Scheduling Policies Based on Dynamic Throughput and Fairness Tradeoff Control in LTE-A Networks

Ioan Sorin Comşa, Mehmet Aydin, Sijing Zhang
Institute for Research in Applicable Computing
University of Bedfordshire
Luton, LU1 3JU, United Kingdom
{Ioan.Comsa, Mehmet.Aydin, Sijing.Zhang}@beds.ac.uk

Pierre Kuonen, Jean-Frederic Wagen, Yao Lu
Institute for Complex Systems
University of Applied Sciences of Western Switzerland
Fribourg, CH-1705, Switzerland
{Pierre.Kuonen, Jean-Frederic.Wagen, Yao.Lu}@hefr.ch

Abstract — **In LTE-A cellular networks there is a fundamental trade-off between the cell throughput and fairness levels for preselected users which are sharing the same amount of resources at one transmission time interval (TTI). The static parameterization of the Generalized Proportional Fair (GPF) scheduling rule is not able to maintain a satisfactory level of fairness at each TTI when a very dynamic radio environment is considered. The novelty of the current paper aims to find the optimal policy of GPF parameters in order to respect the fairness criterion. From sustainability reasons, the multi-layer perceptron neural network (MLPNN) is used to map at each TTI the continuous and multidimensional scheduler state into a desired GPF parameter. The MLPNN non-linear function is trained TTI-by-TTI based on the interaction between LTE scheduler and the proposed intelligent controller. The interaction is modeled by using the reinforcement learning (RL) principle in which the LTE scheduler behavior is modeled based on the Markov Decision Process (MDP) property. The continuous actor-critic learning automata (CACLA) RL algorithm is proposed to select at each TTI the continuous and optimal GPF parameter for a given MDP problem. The results indicate that CACLA enhances the convergence speed to the optimal fairness condition when compared with other existing methods by minimizing in the same time the number of TTIs when the scheduler is declared unfair.**

Keywords- LTE-A, TTI, CQI, throughput, fairness, scheduling rule, policy, MLPNN, RL, MDP, CACLA.

I. Introduction

In the Orthogonal Frequency Division Multiple Access (OFDMA) radio access networks, the system throughput and user fairness tradeoff optimization problem has to maximize the total cell throughput while maintaining a certain level of fairness between user throughputs. One way to maximize the total system throughput subject to fairness constraints is to use opportunistic schedulers such as channel aware based GPF scheduling rule that exploits the multi-user diversity principle. Therefore different tradeoff levels can be obtained by using a proper parameterization of the GPF scheduling scheme [1].

The Next Generation Mobile Networks (NGMN) fairness requirement [2] is used for the fairness criterion adoption which requires a predefined user throughput distribution to be achieved. Based on the NGMN concept, the scheduler is considered to be fair if and only if each user achieves a certain percentage from the Cumulative Distribution Function (CDF) of other normalized user throughputs (NUT). Based on the scheduler instantaneous state (channel conditions, user throughputs and traffic loads), the GPF rule should be adapted in such a manner that the obtained CDF curve of NUTs respects the NGMN fairness condition. By assuming that the NGMN optimality criterion depends only on the previous GPF parameterization, the scheduling procedure can then be modeled as a MDP with the respect of the Markov property.

The innovation of the current work aims to explore the unknown behavior of the scheduler states in order to learn the optimal policy of GPF parameters in such a way that the NGMN fairness requirement is satisfied at each TTI. The CACLA RL algorithm is proposed in this sense to solve given MDP problems by selecting optimal actions. The quality of applying different continuous GPF parameters in different continuous scheduler states is approximated by using a non-linear MLPNN function. The rest of the document is organized as follows: Section II highlights the importance of the fairness-throughput tradeoff optimization problem. Section III presents the elements of the related work. In section IV, the insight elements of the proposed controller are analyzed. Section V presents the results, and the paper concludes with Section VI.

II. User Fairness And System Throughput Tradeoff Optimization Problem

In LTE packet scheduling, a set $\mathcal{U}_t$ of preselected users is scheduled at each TTI t in frequency domain by using a set of $\mathcal{B}$ grouped OFDMA sub-carriers denoted as resource blocks (RBs). The resource allocation procedure in time-frequency domain follows the integer linear programming optimization problem at each TTI t as shown in Eq. (1):

$$\begin{aligned} &\max_{b_{i,j}} \sum_{i\in\mathcal{U}_t}\sum_{j\in\mathcal{B}} b_{i,j}[t]\cdot\left\{ r_{i,j}[t] / \left(\overline{T}_i[t]\right)^{\alpha_k[t]}\right\} \\ &s.t. \quad \sum_{i\in\mathcal{U}_t} b_{i,j}[t]=1, \forall i\in\mathcal{U}_t \\ &\qquad b_{i,j}[t]\in\{0,1\}, \forall i\in\mathcal{U}_t \end{aligned} \qquad (1)$$

where $b_{i,j}$ represents the allocation vector, $r_{i,j}$ is the achievable rate for user i and RB j, and $\overline{T}_i$ denotes the average throughput of user i averaged over a number of TTI by using the exponential moving filter [1]. The fairness-throughput tradeoff is tuned by parameter $\alpha_k[t]$ which can be adapted TTI-by-TTI in order to meet the objective function. When $\alpha_k = 0$ for the entire transmission, the obtained GPF rule maximizes the throughput (MT). If $\alpha_k = 1$, the obtained scheme is entitled Proportional Fair (PF) and when α_k is very large $\left(\alpha_k \to \infty\right)$, the scheduler maximizes the fairness between users and minimizes the system throughput and the obtained rule is entitled maximum fairness (MF). The optimal scheduler state in which the NGMN requirement is respected can be achieved by setting α_k at each TTI t such as:

$$\alpha_k\left[t\right] = \alpha_k\left[t-1\right] + \Delta\alpha_k \tag{2}$$

where $\Delta\alpha_k$ is the optimal $\alpha_k\left[t\right]$ parameter step. Let us define $\mathcal{A}_k\left[t\right] \in \mathcal{A} = \left\{\Delta\alpha_k\right\}, k = 1,..,\left|\mathcal{A}\right|$ as the decision vector of action taken at TTI t in order to close the scheduler nearby or in the optimal state. The action set $\mathcal{A}$ can be discrete or continuous. Obviously, the current action $\mathcal{A}_k\left[t\right]$ should be chosen in such a manner that the tradeoff objective function $\Phi\left[t+1\right]$ in the next state is maximized. By using the estimation operator $\mathbb{E}\left\{\cdot\right\}$, the tradeoff action can be seen as a decision vector of the second linear programming optimization problem which should be solved before Eq. (1):

$$\begin{aligned} &\max_{\mathcal{A}_k} \sum_{k\in\mathcal{A}} \mathcal{A}_k\left[t\right] \cdot \mathbb{E}\left\{\Phi\left[\mathcal{S}_{t+1}^{S}\right]\right\} \\ &\text{s.t.} \quad \sum_{k\in\mathcal{A}} \mathcal{A}_k\left[t\right] = 1, \forall k = 1,..,\left|\mathcal{A}\right| \\ &\qquad\ \mathcal{A}_k\left[t\right] \in \left\{0,1\right\}, \forall k = 1,..,\left|\mathcal{A}\right| \end{aligned} \tag{3}$$

where $\mathcal{S}_{t+1}^{S}$ represents the scheduler state in the next TTI t+1. The NGMN objective function $\Phi\left[\mathcal{S}_t^{S}\right]$ is calculated based on the CDF function of NUT observations set $\left\{\hat{\hat{T}}_i\left[t\right]\right\} \subset \mathcal{S}_t^{S}$, $i = 1,..,\left|\mathcal{U}_t\right|$, as shown by Fig. 1. The NGMN fairness requirement is the oblique continuous line. If the CDF curve is located on the left side of the NGMN requirement (MT rule case), the system is considered unfair $\left(\mathcal{S}_{t+1}^{S} \in \mathcal{UF}\right)$, and when the CDF function lies on the right side (MF and PF rules cases), the system is declared fair $\left(\mathcal{S}_{t+1}^{S} \in \mathcal{F}\right)$. In order to determine the optimal or feasible region in the CDF domain, the superior limit of the NGMN requirement should be imposed (dot oblique line). In this sense, the fair area is divided in two sub-regions: feasible $\left(\mathcal{S}_{t+1}^{S} \in \mathcal{FA}\right)$ and over-fairness $\left(\mathcal{S}_{t+1}^{S} \in \mathcal{OF}\right)$ where $\left\{\mathcal{F}\right\} = \left\{\mathcal{FA}\right\} \bigcup \left\{\mathcal{OF}\right\}$. As seen from Fig.1, only the green curve respects the feasibility condition. Therefore, at each TTI the action $\mathcal{A}_k\left[t\right]$ should be chosen in such a manner that $\mathcal{S}_{t+1}^{S} \in \mathcal{FA}$. Based on Fig. 1, the NGMN objective function is calculated based on Eq. (4)

$$\Phi\left[\mathcal{S}_t^{S}\right] = \left(1/\left|\mathcal{U}_t\right|\right) \sum_{i\in\mathcal{U}_t} \left[\Psi\left(\hat{\hat{T}}_i\right) - \Psi^{Req}\left(\hat{\hat{T}}_i\right)\right] \le 0 \tag{4}$$

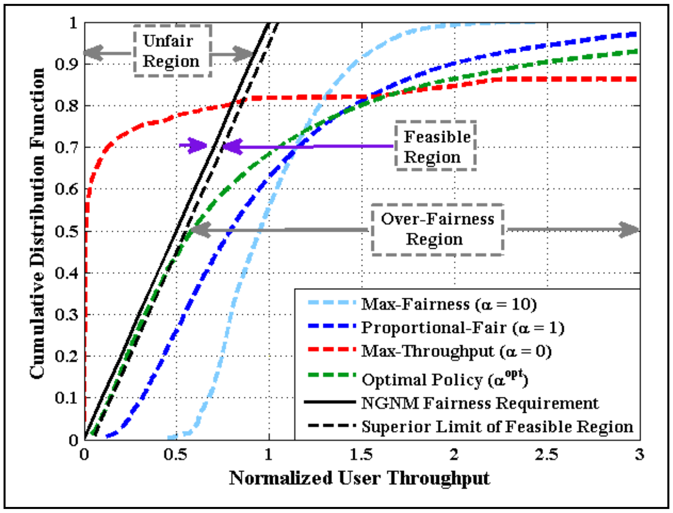

Fig. 1 NGMN Fairness evaluation criteria (benchmarks) for 60 users scenario equally distributed from ENodeB base station to the edge of cell under uniform power allocation and Frequency Division Duplex downlink transmission with a system bandwidth of 20MHz

where Ψ and Ψ^{Req} represent the CDF function and the NGMN fairness requirement, respectively. Equation (4) is the cost function which aims to praise the quality of action $\Delta\alpha_k$ taken in the previous state. Due to the noisy characteristic of Eq. (4) , the CACLA RL algorithm requires the additional function in order to learn the optimal policies of GPF parameters in such a way that $\mathcal{S}_t^{S} \in \mathcal{FA}$.

III. Related Work

The parameterization of the GPF scheduler for the system throughput maximization under NGMN requirement is discussed in [3]. The impact of the traffic load and user rate constraints are considered when the CDF distribution of $\hat{T}_{k,t}$ is determined. Unfortunately, the adaptation process is achieved at different time scales in order to make the proposal suitable for real time scheduling leading to the inflexible behavior when severe changes in the network conditions may occur. In [4] an off-line procedure of adapting the α parameter subject of different temporal fairness indices constraints is proposed. The expected user throughput is calculated at the beginning of each TTI in order to predict the current state of the average user throughput before the scheduling decision. However, the traffic load is not considered and the method cannot be applied to the real systems due to the high complexity cost when the number of active flows increases. In this study, the method from [4] can suffer a slight modification in the sense that $\alpha_k\left[t\right]$ can be adapted based on the NGMN constraint where the CDF function is calculated based on the predicted throughput. The balance of the system throughput and user fairness tradeoff is analyzed in [1], in which the traffic load is categorized based on the CQI reports. The normalized system throughput and Jain Fairness Index are considered as a part of the input state. The Q-Learning algorithm is used to learn different policies that converge very well to different tradeoff levels. However, the concept is not extended to dynamic fairness requirement.

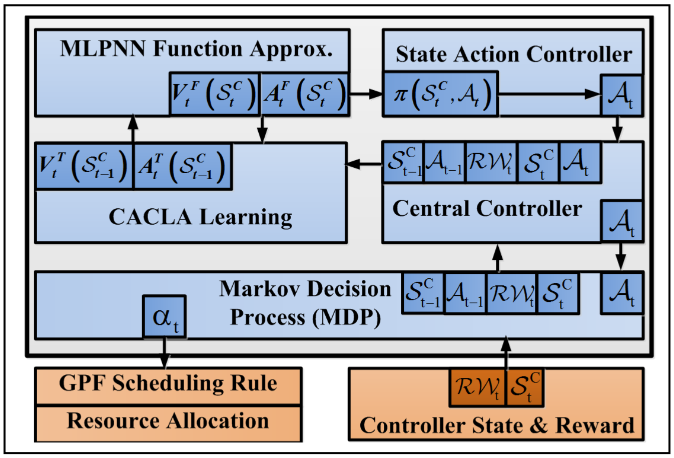

Fig. 2 Proposed Scheduler-Controller Architecture

IV. PROPOSED ARCHITECTURE

The interaction between controller and scheduler shown in Fig. 2 is modeled in two stages: *exploration* and *exploitation*. In the first stage, the LTE controller receives a new state which is the aggregated version of $\mathcal{S}_t^S$ such as $\mathcal{S}_t^C$. Based on the trial and error principle, the controller takes random actions that are mapped into scheduling decisions by the scheduler. The scheduler ranks the previous scheduling decision at the beginning of the next TTI based on the reward function such as $\mathcal{RW}_t\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}^a\right)$. Basically, the reward function $\mathcal{RW}_t$ indicates how far or close is the function $\Phi\left[\mathcal{S}_t^S\right]$ from its objective when compared with the previous state when action $\mathcal{A}_{t-1}^a$ is applied. The exploration stage target is to form a policy of scheduling decisions that follows those actions that maximize the sum of future rewards for every initial state. The exploitation stage applies the learned policy TTI-by-TTI. In order to learn the optimal policy, the MLPNN non-linear function is required to approximate the continuous and multidimensional state $\mathcal{S}_t^C$ in optimal GPF continuous parameters. In this sense, the MLPNN weights are trained by using the gradient descent algorithm with feed-forward (FP) and backward propagation (BP) principles. The BP minimizes the error between the target output and the one which is obtained through FP procedure. The way how the error and target values are calculated determines the type of RL algorithm which is used for the optimal GPF parameterization.

A. Controller State Space

The controller state space contains the relevant information including the previous GPF parameter, a representative compacted state of NUTs and an indication about how close or far the objective function $\Phi\left[\mathcal{S}_t^S\right]$ is from the optimal value. Therefore, the input controller continuous state space is represented by the following set with normalized elements:

$$\mathcal{S}_t^C = \left\{\alpha_{t-1},\ \mu_t^{\hat{T}},\ \sigma_t^{\hat{T}},\ d_t^R,\ |\mathcal{U}_t|,\ \rho_t\right\} \tag{5}$$

where $\mu_t^{\hat{T}}$ and $\sigma_t^{\hat{T}}$ represent the mean deviation and the standard deviation respectively for the log-normal distributions of NUTs, ρ_t is the controller flag which indicates that $\mathcal{S}_t^C \in \mathcal{UF}$ when $\rho_t = -1$, $\mathcal{S}_t^C \in \mathcal{OF}$ when $\rho_t = 0$ and the controller is feasible $\left(\mathcal{S}_t^C \in \mathcal{FA}\right)$ when $\rho_t = 1$. The flag ρ_t is determined based on d_t^R which is the representative CDF distance calculated based on Eq. (6) where $d_t^{i,R} = \Psi_i - \Psi_i^{Req}$:

$$d_t^R = \begin{cases} \max\limits_{i \in \mathcal{U}_t} d_t^{i,R}, if\ \exists d_t^{i,R} > 0, \forall i \in \mathcal{U}_t \\ -\min\limits_{i \in \mathcal{U}_t} d_t^{i,R}, if\ \nexists d_t^{i,R} > 0, \forall i \in \mathcal{U}_t \end{cases} \tag{6}$$

Basically, if there is any $d_t^{i,R} > 0$ in the CDF representation, then $\mathcal{S}_t^C \in \mathcal{UF}$, and when all the percentiles are on the right side of the requirement such as $\nexists d_t^{i,R} > 0$, then $\mathcal{S}_t^C \in \mathcal{F}$. When $d_t^R \in [0, \zeta]$ then $\mathcal{S}_t^C \in \mathcal{FA}$, where ζ is the superior limit of feasible region.

B. Reward Function

The reward function is computed from perspective of the transition area between two consecutive TTIs. When the $\mathcal{S}_t^C \in \mathcal{OF}$ (Fig. 1), any increase of $\alpha_k[t]$ moves the scheduler further away from the optimal region. On the other pole, when $\mathcal{S}_t^C \in \mathcal{UF}$, it is undesirable to decrease $\alpha_k[t]$ parameter. Therefore, for the GPF parameterization when $\mathcal{S}_t^C \in \mathcal{UF}$ then $\alpha_k \nearrow$ and when $\mathcal{S}_t^C \in \mathcal{OF}$ then $\alpha_k \searrow$ until the feasible state is reached. Based on the aforementioned characteristics, the reward function for the GPF parameterization case becomes:

$$\mathcal{RW}_t = \begin{cases} \alpha_{t-1} - \alpha_{t-2},\ if\, \alpha_{t-1} \geq \alpha_{t-2}, \mathcal{S}_{t-1}^C \in \mathcal{UF}, \mathcal{S}_t^C \in \mathcal{UF} \\ -1,\ if\, \alpha_{t-1} < \alpha_{t-2}, \mathcal{S}_{t-1}^C \in \{\mathcal{UF}, \mathcal{FS}, \mathcal{OF}\}, \mathcal{S}_t^C \in \mathcal{UF} \\ 0,\ if\, \alpha_{t-1} > \alpha_{t-2}, \mathcal{S}_{t-1}^C \in \mathcal{UF}, \mathcal{S}_t^C \in \mathcal{OF} \\ 1,\ if\ \mathcal{S}_{t-1}^C \in \{\mathcal{UF}, \mathcal{FS}, \mathcal{OF}\}, \mathcal{S}_t^C \in \mathcal{FS} \\ -1,\ if\, \alpha_{t-1} \geq \alpha_{t-2}, \mathcal{S}_{t-1}^C \in \{\mathcal{FS}, \mathcal{OF}\}, \mathcal{S}_t^C \in \mathcal{OF} \\ \alpha_{t-1} - \alpha_{t-2},\ if\, \alpha_{t-1} < \alpha_{t-2}, \mathcal{S}_{t-1}^C \in \mathcal{OF}, \mathcal{S}_t^C \in \mathcal{OF} \end{cases} \tag{7}$$

The goal of the LTE controller is to find the optimal policy $\pi^*(\mathcal{S}_t^C, \mathcal{A}_t^a)$ at each TTI which permits to select the best action for returning the maximum reward within $\mathcal{S}_{t+1}^C$. The CACLA RL algorithm is used to perform the trained policy as an actor and aims to improve it when necessary as a critic.

C. The CACLA RL Algorithm

The CACLA RL algorithm uses one-dimensional continuous actions $\mathcal{A}_t \in \mathbb{R}_{[0,1]}$ which implies $\Delta\alpha_k \in \mathbb{R}_{[0,1]}$. In order to find optimal policies, one MLPNN is used for the continuous action approximation such as $A_t^F\left(\mathcal{S}_t^C\right)$ and another

MLPNN for forwarding the state value $V_t^F\left(\mathcal{S}_t^C\right)$. The notion of state value implies the approximated accumulated reward value for a given state under some learned policies. The principle of CACLA is to update the action value only if the state target value $V_t^T\left(\mathcal{S}_{t-1}^C\right)$ increases the previous update such as [5]:

$$A_t^T\left(\mathcal{S}_{t-1}^C\right) \xleftarrow{\eta_A} A_t^F\left(\mathcal{S}_{t-1}^C\right) \text{ } if \text{ } V_t^T\left(\mathcal{S}_{t-1}^C\right) > V_t^F\left(\mathcal{S}_{t-1}^C\right) \quad (8)$$

where $V_t^T\left(\mathcal{S}_{t-1}^C\right) = \mathcal{RW}_t\left(\mathcal{S}_{t-1}^C, \mathcal{A}_{t-1}\right) + \gamma \cdot V_t^F\left(\mathcal{S}_{t-1}^C\right)$, γ is the discount factor and η_A is the action value learning rate. Alongside its very simple architecture, CACLA can locate relatively faster the optimal state when compared with other RL algorithms such as: Q-learning, QV-learning, SARSA or ACLA [6], [7], [8] by using predefined GPF parameters steps.

V. Simulation Results

We consider a dynamic scenario with fluctuating traffic load within the interval of [10,120] active data flows/users with infinite buffer. Moreover, the analyzed scheduling policies are running on parallel schedulers that use the same conditions for shadowing, path loss, multi-path loss and interference models. In order to test the impact of the proposed algorithms in the performance metrics, the number of active users is randomly switched at each 1s revealing the generality of the proposed scheduling policy. The rest of parameters are listed in Table I.

The scheduling policy obtained by using CACLA RL is compared against the methods proposed in [4] (MT), [5] (AS) and with other policies obtained by exploring with discrete actions based RL algorithms. The exploration is performed for all RL algorithms by using ε -greedy actions. Figure 3 concludes that the CACLA policy outperforms other policies from the number of TTIs when $\mathcal{S}_t^C \in \{\mathcal{UF}, \mathcal{FA}\}$ points of view.

VI. Conclusions

In this paper the CACLA RL algorithm is used in order to adapt and to apply the best fairness parameter for a dynamic radio environment in LTE-Advanced networks. We proved that CACLA, minimize the number of TTIs when the system is declared unfair being able in the same to fast up the convergence speed by minimizing the number of punishments ($\mathcal{RW}_t = -1$) (Fig. 4) when the number of active users changes dramatically.

TABLE I. Simulation ParAmeters

Parameter Names	Values
System bandwidth	20MHz
Cell radius/ User Speed	1000 m/30km/h
Channel Model	Rayleigh Fading (Vehicular A)
Shadowing std. deviation	8 dB
Path Loss/Penetration Loss	128.1 + 37.6 log(d)/10 dB
Carrier frequency/DL power	2GHz/43dBm
Superior Limit of Feasible Region	$\zeta = 0.05$
Exploration/Exploitation periods	1000 s / 200 s
Learning rate /discount factor/ epsilon	0.01/0.99/0.5
No. hidden layers / No. hidden nodes	1/50

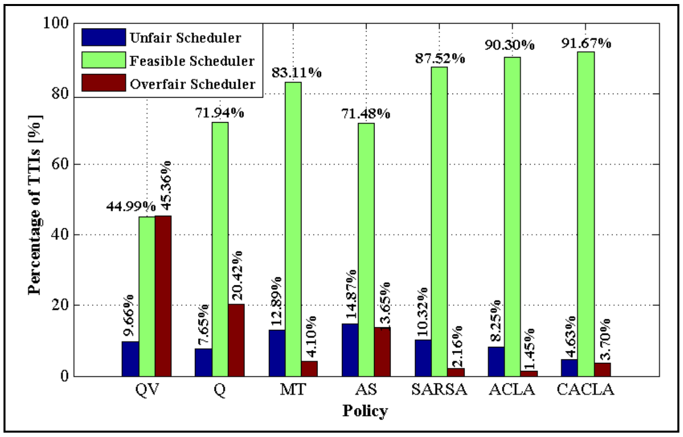

Fig. 3 Reward type percentage and standard deviation

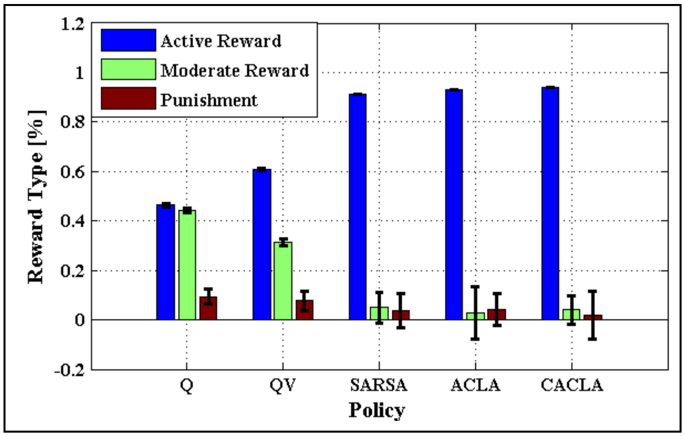

Fig. 4 Number of TTIs when the scheduler state is UF/FEA/OF

References

[1] I.S. Comşa, S. Zhang, M. Aydin, P. Kuonen, and J. F. Wagen, “A Novel Dynamic Q-Learning-Based Scheduler Technique for LTE-Advanced Technologies Using Neural Networks, ” in *37th Annual IEEE Conference on Local Computer Networks (LCN)*, pp. 332-335, Oct. 2012.

[2] R. Irmer, *Radio Access Peiformance Evaluation Methodology*, Next Generation Mobile Networks Std. V 1.3, January 2008.

[3] M. Proebster, C. M. Mueller, and R. Bakker, “Adaptive Fairness Control for a Proportional Fair LTE Scheduler, ” in *IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PMIRC)*, pp. 1504-1509, Sept. 2010.

[4] S. Schwarz, C. Mehlführer, and M. Rupp, “Throughput Maximizing Multiuser Scheduling with Adjustable Fairness,*” in IEEE International Conference on Communications*, pp. 1-5, June 2010.

[5] H. van Hasselt and M. Wiering, "Using Continuous Action Spaces to Solve Discrete Problems, " in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1149 – 1156, June 2009.

[6] C. J. C. H. Watkins and P. Dayan, "{Q}-Learning", in *Machine Learning Journal- Special Issue on Reinforcement Learning*, vol. 8, no. 3/4, May 1992.

[7] S. P. Singh, T. Jaakkola, M. L. Littman and C. Szepesvari, "Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms", in *Machine Learning*, vol. 38, no. 3, pp 287-308, March 2000.

[8] M.A. Wiering and H. van Hasselt, “The QV Family Compared to Other Reinforcement Learning Algorithms,” in *Proceedings of IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL)*, pp. 101-108, March 2009.