A Cooperative Reinforcement Learning Approach for Inter-Cell Interference Coordination in OFDMA Cellular Networks

Mariana Dirani, Zwi Altman

A Cooperative Reinforcement Learning Approach for Inter-Cell Interference Coordination in OFDMA Cellular Networks

Mariana Dirani* and Zwi Altman†

*Altran, 58 Boulevard Gouvion St Cyr - 75017 Paris, France

†Orange Labs, 38/40 rue du General Leclerc,92794 Issy-les-Moulineaux

***Abstract*—Inter-Cell Interference Coordination (ICIC) is commonly identified as a key radio resource management mechanism to enhance system performance of 4G networks. This paper addresses the problem of ICIC in the downlink of cellular OFDMA (LTE and WiMAX) systems in the context of Self-Organizing Networks (SON). The problem is posed as a cooperative Multi-Agent control problem. Each base station is an agent that dynamically changes power masks on a subset of its bandwidth to control interference it produces to its neighbouring cells. The agent learns the optimal coordinated power allocation strategy using information from its own and its neighbouring cells. A Fuzzy Inference System (FIS) is used to handle continuous state space defined by the input quality indicators to the controller performing the ICIC. The FIS is optimized using Reinforcement Learning (RL) with a Fuzzy Q-Learning (FQL) implementation. Simulation results illustrate the important performance gain brought about by the proposed ICIC scheme.**

***Index Terms*—Inter-Cell Interference Coordination, OFDMA, LTE, Reinforcement Learning, Fuzzy Q-Learning, SON.**

[1]

I. Introduction

Emerging standards for B3G and 4G networks have ambitious performance targets [1]) that aim at providing better experience for end users. Orthogonal Frequency Division Multiple Access (OFDMA) technology has been selected for LTE and 4G networks, namely LTE advanced and WiMAX 802.16m ([2], [3]). In the downlink, OFDMA allows assigning frequency subcarriers to users within each cell in an orthogonal manner. With the use of cyclical prefix insertion, intra-cellular interference can be eliminated. Time-frequency scheduling can be implemented allowing to efficiently combat impairments caused by frequency selective channels. Physical Resource Block (PRB) is the smallest time-frequency resource unit that can be allocated to a user. When the same PRB is allocated to two or more neighbouring cells, a collision or interference may occur which can degrade the Signal to Interference plus Noise Ratio (SINR) and Quality of Service (QoS) perceived by the user. The flexibility in assignment of resources in OFDMA allows designing mechanisms to combat interference that can improve spectral efficiency. Interference mitigation techniques including ICIC are considered among the new building blocks of 4G network technologies that will allow to achieve ambitious performance targets set by IMT-Advanced (International Mobile Telecommunications-Advanced) [4].

Different approaches to combat interference have been studied including frequency reuse schemes such as fractional reuse and soft reuse schemes [5]. Fractional reuse schemes are a special case where the reuse-1 is applied to users with good quality (close to the station) and reuse-3 (or higher reuse factors) - to users with poorer quality (close to the cell edge). When different power allocation for the mobile users is associated with different portions of the frequency bandwidth, the frequency reuse is called soft reuse scheme. The power allocation pattern is denoted as the power mask of the cell.

The purpose of this paper is to present a solution for the ICIC based on adaptive soft frequency reuse scheme. The ICIC is presented as a control process that maps system states into control actions. The time scale of the control process is of the order to tens of seconds. The proposed solution is scalable and can be applied to a general network configuration. A RL [6] framework is considered with a Multi-Agent Fuzzy Q-Learning implementation [7]. The basic ICIC control entity comprises a cell and its neighbors. The learning (exploration) phase is cooperative, namely information and utilities can be shared among the base stations. The control (exploitation) phase is fully distributed and the base stations perform local actions in a non-synchronized manner. A recent publication on adaptive soft frequency reuse scheme has been reported in [8]. This work includes the interesting case where information cannot be shared among neighbouring base stations. The model in [8] does not need a learning phase, and requires higher signalling load to operate with respect to the solution proposed in this paper.

To put this paper in perspective, we note that Packet Scheduling (PS) can alleviate as well the impact of interference on the system performance by taking advantage of the channel diversity. Several contributions have been reported on optimal subcarriers' allocation such as [9], that operate

[1]M. Dirani was with Orange Labs at Issy-les-Moulineaux.

on a time scale of the order of a millisecond and can be implemented within a packet scheduler. In a real network implementation, the ICIC solution will operate together with a packet scheduler. An example of the use of PS in the context of SON is described in [10]. The combination of the two mechanisms operating at different time scales will further enhance the system performance, however this is out of the scope of the present work.

The control approach considered here for the interference mitigation problem is a self-optimizing process which falls in the domain of SON. This topic is currently receiving much attention in both industry and academia [11], [12], and is included in new standards [13]. Throughout the paper, the terms eNB and base station can be used interchangeably. Similarly, the terms mobile and mobile user have the same meaning.

The paper is organized as follows: Section II presents the system model based on dynamic soft reuse scheme and describes the allocation strategy of PRBs. Section III models the ICIC control process as a Multi-Agent Reinforcement Learning problem which is solved using a FQL algorithm. The components and the algorithm of the FQL are presented. Section IV describes the simulation environment and provides numerical results of the proposed ICIC scheme. Section V concludes the paper.

II. System model

A. Model description

Consider an OFDMA network with base stations implementing ICIC in the downlink. The ICIC performs adaptive soft reuse-1 scheme, namely the total available bandwidth is reused in all the cells while the transmitted power for a portion of the bandwidth of each cell is dynamically controlled. The PRBs are allocated to the controlled or non-controlled subbands according to their channel quality. Fig. 1 presents the power-frequency allocation model in a seven adjacent cell layout. It is noted that the formulation presented hereafter is applied to a general network layout.

The frequency band is divided into three disjoint subbands. One subband is allocated to mobiles with the worst signal quality and is denoted interchangeably as a protected band or as an edge band. A user with poor radio conditions is often situated at the cell edge, but could also be closer to the base station and experience deep shadow fading. The remaining two frequency subbands are denoted as centre bands. The separation into zones in Fig. 1 is a simplified representation. We use the subscripts e and c for edge and centre respectively. The main interference in the system originates from transmissions on the centre band (of centre cell users) which interfere with neighbouring cell edge users utilizing their protected band. Denote by P the maximum transmission power per subcarrier. When a base station strongly interferes with its neighbours, the ICIC control process reduces the transmission power per subcarrier in the centre band to αP. Resource block allocation is performed based on a priority scheme for accessing the protected subbands. Let s denote the serving base station of

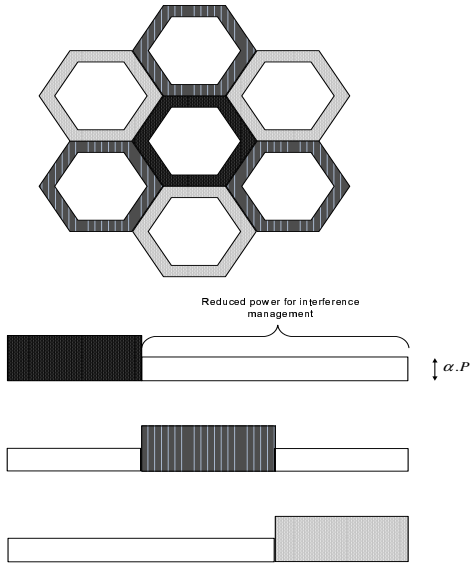

Fig. 1: System model.

the user u. A quality metric h_u is calculated using the pilot channel signal strengths

$$h_u = \frac{Pr_{su}}{\sum\limits_{s' \neq s} Pr_{s'u} + \sigma_n^2}, \tag{1}$$

where Pr_{su} and $Pr_{s'u}$ denote the mean pilot power received by the user u of the signals transmitted by base stations s and s' respectively, and σ_n^2 is the thermal noise power corresponding to the pilot channel. h_u is similar to the SINR with the difference that in the present ICIC scheme, the data channels used to calculate the SINR are subject to power control. The h_u metric is calculated for all users which are then sorted according to this metric. Users with the worst h_u are allocated resources from the protected band and benefit from maximal transmission power of the base station. When the protected subband is full, the resource block allocation continues from the centre band. We assume here that the PRB allocation to subbands is performed at a long time scale, namely that of the ICIC. It is recalled that finer PRBs' allocation at a short time scale can be implemented by the packet scheduler and can further improve the system performance.

Mobility is considered in the system using hard handover and is performed following a path loss criterion. A mobile user u will perform a handover from base station s to s' if the following condition is verified

$$Pr_{s'u} - Pr_{su} > T_{hyst}, \tag{2}$$

T_{hyst} is a fixed hysteresis margin for all base stations and is set to 6 dB in this study.

B. SINR calculation

The channel model used to calculate the SINR is adapted for implementation in a semi-dynamic network simulator that can

assess performance of large network size with tens to hundreds of base stations. Once calculated, the SINR is mapped into spectral efficiency using quality tables (obtained using a link-level simulator) incorporated within the network simulator. A fast fading term is not included in the channel model, and is implicitly taken into account via the quality tables. Within each subband b, namely center ($b = c$) or edge ($b = e$) subbands of base station s, the power allocated to the subcarriers is identical, i.e. $P_{sc} = \alpha_s P$ or $P_{se} = P$ respectively. The allocation of a mobile u to a given subband depends on the quality metric (1).

Consider a mobile u attached to base station s. The average interference perceived by u and produced by eNBs s', $s' \neq s$ is given by

$$I_{ub} = \sum_{s' \neq s} \Lambda(s, s') \eta_{s'b} \frac{P_{s'b} G(s', u)}{L(s', u)}. \quad (3)$$

$\Lambda(s, s')$ equals one if eNBs s and s' use the same frequency bandwidth and zero otherwise. $P_{s'b}$ is the transmitted power per subcarrier belonging to the frequency subband b, $b \in \{c, e\}$. $\eta_{s'b}$ represents the load of subband b of base station s' and is defined as the ratio between the number of PRBs allocated in subband b, $N_{s'b}^{allocated}$, and the total number of PRBs available in this subband, $N_{s'b}^{available}$:

$$\eta_{s'b} = \frac{N_{s'b}^{allocated}}{N_{s'b}^{available}}. \quad (4)$$

The load coefficient (4) expresses the fact that the average interference on a given subchannel belonging to the frequency subband b is proportional to the portion of time the subchannel is used. Hence $\eta_{s'b}$ equals the probability of a collision (i.e. interference) produced by s'. $G(s', u)$ is the antenna gain of base station s' in the direction of the mobile u. The channel loss $L(s', u)$ at a distance d between s' and mobile u is given by

$$L(s', u) = A \chi(s', u) (\frac{1}{d(s', u)})^{\nu}, \quad (5)$$

where A is a constant deduced from the measured path-loss at a reference distance and ν is the path-loss exponent depending on the propagation environment. Having the expression for the interference one can derive the SINR per subcarrier for the mobile u:

$$SINR_{ub} = \frac{P_{sb} G(s, u)}{L(s, u)(I_{ub} + \sigma_z^2)}, \quad (6)$$

where σ_z^2 is the thermal noise per subcarrier and I_{ub} is given by (3).

III. Fuzzy Q-Learning Control (FQLC)

This section describes the FQLC solution for the ICIC problem. We propose a cooperative model for the learning (exploitation) phase to accelerate the learning process, and a distributed model for the exploitation phase which has the advantage of being scalable. The learning model is an approximated Markov Decision Process (MDP) and belongs

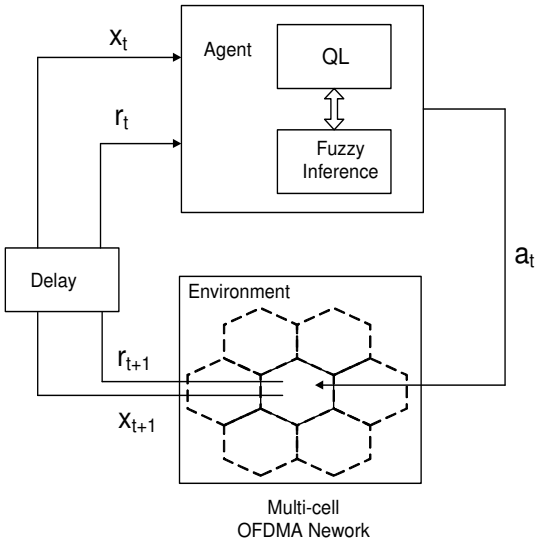

Fig. 2: Fuzzy Q-Learning Controller.

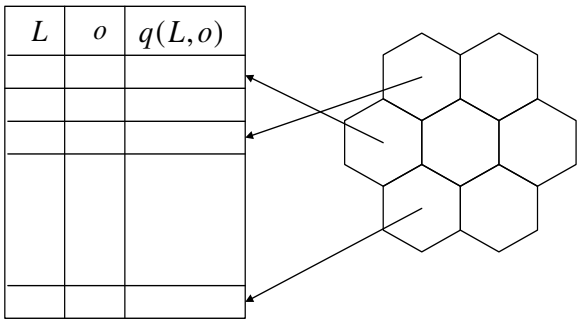

Fig. 3: Cooperative learning of the action-value function.

to decentralized RL models [15]. The diagram of the FQLC is presented in Fig. 2. The agent is located within a base station. It collects information or Key Performance Indicators (KPIs) from its own and its neighbouring cells that define the system state x_t at time t, and performs a local action a_t. The agent then transits to a new state x_{t+1} and receives a reward r_{t+1}. The KPIs are filtered to remove short term fluctuations and to stabilize the control process.

The Q-Learning (QL) algorithm is capable of learning the optimal policy that maps states to control actions [6]. For continuous state space (as in the present case) we propose to use the FQL which combines fuzzy logic with the QL algorithm [7]. In the FQL algorithm, the controlled system is presented as a FIS [15]. The system model is considered unknown (i.e. the transitions' probability between states and the rewards) and therefore the problem is solved using a Temporal Difference (TD) solution. The cooperative learning model is enforced by using a global reward comprising the sum of rewards of all the learning agents. In the FQL nomenclature

we use labels, L, which represent discrete states and, together with actions o defining the FIS rules. The agents learn together a common strategy by feeding a single q-table as shown Fig. 3. In addition to a fast convergence time, this model benefits from a diversified experience learned by the cooperating agents.

A. FQLC components

The components of the ICIC FQLC are described below.

- **State:** The input state vector to the FQL controller is defined as follows:

$$x_s = [\ P_{sc} \quad SE_c^s \quad SE_e^{n(s)}\ ]. \tag{7}$$

P_{sc} is the transmitted power on PRBs belonging to the center subbands of base station s. SE_c^s is the mean spectral efficiency of the users served on subcarriers belonging to the center subbands of the controlling cell. $SE_e^{n(s)}$ is an aggregated mean spectral efficiency of users served within the protected subbands of the neighbouring cells. The aggregated KPI of neighbouring cells, $X^{n(s)}$, is defined as a weighted sum over the KPIs (of the same type), $X^{s'}$. s' belongs to the neighbourhood of s, $\mathcal{N}(s)$, namely the neighbouring cells. $X^{n(s)}$ is given by:

$$X^{n(s)} = \sum_{s' \in \mathcal{N}(s)} w_{s's} X^{s'}, \tag{8}$$

$w_{s's}$ are weighting coefficients satisfying $\sum_{s'} w_{s's} = 1$, and reflecting the degree of "neighbourhood" of cell $s' \in \mathcal{N}(s)$ with s. They are calculated once and stored in a table. $w_{s's}$ represent the normalized traffic flux between cells s and s' with respect to the total traffic flux between s and all its neighbours, and can be calculated off-line. The power of the controlling base station, P_{sc}, is included in (7) to allow the controller to deduce the cause of degraded values of $SE_e^{n(s)}$. It helps to decorrelate the impact of the controller from that of other interfering base stations on the degraded signal quality of the neighbouring cells. For example, a poor value of SE_c^s can result in from a low value of P_{sc}, or conversely, from high interference from second tier base stations.

- **Actions and strategies:** The action is the reduced transmit power allocated by a base station to its center subband. The strategy of base station s, π_s, is a mapping between the state of base station s, x_s, and the action $a_s \in \mathcal{A}$, $\mathcal{A}$ being the set of possible actions (transmitted powers in our case) for the base station s:

$$\pi_s : x_s \rightarrow a_s. \tag{9}$$

- **Utility function:** The controller of each base station aims at optimizing a utility function defined by a long term sum of discounted rewards. The optimization problem is formulated as follows:

$$\max_{\pi_s \in \Pi_s} :\ R_s = E_{\pi_s}[\sum_{t=0}^{\infty} \gamma^t r_s(x_{s,t}, a_{s,t})], \tag{10}$$

Π_s is the set of allowable policies for base station s, $r_s(x_{s,t}, a_{s,t})$ is the instantaneous reward as seen by base station s in state $x_{s,t}$ when taking the action $a_{s,t}$ at time t. γ is a discount factor ranging in the interval $[0, 1)$. The smaller γ, the greater the emphasis given by the controller to present rewards with respect to future ones.

- **Instantaneous rewards:** The harmonic mean throughput is chosen as the reward function, which reflects user's satisfaction in a network with data transfer applications. It is recalled that the harmonic mean fairness is a special case of the generalized fairness model [16] in eq. (11) where the parameter α_f defines the degree of fairness and is chosen here as $\alpha_f = 2$. Th denotes a vector of achieved rates in which the component Th_k stands for the throughput achieved by user k.

$$r(Th) = \sum_k \frac{Th_k^{1-\alpha_f}}{1-\alpha_f}. \tag{11}$$

Denote by $Th_{u,t}$ the instantaneous throughput of a user u belonging to $\mathcal{U}(s')$, the users served by cell s'. The instantaneous global reward is a sum over all the individual rewards of the set of cooperating base stations $\mathcal{S}$:

$$r_{s,t} = -\sum_{s' \in \mathcal{S}} \sum_{u \in \mathcal{U}(s')} \frac{1}{Th_{u,t}}. \tag{12}$$

B. FQLC algorithm

In the QL algorithm, the solution to the maximization problem defined in (10) uses the action-value function under the policy π that is defined as the expected sum of the discounted rewards when starting from state $x_0 = x$ at t_0, and is given by:

$$Q^\pi(x, a) = E_\pi[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) | x_0 = x, a_0 = a]. \tag{13}$$

and can be solved iteratively by the following TD QL update equation [6]

$$Q_{t+1}(x_t, a_t) = (1-\kappa) Q_t(x_t, a_t) + \kappa(r_{t+1} + \gamma \max_{a'} Q_t(x_{t+1}, a')), \tag{14}$$

where κ is a learning rate. In the FQL the controlled system is represented as a FIS that receives as input continuous states. The idea of the FQL algorithm is to use a q-value function (or q-function for brevity) defined over a discrete set of states together with special member functions to derive (interpolate) the Q function over the continuous state space. In the FQL, the FIS is presented by a set of rules $\mathcal{J}$ with a rule $j \in \mathcal{J}$ defined as [7]

IF ($x^1$is L_j^1) ...AND ($x^n$is L_j^n) ...AND ($x^N$is L_j^N)
THEN $a = o_j$ with $q(L_j, o_j)$.

L_j^n is a fuzzy label that corresponds to a distinct fuzzy set defined in the domain of the nth component x^n of the state vector $x = [x^1, ..., x^n, ..., x^N]$, and o_j is the output action of the rule j. The vector $L_j = [L_j^1, ..., L_j^n, ..., L_j^N]$ is called the modal vector corresponding to the rule j. $q(L_j, o_j)$ is

called the q-value function of state L_j an action o_j of the rule j. The dimension of the state vector and the number of its corresponding fuzzy labels sets a tradeoff between the accuracy of the system model and the speed of convergence of the learning process. Three labels for each component of the state vector have been used in the present work.

The membership function maps a continuous state component into the degree of membership to a fuzzy set corresponding to a given label. Triangular fuzzy sets are used here, as depicted in Figure 4.

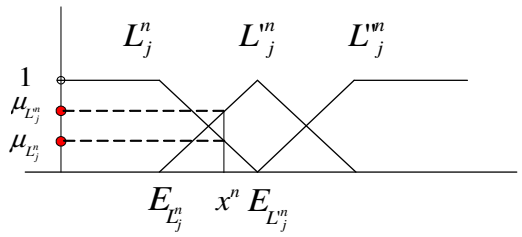

Fig. 4: Three fuzzy sets.

Let J_x denote the set of all rules for which the state x has a positive member function value for each component of the state vector. The membership of a vector x, or the *degree of truth* in the fuzzy logic terminology, with respect to the rule j, $j \in J_x$ is defined as the product of the corresponding member functions of the rule:

$$\alpha_j(x) = \prod_{n=1}^{N} \mu_{L_j^n}(x_n). \tag{15}$$

The q-values are initially set to zero. In the FQL algorithm, the actions are chosen from the set of permissible actions according to an Exploration/Exploitation Policy (EEP). The ϵ-greedy method is used as the EEP policy for choosing the actions:

$$\begin{cases} \text{with prob. } \epsilon: & \forall j \in J_x\text{: } o_j = \underset{k \in \mathcal{K}}{argmax}\, q(L_j, o_k) \\ \text{with prob. } 1-\epsilon: & \forall j \in J_x\text{: } o_j = \underset{k \in \mathcal{K}}{random}(o_k) \end{cases} \tag{16}$$

where ϵ determines the tradeoff between exploration and exploitation, and is typically chosen close to and below 1. $\mathcal{K}$ is the set of action indices and is assumed to be the same for all eNBs.

The relation between the inferred action for an input vector x and the applied rule actions o_j, $j \in J_x$ is given by the following equation

$$a(x) = \sum_{j \in J_x} \alpha_j(x) o_j. \tag{17}$$

$Q(x, a)$ for any input vector x is calculated as an interpolation of the q-value function at the modal points of the activated rules:

$$Q(x, a(x)) = \sum_{j \in J_x} \alpha_j(x) \cdot q(L_j, o_j). \tag{18}$$

We use the value function for the state x which is defined here as

$$V(x) = \sum_{j \in J_x} \alpha_j(x) \cdot \max_k q(L_j, o_k). \tag{19}$$

To update the q function, the quantity ΔQ is defined as the difference between the old and the new value of $Q(x, a(x))$. Denote by y the new state after taking the action $a(x)$ in the state x and receiving the reward r. ΔQ is calculated by:

$$\Delta Q = r + \gamma \cdot V(y) - Q(x, a(x)). \tag{20}$$

The update equation for the q function is given by eq. (21). The subscript t is added to highlight the time dependency in the update equation.

$$\begin{aligned} q_{t+1}(L_j, o_j) = \quad & q_t(L_j, o_j) \\ & + \kappa \cdot \alpha_j(x_t) \cdot (r_{t+1} + \gamma \cdot V_t(x_{t+1}) - Q_t(x_t, a(x_t))), \end{aligned} \tag{21}$$

where κ is a learning rate. Table I presents the FQL algorithm.

TABLE I: Fuzzy Q-Learning Algorithm

1. Initialize $q(L_j, o_k)$ for all $j \in \mathcal{J}$ and $k \in \mathcal{K}$.
2. Calculate the degree of truth of the initial state $\alpha_j(x_0)$ for all $j \in J_{x_0}$ (Eq. (15)).

Repeat (at each time t):

3. For each activated rule L_j, $j \in J_{x_t}$ select an action o_j with the EEP policy (Eq. (16)).
4. Calculate the inferred action $a(x_t)$ corresponding to x_t and o_j, $j \in J_{x_t}$ (Eq. (17)).
5. Calculate the corresponding quality $Q(x_t, a(x_t))$ (Eq. (18)).
6. Execute the action $a(x_t)$ and observe new state x_{t+1} and reinforcement r_{t+1}.
7. Calculate the membership functions $\alpha_j(x_{t+1})$ for $j \in J_{x_{t+1}}$ (Eq. (15)).
8. Calculate the value function of the new state (Eq. (19)).
9. Calculate the variation of the quality ΔQ (Eq. (20)).
10. Update the elementary quality $q(L_j, o_j)$ for each activated rule j, $j \in J_{x_t}$ (Eq. (21)).
11. $t \leftarrow t + 1$

If convergence is attained then stop learning.

IV. Numerical results

A. Simulation environment

The results presented in this section have been obtained using a semi-dynamic network simulator. The simulator performs correlated snapshots to account for the time evolution of the network with a time resolution of the order of a second. FTP-type data traffic is considered. During a time interval between two consecutive snapshots, the following operations are performed: users' arrivals and departures and update of

base station loads; admission control algorithm is executed for each new arrival; the user arrival follows a Poisson process, and the data volume for download is fixed to 5 Megabits. After download completion the user leaves the network and the QoS is calculated for the terminating call. At the end of each time interval, the simulator computes the new positions, radio conditions and handovers of the users. More details on the semi-dynamic network simulator can be found in [17].

A LTE network comprising 45 base stations positioned on a non-regular grid is considered. The inter-site distance varies from 1.5 to 2 km. Each base station has a capacity of 15 PRBs per subband, namely a total of 45 PRBs per base station. Users requesting to initiate a file transfer are allocated 1 to 3 PRBs according to resource availability. Users' speed of 13.88 m/sec is chosen. The noise power spectral density is taken as -173 dBm/Hz. The Okumura-Hata propagation model is used. The path loss at a reference distance of 1 km and the path loss exponent are chosen as -128 dB and 3.76 respectively. The standard deviation of the shadowing process is 6 dB.

The periodicity of the FQLC is of 40 seconds, namely the base stations independently apply the power masks over the protected subbands every 40 seconds. The KPIs serving as inputs to the FQL controller are filtered using an averaging filter over the same duration of 40 seconds. The learning rate κ is set to 0.1 and the discount factor γ - to 0.95. Three fuzzy sets are used per state component and are equally spaced in the domain of variation of the component. The transmit power per PRB can vary from 24 to 32 dBm. Hence the q-table has a total amount of 81 state-action pairs (entries). The control process is carried out over a time period of 170000 iterations (i.e. simulator seconds).

B. Simulation results

The results obtained using the ICIC FQLC approach are compared with two reference systems, both using maximum transmit power over the entire bandwidth: The first utilizes a reuse-1 scheme in which the entire bandwidth is used by each base station. The second utilizes a reuse-3 scheme achieving a complete interference mitigation among first tier of neighbouring cells.

Fig. 5 compares the mean file transfer time of the three systems. Significant improvement with respect to the reuse-1 scheme is obtained. For traffic intensity up to 13.5 arrivals/sec, the ICIC scheme outperforms the reuse-3 scheme, and then the tendency reverses.

Fig. 6 depicts the blocking rates for the three systems as a function of traffic intensity. The blocking rate provides an indicator of capacity, namely the traffic intensity that can be served by the network for a given blocking rate. The ICIC FQLC approach offers a better system capacity with respect to the two reference systems. The poor performance of reuse-3 scheme is explained by the fact that only a third of the available bandwidth resources are allocated within each cell.

Fig. 7 presents the comparison for the mean SINR perceived by all users. The ICIC FQLC solution is better than the reuse-3 solutions for moderate traffic intensity and for

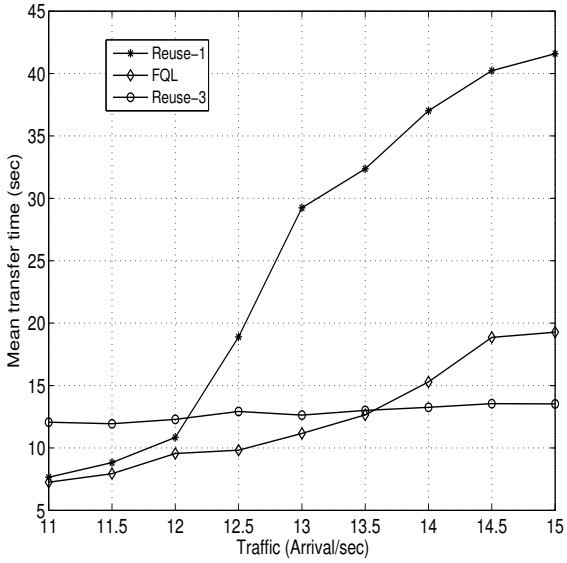

Fig. 5: Mean file transfer time as a function of traffic intensity.

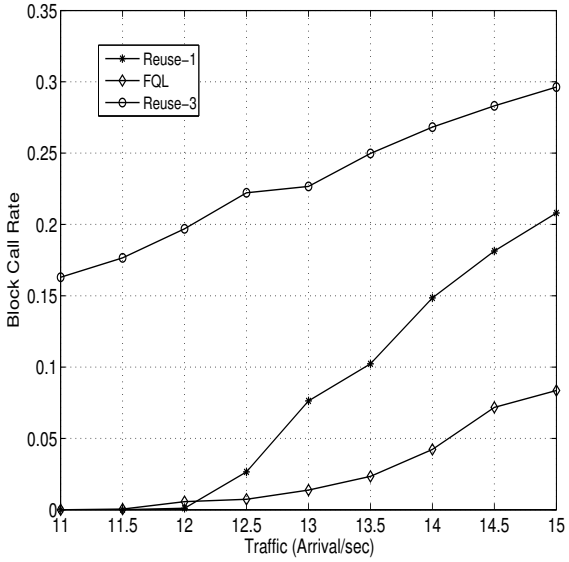

Fig. 6: Blocking rates in the system as a function of traffic intensity.

acceptable blocking rates. The tendency reverses for higher traffic intensity for which blocking rate is high. The planning (operation) point will be selected for traffic intensity for which the blocking rate is relatively small, e.g. below five percent or 14 arrivals per second. The reuse-1 suffers from low average SINR due to inter-cell interference. The results for the average throughput per PRB follow the same tendency.

Fig. 8 compares the cumulative distribution function (cdf) of users' file transfer time for the three systems. The ICIC FQLC solution achieves lower values for the file transfer time, even when compared with the system implementing the reuse-3 scheme. This does not come as a surprise since the harmonic

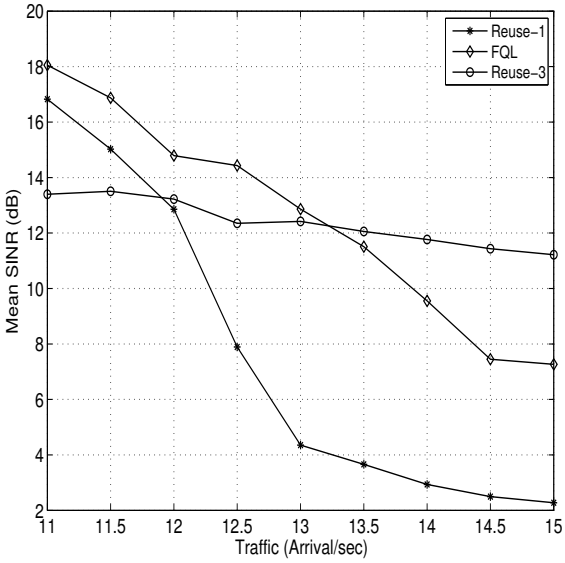

Fig. 7: Mean SINR as a function of traffic intensity.

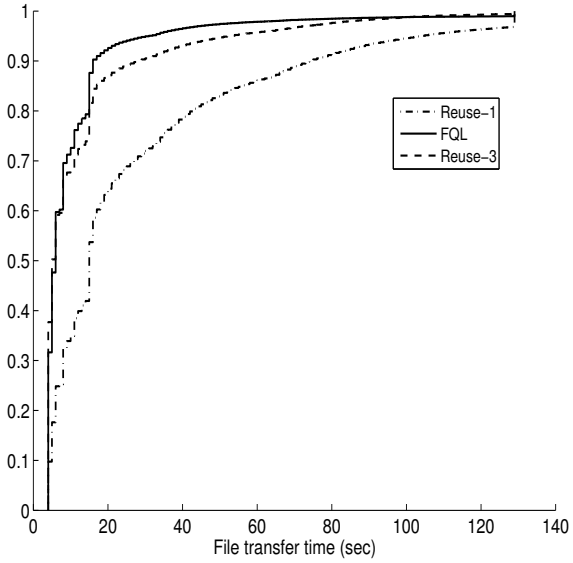

Fig. 8: Cdf of the file transfer time.

throughput has been used as the objective that guides the FQL process. The ICIC mechanism enforces harmonic mean fairness and achieves joint improvement for all users. In other words, improvement of users with bad quality does not come at the expense of users with good quality.

V. Conclusion

This paper has presented a distributed solution for ICIC in OFDMA networks based on a Reinforcement Learning approach with a Fuzzy Q-Learning implementation. The learning phase is cooperative, i.e. the agents share their learned experience. The exploitation phase is fully distributed and can be implemented in a non-synchronized manner, resulting in a scalable solution that can be implemented in a real network. The numerical results have shown important enhancement brought about by the ICIC FQLC in terms of system capacity and file transfer time with respect to two reference systems implementing reuse-1 and reuse-3 schemes.

References

[1] 3GPP TR 36.913 v8.0.0, "Requirements for Further Advancements for E-UTRA (LTE-Advanced)", (Release 8), June 2008.

[2] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, 3G Evolution: HSPA and LTE for Mobile Broadband, Academic Press 2007.

[3] JG Andrews, A. Ghosh, and R. Muhamed, Fundamentals of WiMAX, Prentice-Hall, 2007.

[4] ITU-R M.1645, "Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000".

[5] IST WINNER II project, Interference avoidance concept, Delivrable D4.7.2, June 2007.

[6] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, 1998.

[7] P. Y. Glorennec, "Reinforcement Learning: an overview," European Sym. on Intelligent Techniques, Aachen, Germany, 2000.

[8] A. L. Stolyar, H. Viswanathan, "Self-organizing dynamic fractional frequency reuse for best-effort traffic through distributed inter-cell coordination," Proc. IEEE Proc. INFOCOM, Rio de Janeiro, April 2009.

[9] J. Gross et al, "Comparison of heuristic and optimal subcarrier assignment algorithms," Proc. ICWN'03, June 2003.

[10] R. Combes Z. Altman, E. Altman, "On the use of packet scheduling in self-optimization processes: application to coverage-capacity optimization", to be published in WiOpt 2010.

[11] J.L. van den Berg et al, "Self-organizion in future mobile communication networks," ICT - Mobile Summit 2008, Stockholm, Sweden, June 10-12, 2008.

[12] "NGMN recommandation on SON and O&M requirements", a requirement Specification by the MGMN Alliance, Dec. 2008.

[13] 3GPP TR 36.902, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self configuration and self-optimization network use cases and solutions," Release 9, May 2009.

[14] L. Jouffe, "Fuzzy inference system learning by reinforcement methods," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 28, no. 3, pp. 338-355, Aug. 1998.

[15] L. Matignon, G.J. Laurent, and N. Le Fort-Piat, "Hysteretic Q-Learning: an algorithm for decentralized Reinforcement Learning in Cooperative Multi-Agent teams," IEEE/RSJ IROS 2007 conference, San Diego, California, Nov. 2007.

[16] J. Mo, J. Walrand, "Fair end-to-end window-based congestion control," IEEE/ACM Trans. on Networking, vol. 8, Oct. 2000, pp. 556 - 567.

[17] A. Samhat, Z. Altman, M. Francisco, B. Fouresti, "Semi-dynamic simulator for large scale heterogeneous wireless networks", International Journal on Mobile Network Design and Innovation (IJMNDI), Vol. 1, N. 3-4, pp. 269-278, 2006.