A New Fairness Index and Novel Approach for QoS-Aware Resource Allocation in LTE Networks Based on Utility Functions

M. J. Rezaei¹, M. F. Sabahi²*, K. Shahtalebi², R. Mahin Zaeem¹ and R. Sadeghi³

¹- MSc. Student, Electrical Engineering Department, University of Isfahan, Isfahan, Iran.
²- Assistant Professor, Electrical Engineering Department, University of Isfahan, Isfahan, Iran.
³- Assistant Professor, Department of Electrical Engineering, Dolatabad Branch, Islamic Azad University, Isfahan, Iran

ABSTRACT

Resource allocation techniques have recently appeared as a widely recognized feature in LTE networks. Most of existing approaches in resource allocation focus on maximizing network’s utility functions. The great potential of utility function in improving resource allocation and enhancing fairness and mean opinion score (MOS) indexes has attracted large efforts over the last few years. In this paper, a new fairness index is proposed to measure resource allocation performance for real-time/delay-tolerant applications. This index can suggest a new approach for resource allocation. There are several methods in resource allocation of cellular networks which employ fairness index for performance evaluation. Here, we focus on utility-function-based resources allocation and related algorithms. According to the suggested method, the base station (BS) allocates resources based on different service requirements. Appropriate utility function for each application is defined, and the requested quality-of-services (QoS) are satisfied through solving the corresponding optimization problem. The new well-defined fairness index shows that the proposed method has a good performance for different real-time/delay-tolerant applications. Additionally, numerical results show that this approach is able to improve other important indicators such as throughput and MOS as well.

KEYWORDS

Resource Allocation, Fairness Index, MOS, Throughput, Utility Function.

*Corresponding Author, Email: sabahi@eng.ui.ac.ir
1. INTRODUCTION

Increasing in variety of telecommunications applications, growing number of users and limited network resources are the most important challenges in next-generation cellular networks [1]. Different approaches can be designed and implemented by appropriate allocation of resources to achieve optimal values of main indicators. Important indicators for resource allocation are MOS, fairness index, throughput, ease of implementation and interference reduction [2]. From the perspective of network servers, increase in throughput of the system leads to increase in revenue. However, this increase does not necessarily result in enhancement in other indicators such as MOS, which represents network user’s satisfaction rate [3]. The existing cellular networks tend to increase user satisfaction and fairness during network resource allocation. Therefore, the resource allocation approaches not only should be based on maximizing system throughput but also they should provide user satisfaction based on their demands. In [4], the MOS index is defined for measuring user satisfaction in various applications. In [5], resource allocation is applied for each user by employing proportional approach. This method can also focus on the types of user’s services demands. The authors of [6]-[7] have investigated the tradeoff of between different approaches which increase one index and reduce other indicators.

In [8], it is discussed that user satisfaction can be intensified using utility theory, which comes from economics. The authors of [8] have introduced an exact model for different requirements of user’s services. In [9], a good method for scheduling the resource allocation process is proposed. In [10], for the first time, utility functions are listed in four categories including elastic, real time, rate adaptation and step wise. The method described in [11] allocates the resources based on requested types of user’s QoS. Moreover, an algorithm to maximize total utility functions has been introduced in [11].

There are some well-known indexes for evaluating resource allocation algorithms. Throughput index represents the ratio of bandwidth to the time that is taken for allocating user’s rate [12]. As a result, the most favorable status for BS is obtained when this time being minimum and leads to maximum throughput. User satisfaction index or MOS is fully listed in [13] for different services in terms of rate, delay and signal to noise ratio. The indexes are applied in unique function as used in [7] or stand in the utility functions. Fairness indexes in [13]-[14] are also widely used for utility function-based approaches. Jain index and MMF index are presented respectively in [14] and [15]. In these two approaches, the best situation will occur while all users have the same utility coefficients.

In addition, there are some works which exploit different scheduling algorithm [16]-[19]. In [16], the authors suggested first-in-first-out (FIFO) queue for resource allocation wherein each user can access the resources with the priority according to its time of respond. Round Robin (RR) is the next scheduling method which is used in [17]. Its operation is similar to FIFO, but it includes a memory to allocate more resources to the user which receives fewer resources in the previous resource allocation [17]. In another approach, blind equal throughput (BET) was employed in [18]. In this approach, the main goal of resource allocation is to achieve more throughput. Weighted fairness queuing (WFQ) is the next scheduling scheme which was investigated in [19] and indicated the best performance in comparison of FIFO, RR and BET algorithms [19].

Nevertheless, the existing indexes were not able to evaluate the resource allocation methods based on services quality. Therefore, we need to define a new index to measure fairness accurately. Indeed, real-time users can reach a high utility by small amount of rate allocated, however, for delay-tolerant users, more utility may results in huge costs on resource allocation. Therefore, the Jain and MMF indexes cannot appropriately measure the fairness. Here, we seek to provide a new resource allocation method that has ability to maximize overall utilities by applying allowed delay and services qualities. Besides, a new index to measure the fairness of the proposed approach is presented.

The rest of the paper is organized as follows. Section II is dedicated to the problem formulation and its solution. In section III, the numerical results for the proposed index and other indicators are presented. Finally there is a conclusion in section IV.

2. PROBLEM FORMULATION AND ITS SOLUTION

A. Problem Formulation

In this section, we generally study the utility-based resource allocation process. The main objectives of the resource allocation problem are maximizing aggregated utility and providing fairness during the allocation of the resources among the users, subject to limited resources. Two types of real-time and delay-tolerant services should be handled differently in a network to maximize the overall satisfaction of the users. In each cell of a cellular
network, it is assumed that there is a BS and a lot of users with different utility functions (Fig. 1). Each utility function is either sigmoidal-like for real-time applications or logarithmic type for delay-tolerant applications (Fig. 2). These two cases are presented in [10]. Maximizing overall utility functions in [10] is unaware of exact value of allowed delay for each user. Although users with a logarithmic utility functions are resistant against delay, we can consider delay parameter for them in order to have a unified form for utility functions of all users. Table 1 indicates the QoS classes of 3GPP standard. The allowable delay for some services are shown in this table. The utility function parameters for each user are shown in table 2.

![System Model of LTE cellular network: a single cell with one eNodeB and 6 UEs having delay-tolerant or real-time application.](image1)

![The user’s utility functions (three sigmoidal-like functions and three logarithmic functions).](image2)

### Table 1. 3GPP Standardized QoS Classes

<table>
<thead>
<tr>
<th>QoS Class</th>
<th>Bit Rate Requirement</th>
<th>PDR (ms)</th>
<th>Packet Error Rate</th>
<th>Service Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GBR</td>
<td>100</td>
<td>10^{-2}</td>
<td>Live Voice Streaming</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>150</td>
<td>10^{-3}</td>
<td>Live Video Streaming</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>50</td>
<td>10^{-3}</td>
<td>Real-time gaming</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>300</td>
<td>10^{-6}</td>
<td>Buffered Video Streaming</td>
</tr>
<tr>
<td>5 Non-GBR</td>
<td></td>
<td>100</td>
<td>10^{-6}</td>
<td>IMS signaling</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>300</td>
<td>10^{-6}</td>
<td>Web traffic for privileged users</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>100</td>
<td>10^{-3}</td>
<td>Interactive gaming</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>300</td>
<td>10^{-6}</td>
<td>Web traffic for standard users</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td>Elastic traffic</td>
</tr>
</tbody>
</table>

### Table 2. Values for Utility Function Models

<table>
<thead>
<tr>
<th>Model type</th>
<th>Number of Models</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sigmoidal</td>
<td></td>
<td>a_1 = 5, b_1 = 0</td>
<td>a_2 = 3, b_2 = 20</td>
<td>a_3 = 1, b_3 = 30</td>
</tr>
<tr>
<td>Logarithmic</td>
<td></td>
<td>k_1 = 15, D_1 = 150</td>
<td>k_2 = 3, D_2 = 100</td>
<td>k_3 = 5, D_3 = 50</td>
</tr>
</tbody>
</table>

Appropriate index for evaluating a delay-aware resource allocating scheme should have an inverse relationship with allowable delay, which means more allowed delay for user, decreases its priority in the resource allocation process. Assume that there are N users in a cell and the utility function of user i, 1 ≤ i ≤ N, is represented by \( U_i(r_i) \) where \( r_i \) is specific rate for the i-th user. Regarding to being delay-tolerant or delay-sensitive service, the utility function has one of the forms in (1).

\[
U_i(r_i) = \begin{cases} 
  c_i \left( \frac{1}{1 + a^i r_i - b_i} - d_i \right): & \text{Sigmoidal} \\
  \log(1 + k_i r_i) / \log(1 + k_max): & \text{Logarithmic} 
\end{cases} 
\]

where \( c_i = \frac{1 + e^{a_i b_i}}{e^{a_i b_i}}, \quad d_i = \frac{1}{1 + e^a b_i} \) and \( \{a_i, b_i\} \) and \( k_i \) represent the parameters of sigmoidal and logarithmic functions respectively. These parameters are determined by the applications. \( r_{\text{max}} \) represents the maximum rate allocated in base station. These two functions are increasing. Also we have \( U_i(r_i) = 0 \) for \( r_i = 0 \) and \( U_i(r_i) = 1 \) for \( r_i = \infty \). For the i-th user, maximum allowable delay is denoted by \( D_i \).

We are seeking for a new index such that it can maximize total utility functions with respect to permissible delay. Our proposed index is utility-proportional delay-aware (UPD) and is defined as follows:

\[
I_{UPD} = \frac{\sum_{i=1}^{N} U_i(r_i)}{\sum_{i=1}^{N} D_i} \tag{2}
\]

It is worth to be noted that the resource allocation is performed by maximizing this index subject to the maximum rate for BS \( r_{\text{max}} \). The term \( \sum_{i=1}^{N} \frac{1}{D_i} \) in the denominator is for limiting the maximum value of index to one. By maximizing \( I_{UPD} \), a weighted sum utility is maximized where the weights are related to delay parameters. The constraint to this maximization is \( \sum_{i=1}^{N} r_i \leq r_{\text{max}} \). As a result, the optimization problem can be formulated in the following form:

\[
r_i^* = \max_{r_i \leq r_{\text{max}}} \frac{I_{UPD}}{c_i \sum_{i=1}^{N} r_i D_i} \tag{3}
\]
This problem is expressed in [11] wherein, the effort is to maximize sum of utility functions and could not able to consider user’s delay. In our problem we stipulate both of utility and delay requirements, in fact, (3) can allocate resources based on both of the utility functions and delays. More allowed delay leads to less resource in compared to another user with the same utility function. In the next section we introduce an algorithm to solve our problem.

B. Solution

In this section, we present solution of (3). As is shown in (3), there is just one constraint to satisfy. Many methods [3]-[14], [20] are suggested to solve this equation, but we select bi-section algorithm [20]. It has been shown that the speed of this algorithm stands in an appropriate situation.

In appendix we show that $I_{UPD}$ is a concave function of $\{r_i^t\}_{1\leq i \leq N}$. Therefore, the optimal solution of (3) can be obtained using Lagrange multipliers. Forming the Lagrangian, we have:

$$\mathcal{L}(r_i) = I_{UPD} - \lambda \left( \sum_{i=1}^{N} r_i - r_{max} \right)$$  \hspace{1cm} (4)

Where, $\lambda \geq 0$ is a Lagrange multiplier to establish constraint in (3). By calculating the derivations we have:

$$\frac{\partial \mathcal{L}(r_i)}{\partial r_i} = \frac{\partial u_i(r_i)}{\partial r_i} - \lambda = 0 \Rightarrow \frac{\partial u_i(r_i)}{\partial r_i} = \lambda \frac{D_i}{\Sigma_{i=1}^{N} \frac{1}{D_i}}$$  \hspace{1cm} (5)

Equation (5) has a very important interpreting. In fact, the BS not only should consider the $i$-th user delay, but also should pay attention to sum of the inverse delays. From the Equation (5) it can be concluded that more delay value causes to be placed in a part with more slope. Based on the utility figures, more delay cause less rate.

In [10], $\frac{\partial u_i(r_i)}{\partial r_i}$ were calculated for the both types of utility functions introduced in equation (1). Using the well-known bi-section method [21], we can solve (3). Solution of (5) can be done by algorithm 1. In this algorithm, there are a lot of parameters that we introduce them as below:

- $\delta$ and $\varepsilon_0$ are small threshold values (around zero)
- $\{\lambda_j\}_{1 \leq j \leq 3}$ are values that we use in the bi-section method.
- R is the total rate of the BS covering the N UEs
- $g_{\lambda_j} = (\Sigma_{i=1}^{N} r_i^t) - R$ for $1 \leq j \leq 3$ is function that compute constraint in (2) for corresponding $\{\lambda_j\}_{1 \leq j \leq 3}$

<table>
<thead>
<tr>
<th>Algorithm 1. Determine optimized $r_i^t$</th>
</tr>
</thead>
</table>
| **Input:** Identify $\{U_i(r_i), D_i\}_{1 \leq i \leq N}$ for each user, $\delta << 1$, $\lambda_1 = \varepsilon_0$, $\lambda_2 = \frac{1}{\varepsilon_0}$ ($\varepsilon_0 \approx 0$), $\lambda_3 = \frac{\lambda_1 + \lambda_2}{2}$ and select a value such that $|g_{\lambda_3}| > \delta$
| 1. while $(|g_{\lambda_3}| > \delta)$
| 2. for $1 \leq j \leq 3$
| 3. $\frac{\partial u_i(r_i)}{\partial r_i} = \lambda_j D_i \Sigma_{i=1}^{N} \frac{1}{D_i}$ → $\{r_i^t\}_{1 \leq i \leq N}$
| 4. $g_{\lambda_3} = (\Sigma_{i=1}^{N} r_i^t) - R$
| 5. end for
| 6. if $g_{\lambda_3}; g_{\lambda_3} < 0$
| 7. $\lambda_2 = \lambda_3$;
| 8. else if $g_{\lambda_3}; g_{\lambda_3} < 0$
| 9. $\lambda_1 = \lambda_3$;
| 10. end if
| 11. $\lambda_3 = \frac{\lambda_1 + \lambda_2}{2}$
| 12. end while
| **Output:** $\lambda^* = \lambda_3$ so $\frac{\partial u_i(r_i)}{\partial r_i} = \lambda^* D_i \Sigma_{i=1}^{N} \frac{1}{D_i}$ → $\{r_i^t\}_{1 \leq i \leq N}$

As we can see in algorithm 1, initial values allocate to $\{\lambda_j\}_{1 \leq j \leq 3}$ and (3) denotes accuracy of rate allocation. Then we compute rate by (5) for $\{\lambda_j\}_{1 \leq j \leq 3}$. In next step, $g_{\lambda_3} = (\Sigma_{i=1}^{N} r_i^t) - R$ for $1 \leq j \leq 3$ is computed. Now we decide where the answer stands. For this purpose, the steps 6 to 9 in the algorithm should be performed. This procedure repeats until $|g_{\lambda_3}| \leq \delta$. Outputs of this algorithm are resources (rates) allocated to users.

3. NUMERICAL RESULTS

For simulation, we use two other well-known indicators and a comparison is made. The first one is a Jain’s index that is defined as follows [14].

$$I_{\text{jain}} = \frac{\sum_{i=1}^{N} U_i(r_i)}{\left| \sum_{i=1}^{N} U_i(r_i) \right|}, 1 \leq i \leq N$$ \hspace{1cm} (6)

The maximum value of this indicator is achieved when the same value is obtained for all utility functions. The second indicator is max-min-fairness (MMF) index which is defined as follows.

$$I_{\text{MMF}} = \min \left( \frac{U_i(r_i)}{\max U_i(r_i)} \right), 1 \leq i \leq N$$ \hspace{1cm} (7)

This index is also looking for resource allocation in the way that the same value for utility functions are obtained [14]. The optimal solution for the two approaches would be:

$$U_i(r_i) = \frac{r_{\max}}{N}, 1 \leq i \leq N$$ \hspace{1cm} (8)

In fact, the resource allocation schemes based on these indexes are not able to allocate resources according
to the delay requirements of the requested services; therefore, they have to spend more cost for delay-tolerant users in order to establish (8). In the other words, these approaches cannot determine supremacy of delay-tolerant users. Hence, while resource allocation is performed through maximizing total utility functions, these indexes are not suitable tools for evaluating the approach.

The simulation results are listed as below:

In Fig. 3 and Fig. 4, \( I_{Jain} \) and \( I_{MMF} \) are shown respectively based on two user’s utility functions. It is assumed that there are two users with utility functions \( U_1 \) and \( U_2 \). As can be seen, the maximum can be achieved when \( U_1 = U_2 \). The difference between these two figures is in type of convergence to \( U_1 = U_2 \). The Jain approach has a quadratic form and has a faster convergence. In both figures, it can be seen that the maximum values of the indexes are not necessarily corresponding to the maximum value of total utility functions. In fact, for the small equal values of \( U_1 \) and \( U_2 \), these indexes will be maximized.

In Fig. 5, the proposed index is shown versus \( U_1 \) and \( U_2 \) for two users with two different values of delay \( (D_1 = 100\text{ms} \text{ and } D_2 = 300\text{ms}) \). The first user is more sensitive to delay. So, increasing in its requested rate has more effect on the \( I_{UPD} \). It should be noted that the rate allocation is performed based on both delay and utility function such that the best value of this index is achieved.

In Fig. 6, we compare the fairness index, resulted from the proposed method and four well-known resource allocation algorithms including FIFO, RR, BET, WFQ, [16]-[19]. These resource allocation approaches are also presented in [22-24]. The results of the algorithms based on UPF method [10] and MMF method [13] are also shown. As it is expected, the proposed method provides a better fairness compared to the others. In all simulations, each user selects one of the utility functions (introduced in (1)) randomly. Other parameters of utility functions are available in Table 2. Also, the maximum resource is \( r_{max} = 1500 \).

The MOS indexes for all approaches are shown in Fig. 7. For a single user, the MOS index is five times of utility function [6]. As a result, the MOS index is equal to average of MOS’s for all users. As is shown in Fig. 7, the proposed method has the best performance, due to considering allowable delay in \( I_{UPD} \).

In Fig. 8, the throughput of all approaches have been compared. As it is shown, the proposed method has slightly worse throughput than some others. This is because of the extra time spent for delay-aware rate allocation process in the new \( I_{UPD} \)-based approach.
4. CONCLUSION

In this paper, we proposed a new index based on utility function that not only maximize total utility functions and allocate rate according to user’s requirements, but also consider maximum allowable delay for each user. The proposed index is very suitable for new utility-based resource allocation approaches and can improve the poor performance of existing indexes. We focused on investigation of a method for measuring fairness of resource allocation schemes. Besides, this method provides a way to maximize this index in delay-aware resource allocation. The proposed approach has an improved performance for some important existing indexes such as throughput, and MOS too. Briefly speaking, spending small amounts of cost in throughput, the MOS and fairness will be increased.

APPENDIX

We show that the objective function of (3) is concave. It is necessary to prove that the second derivative of this function is always negative.

\[
\frac{\partial^2 U_{\text{UPD}}}{\partial r_i^2} = \frac{\partial}{\partial r_i} \left( \frac{d_i}{\sum_{1 \leq j \leq N} d_j} \right) = \frac{\partial^2 f_i(r_i)}{\partial r_i^2}, \quad 1 \leq i \leq N
\]  

(9)

In [10], it has been shown that for the both types of utility functions introduced in (1) we have \( \frac{\partial^2 f_i(r_i)}{\partial r_i^2} \leq 0 \) in certain interval of \( r_i \), so \( \frac{\partial^2 U_{\text{UPD}}}{\partial r_i^2} \leq 0 \) and the objective function is concave.

REFERENCES


A New Fairness Index and Novel Approach for QoS-Aware Resource Allocation in LTE Networks Based on Utility Functions


