offloaded task of 30 MB (Fig. 10b).

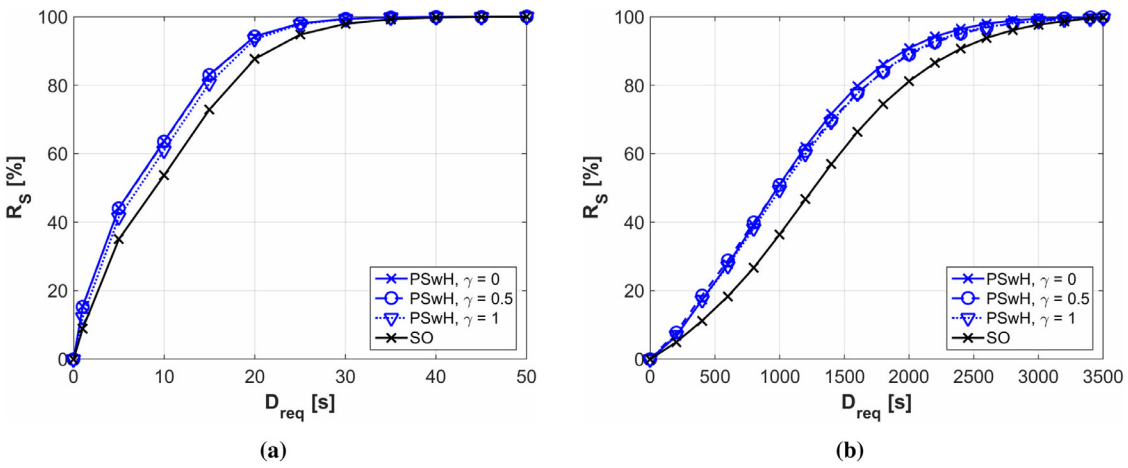

Fig. 9. Ratio of users satisfied with experienced delay, R_S, for DSL backhaul for offloaded task of 300 kB (a) and 30 MB (b).

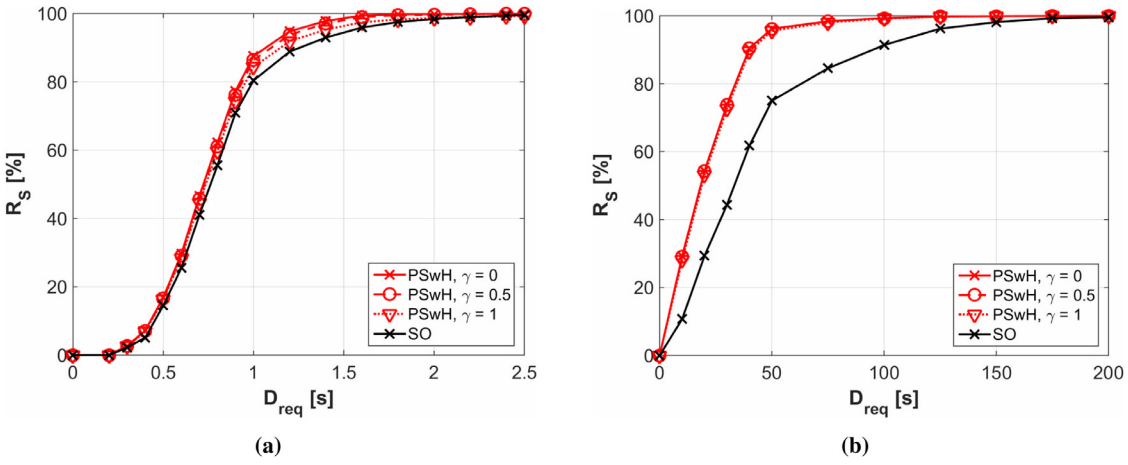

Fig. 10. Ratio of users satisfied with experienced delay, R_S, for optical fiber backhaul, for offloaded task of 300 kB (a) and 30 MB (b).

5.4. Load of small cell's backhaul

The proposed algorithm takes advantage of handovers to speed up data delivery and also to offload the backhaul of SCeNBs. The load of backhaul (μ_T) is represented by a mean number of the offloaded tasks transmitted per backhaul link and time. For the PSwH, the backhaul load increases with γ since a priority is given to lowering the UE's energy consumption while backhaul capacity (represented by transmission delay) is of a lower priority. This behavior results from the lowering energy consumption (high γ), which leads to selection of less energy consuming radio link even if backhaul has to be used. Contrary, the most of the traffic is transmitted directly to the computing SCeNB in radio communication range if the users prefer low delay (low γ).

In Fig. 11a, we can see that the PSwH reduces the DSL backhaul load by up to 32% comparing to the SO for both uplink and downlink for the offloaded task of 300 kB. For the offloaded task of 30 MB, more than 9% decrease in DSL backhaul utilization is observed as well for both directions as shown in Fig. 11b. The decrease in backhaul load by the PSwH is due to exploitation of the radio link rather than low quality backhauls.

In case of the optical fiber, the PSwH lowers the backhaul load by up to 11% for the offloaded task with a size of 300 kB and by up to 15.5% for the offloaded task with a size of 30 MB, as shown in Figs. 12a and 12b, respectively.

5.5. Number of performed handovers caused by the proposed algorithm

The ratio of additional handovers introduced by the PSwH for the transmission of offloaded task (R_H) is shown in Fig. 13. If the PSwH is used, the number of handovers for delivery of the offloaded task of 300 kB is increased by 53–56% for the optical fiber backhaul and by 53–55% for the DSL backhaul (see Fig. 13a). If the size of offloaded task is 30 MB (Fig. 13b), the number of handovers is increased by 5% for DSL backhaul and by 12.5% for optical fiber backhaul. For both backhauls, the number of handovers increases with γ for low values of γ and then decreases for high values of γ. This behavior is a result of combination of handovers initiated for minimization of the energy consumption as well as for minimization of the delay for $0 < \gamma < 1$. For $\gamma = 0$ or $\gamma = 1$, the handover is initiated less often as only either energy consumption or delay are targeted. Note that the impact of γ on the number of additional handovers is very low (below 3.5%).

Less significant increase in the number of handovers for large offloaded tasks (30 MB) is caused by more time required for transmission of such task. Therefore, the radio links of the SCeNBs in communication range of the UE (included in set I) are heavily loaded for a longer period of time. Consequently, allocation of resources at the overloaded neighboring SCeNBs for the users associated to another cell is not feasible.

The proposed algorithm introduces additional handovers, which can lead to redundant signaling and interruption in communica-

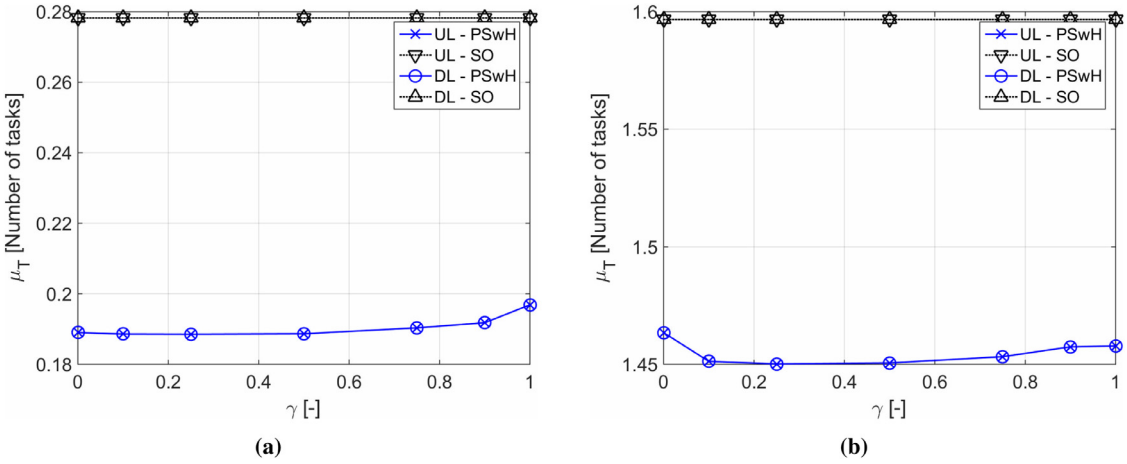

Fig. 11. Mean number of the offloaded tasks, μ_T, transmitted over DSL backhaul for the offloaded task size of 300 kB (a) and 30 MB (b).

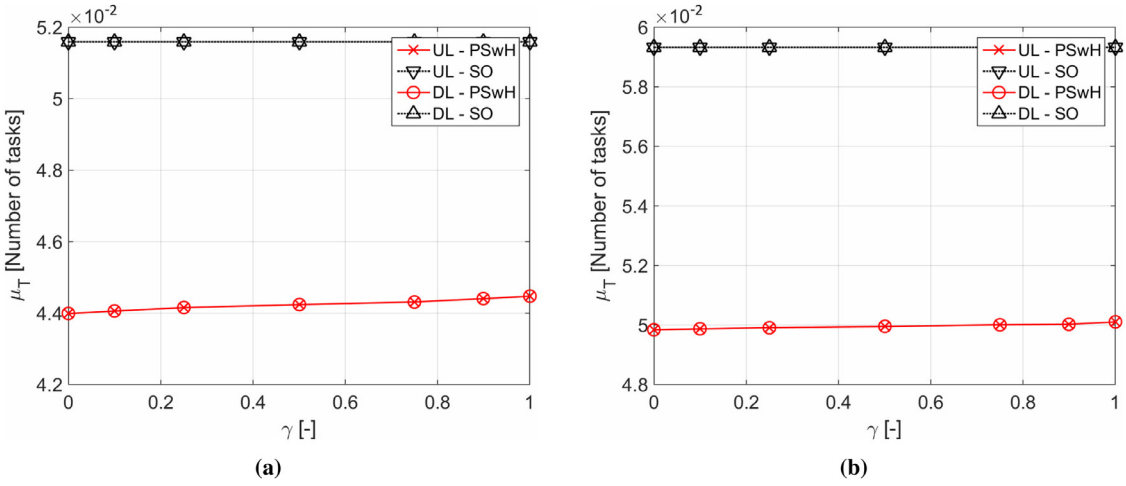

Fig. 12. Mean number of the offloaded tasks, μ_T, transmitted over optical fiber backhaul for the offloaded task size of 300 kB (a) and 30 MB (b).

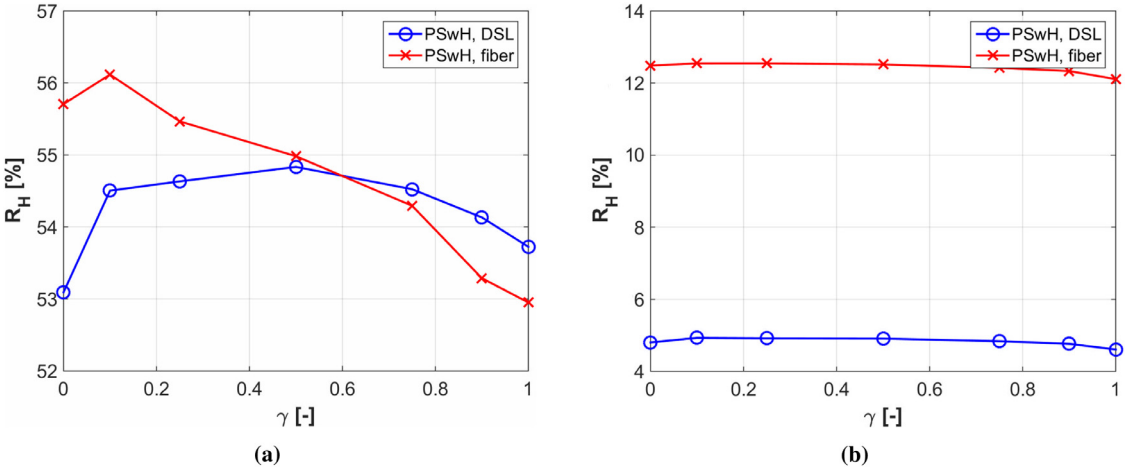

Fig. 13. Ratio of additional handovers generated by the PSwH algorithm, R_H, with respect to the SO for offloaded tasks of 300 kB (a) and 30 MB (b).

tion due to performing handover (known as handover interruption). The signaling overhead generated per handover is in order of kb [40]. The overall number of handovers per one offloaded task is very low (roughly 0.8 in average). Hence, total handover overhead is in order of kb per offloaded task and can be considered negligible. The second problem, handover interruption, is not related to the SCC services as the users do not care about interruption in transmission of the offloaded task if the computation results are delivered within desired delay D_{req}. The handover interruption is considered in the PSwH algorithm (see (4)). Thus, all above-presented results already include impact of the handover interruption. Of course, the handover interruption introduced by the PSwH can degrade quality of conventional services (voice, video, etc.) running at the UE simultaneously with offloaded SCC services.

Table 4
Summarized improvement (green color) in performance metrics introduced by the PSwH comparing to the SO for proper values of γ.

Backhaul type	Size of task	Proper γ	ΔD [%]	ΔE [%]	$\Delta \mu_T$ [%]	R_H [%]
DSL	300 kB	1	-20.2	-3.2	-29.2	+53.5
DSL	30 MB	0.5	-19.1	-0.9	-9.1	+5.1
optical fiber	300 kB	0.75	-3.8	-0.8	-14.3	+54.3
optical fiber	30 MB	0	-54.3	-7.5	-16	+12.4

In case of simultaneous usage of the SCC offloading service and common real-time service, the user must indicate priority for one type of services. If the preference is given to the SCC service, the user should not be disappointed with lower quality of the secondary service. Contrary, if the preference is given to the conventional real-time (non-SCC) service, the SCC service will be handled in conventional way (i.e., by means of SO algorithm) with no gain in delay or energy consumption but also with no degradation in QoS for the non-SCC service. Note that the SCC is intended mainly for delay sensitive and real-time services (applications). Therefore, simultaneous usage of the SCC service and common non-SCC service is not very likely.

5.6. Discussion of results and selection of proper γ

In this section, proper selection of γ for the proposed algorithm is discussed along with gain in above-mentioned performance metrics introduced by the PSwH.

The proper γ is selected in such a way that the delay reduction comparing to the SO is maximal while energy consumption is still lowered or at least not impaired comparing to the SO. The selected values of γ are shown in Table 4. The proper value of γ spans over the whole range (i.e., from 0 to 1) and individual proper value depends on combination of backhaul quality and a size of the offloaded task. Consequently, also the gain (Δ) introduced by the PSwH comparing to the SO varies for backhaul types and a size of the offloaded tasks. The gain Δ is defined as improvement introduced by the PSwH with respect to the SO for each performance metric. For example, the gain in delay is defined as $\Delta D = (D_{PSwH} - D_{SO})/D_{SO}$. Therefore, the negative numbers (green color) in this table represent improvement introduced by the PSwH comparing to the SO (e.g., the PSwH reduces delay by 20.2%) while the positive numbers (red color) indicate worsened performance (additional handovers introduced by the PSwH).

The variation of gain for different backhauls and offloaded task size is caused by availability of each backhaul for data transmission and its ability to handle given level of load introduced by the offloading tasks. For high quality optical fiber backhaul, the small tasks (300 kB) can be handled even by the SO algorithm as the optic is able to distribute such small amount of data easily. Thus, to reach a gain by the PSwH, a high number of handovers must be performed to find more suitable way of data distribution. However, for other scenarios optical fiber with large tasks or DSL with both sizes of the tasks), the SO fails in distribution of the tasks over backhaul, which can be easily overloaded by the offloaded tasks. Consequently, the gains introduced by the PSwH become more significant.

6. Conclusion

In this paper, we have proposed a new path selection algorithm for delivery of the offloaded tasks between the UE and the cloud-enhanced small cells. The algorithm forces the UE to perform handover if it is efficient in terms of the overall transmission delay (considering radio and backhaul) and/or energy consumption of the UE. In order to find a trade-off between transmission delay and energy efficiency, weighting of both metrics is introduced. The proposed algorithm reduces the transmission delay by up to 20.2% and 54.3% in scenario with small cells connected to the operator's network by the DSL backhaul and optical fiber, respectively. At the same time, the energy consumption of the UE can be lowered by 3.2% and by 7.5% for DSL and optical fiber backhauls, respectively. Notice that the improvement accomplished by the PSwH depends on the size of offloaded task together with used backhaul connection. The proposed algorithm also increases user's satisfaction with experienced delay (up to 29%) and lowers backhaul load (up to 32%). The improvements reached by the proposed algorithm are at the cost of additional handovers. Nevertheless, delay introduced by these additional handovers is already considered in the path selection algorithm. Therefore, the handovers do not decrease QoS but leads only to negligible additional overhead (few kb per offloaded task).

As the algorithm can select efficient path for downlink and uplink independently, it is suitable also for mobility management of moving user's exploiting the SCC services.

In the future, the proposed algorithm can be extended to scenario with over-the-air communication and the combination of the PSwH with possible migration of the VMs among the SCeNBs in order to shorten the delay.

Acknowledgment

This work has been performed in the framework of the FP7 project TROPIC IST-318784 STP, which is funded by the European Community. The Authors would like to acknowledge the contributions of their colleagues from TROPIC Consortium (http://www.ict-tropic.eu).

References

[1] S. Carlaw, Connected World of Tomorrow, Predictions for 2014 and 2015, Technical Report, ABI research, 2014.
[2] W. Li, Y. Zhao, S. Lu, D. Chen, Mechanisms and challenges on mobility-augmented service provisioning for mobile cloud computing, Commun. Mag. IEEE 53 (3) (2015) 89–97.
[3] S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, R. Buyya, Cloud-based augmentation for mobile devices: motivation, taxonomies, and open challenges, Commun. Surv. Tut. IEEE 16 (1) (2014) 337–368.
[4] M. Barbera, S. Kosta, A. Mei, J. Stefa, To offload or not to offload? the bandwidth and energy costs of mobile cloud computing, in: INFOCOM, 2013 Proceedings IEEE, 2013, pp. 1285–1293, doi:10.1109/INFCOM.2013.6566921.
[5] B. Addis, D. Ardagna, A. Capone, G. Carello, Energy-aware joint management of networks and cloud infrastructures, Comput. Netw. 70 (0) (2014) 75–95. http://dx.doi.org/10.1016/j.comnet.2014.04.011.
[6] F. Lobillo, Z. Becvar, M. Puente, P. Mach, F. Lo Presti, F. Gambetti, M. Goldhamer, J. Vidal, A. Widiawan, E. Calvanesse, An architecture for mobile computation offloading on cloud-enabled lte small cells, in: Wireless Communications and Networking Conference Workshops (WCNCW), 2014 IEEE, 2014, pp. 1–6, doi:10.1109/WCNCW.2014.6934851.
[7] M. Molina, O.M. Medina, A.P. Iserte, J. Vidal, Joint scheduling of communication and computation resources in multiuser wireless application offloading, in: IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2014, pp. 1093–1098.
[8] S. Barbarossa, S. Sardellitti, P. Di Lorenzo, Communicating while computing: distributed mobile cloud computing over 5g heterogeneous networks, Sig. Process. Mag. IEEE 31 (6) (2014) 45–55, doi:10.1109/MSP.2014.2334709.
[9] M. Puente, Z. Becvar, M. Rohlik, F. Lobillo, E. Calvanese-Strinati, A seamless integration of computationally-enhanced base stations into mobile networks towards 5g, in: Vehicular Technology Conference (VTC Spring) Workshop on 5G Architecture, 2015 IEEE, 2015.

[10] R. Sharma, S. Kumar, M. Trivedi, Mobile cloud computing: a needed shift from cloud to mobile cloud, in: Computational Intelligence and Communication Networks (CICN), 2013 5th International Conference on, 2013, pp. 536–539, doi:10.1109/CICN.2013.116.

[11] V. Di Valerio, F. Lo Presti, Optimal virtual machines allocation in mobile femtocloud computing: an mdp approach, in: Wireless Communications and Networking Conference Workshops (WCNCW), 2014 IEEE, 2014, pp. 7–11, doi:10.1109/WCNCW.2014.6934852.

[12] S. Barbarossa, S. Sardellitti, P. Di Lorenzo, Computation offloading for mobile cloud computing based on wide cross-layer optimization, in: Future Network and Mobile Summit (FutureNetworkSummit), 2013, 2013, pp. 1–10.

[13] 3GPP, Mobility enhancements in heterogeneous networks, TR, 36.839, 3rd Generation Partnership Project (3GPP), 2013.

[14] D. Xenakis, N. Passas, L. Merakos, C. Verikoukis, Mobility management for femtocells in lte-advanced: key aspects and survey of handover decision algorithms, Commun. Surv. Tut. IEEE 16 (1) (2014) 64–91, doi:10.1109/SURV.2013.060313.00152.

[15] Z. Becvar, P. Roux, P. Mach, Fast cell selection with efficient active set management in ofdma networks with femtocells, EURASIP J. Wireless Commun. Netw. 2012 (1) (2012), doi:10.1186/1687-1499-2012-292.

[16] D. Xenakis, N. Passas, C. Verikoukis, An energy-centric handover decision algorithm for the integrated LTE macrocell-femtocell network, Comput. Commun. 35 (14) (2012) 1684–1694. Special issue: Wireless Green Communications and Networking. http://dx.doi.org/10.1016/j.comcom.2012.04.024.

[17] Z. Becvar, J. Plachy, P. Mach, Path selection using handover in mobile networks with cloud-enabled small cells, in: IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2014, pp. 1480–1485.

[18] J. Al-Karaki, A. Kamal, Routing techniques in wireless sensor networks: a survey, Wireless Commun. IEEE 11 (6) (2004) 6–28, doi:10.1109/MWC.2004.1368893.

[19] S. Mueller, R. Tsang, D. Ghosal, Multipath routing in mobile ad hoc networks: Issues and challenges, Performance Tools and Applications to Networked Systems, 2004.

[20] M.K. Marina, S.R. Das, Ad hoc on-demand multipath distance vector routing, Wireless Commun. Mob. Comput. 6 (7) (2006) 969–988, doi:10.1002/wcm.432.

[21] S. Kumar, S. Khimsara, K. Kambhatla, K. Girivanesh, J.D. Matyjas, M. Medley, Robust on-demand multipath routing with dynamic path upgrade for delay-sensitive data over ad hoc networks, J. Comput. Netw. Commun. 2013 (2013), doi:10.1155/2013/791097.

[22] S. Othmen, A. Belghith, F. Zarai, M. Obaidat, L. Kamoun, Power and delay-aware multi-path routing protocol for ad hoc networks, in: Computer, Information and Telecommunication Systems (CITS), 2014 International Conference on, 2014, pp. 1–6, doi:10.1109/CITS.2014.6878956.

[23] H. Wu, Q. Wang, K. Wolter, Methods of cloud-path selection for offloading in mobile cloud computing systems, in: Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on, 2012, pp. 443–448, doi:10.1109/CloudCom.2012.6427587.

[24] J. Oueis, E. Calvanese-Strinati, S. Barbarossa, Small cell clustering for efficient distributed cloud computings, in: IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2014, pp. 1480–1485.

[25] S. Wang, R. Urgaonkar, T. He, M. Zafer, K. Chan, K. Leung, Mobility-induced service migration in mobile micro-clouds, in: Military Communications Conference (MILCOM), 2014 IEEE, 2014, pp. 835–840, doi:10.1109/MILCOM.2014.145.

[26] S. Wang, R. Urgaonkar, K. Chan, T. He, M. Zafer, K.K. Leung, Dynamic service placement for mobile micro-clouds with predicted future costs, in: 2015 IEEE International Conference on Communications (ICC), 2015, pp. 5504–5510, doi:10.1109/ICC.2015.7249199.

[27] H. Maziku, S. Shetty, Network aware vm migration in cloud data centers, in: IEEE Third GENI Research and Educational Experiment Workshop (GREE), 2014, IEEE, 2014, pp. 25–28.

[28] Z. Becvar, P. Mach, M. Vondra, Handover Procedure in Femtocells, in: R.A. Saeed (Ed.), Chapter in Femtocell Communications and Technologies: Business Opportunities and Deployment Challenges, Information Science Reference, 2012.

[29] 3GPP, Handover Procedures, TS, 23.009, 3rd Generation Partnership Project (3GPP), 2014.

[30] J. Oueis, E. Calvanese-Strinati, S. Barbarossa, Multi-parameter decision algorithm for mobile computation offloading, in: Wireless Communications and Networking Conference (WCNC), 2014 IEEE, 2014, pp. 3005–3010, doi:10.1109/WCNC.2014.6952959.

[31] O. Munoz, A. Pascual Iserte, J. Vidal, M. Molina, Energy-latency trade-off for multiuser wireless computation offloading, in: Wireless Communications and Networking Conference Workshops (WCNCW), 2014 IEEE, 2014, pp. 29–33, doi:10.1109/WCNCW.2014.6934856.

[32] O. Munoz-Medina, A. Pascual-Iserte, J. Vidal, Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading, Veh. Technol. IEEE Trans. PP (99) (2014). 1–1, 10.1109/TVT.2014.2372852.

[33] G. Bolch, S. Greiner, H. de Meer, K.S. Trivedi, Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications, Wiley-Interscience, New York, NY, USA, 1998.

[34] 3GPP, Policy and Charging Control (PCC); Reference points, TS, 29.212, 3rd Generation Partnership Project (3GPP), 2015.

[35] 3GPP, layer 2 - measurements, TS, 36.314, 3rd Generation Partnership Project (3GPP), 2014.

[36] S. Ahmadi, LTE-Advanced: A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies, 1st, Academic Press, 2013.

[37] M. Lauridsen, A. Jensen, P. Mogensen, Reducing lte uplink transmission energy by allocating resources, in: Vehicular Technology Conference (VTC Fall), 2011 IEEE, 2011, pp. 1–5, doi:10.1109/VETECF.2011.6092935.

[38] M. Lauridsen, L. Noël, T.B. Sørensen, P. Mogensen, An empirical lte smartphone power model with a view to energy efficiency evolution, Intel Technol. J. 18 (1) (2014) 172–193.

[39] E. Tejaswi, S. B, Survey of power control schemes for lte uplink, Int. J. Comput. Sci. Inf. Technol. (IJCSIT) (2013).

[40] 3GPP, User Equipment (UE) radio transmission and reception, TS, 36.101, 3rd Generation Partnership Project (3GPP), 2015.

[41] 3GPP, Medium Access Control (MAC) protocol specification, TS, 36.321, 3rd Generation Partnership Project (3GPP), 2015.

[42] 3GPP, Radio Resource Control (RRC), Protocol specification, TS, 36.331, 3rd Generation Partnership Project (3GPP), 2015c.

[43] J. Andrews, H. Claussen, M. Dohler, S. Rangan, M. Reed, Femtocells: Past, present, and future, Select. Areas Commun. IEEE J. 30 (3) (2012) 497–508, doi:10.1109/JSAC.2012.120401.

[44] 3GPP, Telecommunication management; Self-Organizing Networks (SON); Self-healing concepts and requirements, TS, 32.541, 3rd Generation Partnership Project (3GPP), 2014.

[45] 3GPP, Further Advancements for E-UTRA Physical Layer Aspects, TS, 36.814, 3rd Generation Partnership Project (3GPP), 2010.

[46] N. Meghanathan, Mobility models for wireless ad hoc networks, Presented at the REU 2010, Jackson State University, 2010.

[47] T. project deliverable, Deliverable D21, Scenarios and requirements, Confidential Deliverable, Distributed computing, storage and radio resource allocation over cooperative femtocells, 2014. www.ict-tropic.eu.

[48] K. Ha, P. Pillai, W. Richter, Y. Abe, M. Satyanarayanan, Just-in-time provisioning for cyber foraging, in: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, in: MobiSys '13, ACM, New York, NY, USA, 2013, pp. 153–166, doi:10.1145/2462456.2464451.

[49] M.P. Wylie-Green, T. Svensson, Throughput, capacity, handover and latency performance in a 3gpp lte fdd field trial, in: IEEE Global Telecommunications Conference (GLOBECOM 2010), IEEE, 2010, pp. 1–6.

Jan Plachy received the BSc and MSc degree in telecommunication engineering from the Czech Technical University in Prague, Czech Republic in 2012 and 2014, respectively. Currently, he works towards the PhD degree at the Department of Telecommunication Engineering at the same university with a topic Big data in 5G mobile networks. He was on internships at CEA-Leti, France (2014), and EURECOM, France (2016). Participated in FP7 TROPIC project founded by European Commission, with a goal of designing path selection algorithm for distributed computing. His research interests cover optimization of radio resource management in future mobile networks and transfer of big data over mobile networks.

Zdenek Becvar received the MSc and PhD degree in telecommunication engineering from the Czech Technical University in Prague, Czech Republic in 2005 and 2010, respectively. Currently, he is Associate professor at the Department of Telecommunication Engineering at the same university. From 2006 to 2007, he joined Sitronics R&D centre in Prague focusing on speech quality in VoIP. Furthermore, he was involved in research activities of Vodafone R&D center at Czech Technical University in Prague in 2009. He was on internships at Budapest Politechnic, Hungary (2007), CEA-Leti, France (2013), and EURECOM, France (2016). Since 2007, he participates in national projects and project founded by European Commission. In 2013, he becomes representative of the Czech Technical University in Prague in ETSI and 3GPP standardization organizations. In 2015, he founded 5Gmobile research lab at CTU in Prague focusing on research towards 5G mobile networks. He is a member of more than 15 program committees at international conferences or workshops and he published 3 book chapters and more than 60 conference or journal papers. He works on development of solutions for future mobile networks (5G and beyond) with special focus on optimization of radio resource management, mobility support, device-to-device communication, self-optimization, power control, architecture of radio access network, and small cells.

Pavel Mach received his MSc and PhD degree in telecommunication engineering from Czech Technical University in Prague, Czech Republic in 2006 and 2010 respectively. During his study he joined research groups at Sintronics and Vodafone R&D centers focusing on wireless mobile technologies. He is a member of more than 15 program committees of international conferences. He has published more than 40 papers in international journals and conferences. He has been actively involved in several national and international projects. He participated in several projects founded by European Commission. His research interests include MAN/LAN networks based on relay architectures. He is dealing with aspects relating to radio resource management in emerging wireless technologies and focuses on cross layer optimization processes.