Low Complexity Utility Based Resource Allocation for 802.16e OFDMA Systems

Guowang Miao, Georgia Institute of Technology (gmiao3@gatech.edu)

Nageen Himayat, Intel Corporation (nageen.himayat@intel.com)

Abstract— A utility based resource allocation scheme is investigated for improving system performance of 802.16 OFDMA systems supporting a mix of multi-media traffic types. We show that a utility based approach can improve system capacity over conventional proportional-fair resource allocation schemes through more efficient allocation of only the required resources demanded by the QoS profile of each traffic class. It is also shown that using average channel-aware metrics for resource allocation allows for local linearization of the utility function, which results in linear complexity OFDMA resource allocation schemes. Further, performance with the reduced complexity scheme is similar to a previously known near-optimal algorithm with cubic complexity.

Index Terms— Cellular systems, WiMAX, OFDMA, QoS, utility based resource allocation, 802.16e, linearization, multimedia traffic, mixed traffic

I. Introduction

We investigate a utility based resource allocation framework for 802.16 OFDMA systems. This frame-work is based on assigning a service-dependent utility to the resources assigned to each subscriber, and allocating resources such that the overall utility across the network is maximized. The utility metric quantifies the level of satisfaction a user experiences towards meeting its quality-of-service (QoS) requirements with each additional assigned resource. Examples of utility functions are given in Figure 1. These functions may be determined through subjective measurements of user QOS experience, or they can be derived quantitatively through characterization of the traffic statistics of each service class.

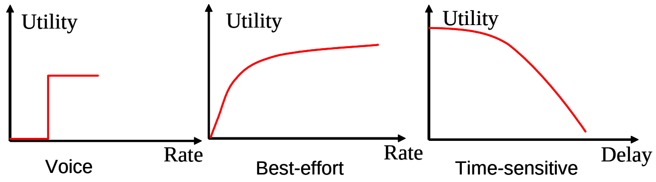

Figure 1: Example utility functions for voice, best-effort and time-sensitive traffic functions.

Utility based resource allocation has been studied in the context of OFDMA systems in [1]-[2], where optimal framework for utility based sub-channel and power allocation is investigated, as well as practical, reduced-complexity resource allocation algorithms are derived. We note that a utility based framework may be considered as a QoS-aware framework for resource allocation, where differing QoS requirements across a mix of service types may be directly included in the optimization objective. Hence, utility based resource allocation can serve as a unifying framework for simultaneous allocation of resources across a mix of multimedia traffic types.

In this paper, we apply the utility-based framework to evaluate system performance improvements for a downlink 802.16 OFDMA wireless communication system. We consider scheduling across a mix of rate-adaptive as well as real-time service classes. We note that although it is possible to consider both channel and queue statistics in the utility based optimizing framework, [6], here we focus on channel-aware scheduling only.

II. Utility Based Resource Allocation

A. Framework & Notation

The objective in utility based resource allocation is to maximize the overall network utility, thereby maximizing the overall satisfaction experienced by all the users in a network. Detailed analysis of this allocation problem may be found in [1]-[2]. Here a brief description is provided. It is assumed that resource allocation decisions are made at the base-station (BS), which makes it decisions on a frame-by-frame basis. Some relevant notation is as follows.

Q_i: is the set of Quanta (resources) assigned to the i^{th} user.

c, s, t : are the sub-channel, symbol and frame indices identifying a Quantum.

$R_i(t,c,s)$: is the rate allocated to the i^{th} user on Quantum c,s for frame t

$I_i(t,c,s)$: is an indicator function indicating whether sub-channel c in time slot s is allocated to the i^{th} user.

$U_i(.)$: is the utility function of the i^{th} user.

$U_i^{'}(.)$: is the derivative of the utility function (also known as the marginal utility function).

M: the total number of users

The data rate per frame of the i^{th} user after allocation is then given as

$$r_i = \sum_{(c,s)\in Q_i} R_i(t,c,s) I(i,t,c,s)$$

Also note that each Quantum is assigned to only one user such that:

$$Q_i \bigcap Q_j = 0 \quad \forall i \neq j \quad and \quad \bigcup_i Q_i = Q$$

The objective is to find the optimal scheduling policy, *I*, satisfying

$$I^* = \arg_I \max \sum_i U_i(r_i) \quad s.t. \quad \sum_i I(i,t,c,s) = 1$$

For the class of concave, continuously differentiable utility functions the optimal solution is given as

$$I^* = \arg_I \max \sum_i U'_i(r_i^*) R_i(t,c,s)$$

It can be seen that achieving the global optimal solution can be complex, as the optimization metric relies on the derivative of the utility function at the total optimal rate allocated to each user in a frame. Hence, the optimal solution must be found by searching over M^Q different possible Quanta assignments. Several, sub-optimal solutions are described in [2], which yield good performance with cubic complexity. Finally, note that if all utility functions are linear then the optimal allocation strategy assigns each Quantum to the user with the highest $U'_i R_i(t,c,s)$. Here, resource assignment can proceed sequentially, Quantum-by-Quantum, as the marginal utility function is constant and no longer a function of the optimal rate per user.

B. *Low Complexity Utility Based Resource Allocation*

In the previous section we observed that "linear utility" function leads to a low-complexity "Quantum-by-Quantum" allocation algorithm that is globally optimal. Therefore it is of interest to use a *locally linear approximation to any feasible utility function, via a Taylor series expansion*. To ignore the higher order terms in the Taylor expansion, we need to ensure that any additional allocation to a user causes only a small perturbation around the current operating point of the expansion. Therefore, we measure utility as function of "average user throughput" rather than as a function of the instantaneous rate assigned to a user. With this assumption, we ensure that additional resources allocated to a user for each Quantum, cause the throughput to change nominally around the current value, allowing for a local linearization of the utility function. Later we show that measuring overall utility as a function of the average throughput, still gives good performance when the instantaneous rate metric is used.

Specifically, the average throughput of the i^{th} user at frame (t+1) can be obtained using an exponentially weighted low-pass filter:

$$T_i(t+1) = (1-\frac{1}{\tau})T_i(t) + \frac{1}{\tau}\sum_{c,s} R_i(t,c,s) I(i,t,c,s)$$

Typically it can be assumed that

$$r = \frac{1}{\tau}\sum_{c,s} R_i(t,c,s) I(i,t,c,s) << (1-\frac{1}{\tau})T_i(t)$$

a desirable outcome for the linear approximation to be valid. Now the optimization problem can be expressed as

$$I^* = \arg_I \max \sum_i U_i(T_i(t+1)) \quad s.t. \quad \sum_i I(i,t,c,s) = 1$$

Using the Taylor series expansion, we can express

$$\begin{aligned} U_i(T_i(t+1)) &= U_i\left((1-\frac{1}{\tau})T_i(t) + \frac{1}{\tau}\sum_{c,s} R_i(t,c,s) I(i,t,c,s) \right) \\ &\approx U_i((1-\frac{1}{\tau})T_i(t)) + \frac{1}{\tau}\sum_{c,s} R_i(t,c,s) I(i,t,c,s) \end{aligned}$$

Hence the optimization problem thus becomes the one of maximizing the following equations:

$$I^* = \arg_I \max \sum_i U'_i((1-\frac{1}{\tau})T_i(t)) R_i(t,c,s) I(i,t,c,s)$$

$$s.t. \sum_i I(i,t,c,s) = 1$$

Therefore the optimal scheduling policy is simply to allocate sub-channel *c* in time-slot *s* to i^{th} user with the largest $U'_i((1-\frac{1}{\tau})T_i(t))R_i(t,c,s)$.

Note that this approximation is applicable to a large class of utility functions for which local linear approximations are possible and not just concave utility functions.

III. System Assumptions

Before we provide simulation results, we briefly describe the system parameters and assumptions.

A. *Utility Functions*

As mentioned, the design of the utility function can be based on subjective measures of a user's QOS experience. However, it is also possible to determine the characteristics of the utility function based on the traffic statistics of each service. Intuitively, if a certain average service rate is not maintained then the queue for a particular service flow will overflow, thereby causing the loss of the queued packet. The net throughput is dominated by the factor $Throughput\ (1 - PacketLossRate)$. Hence, when the allocated service rate is low, the packet loss rate is high, resulting in a low utility for the user for the given rate. For high service rates, the packet loss rate will be negligible, resulting in a utility of almost 1 for the higher service rate. Therefore a utility function may be derived based on traffic characteristics and the average service rate required to maintain a stable queue for the particular service class. The following functions illustrate utility functions described in the literature (see [3][4]) for service classes of interest.

The utility function for rate-adaptive applications (see Figure 2), like file-transfer, email etc., is described by the following equation.

$$U(r)=\begin{cases} U_o \dfrac{(r)^2}{(r_0)^2} & 0<r\le r_0 \\ 1-\dfrac{(1-U_o)r_o^2}{r^2} & r>r_0 \end{cases}$$

Here U_o is the basic utility when user is assigned a rate r_o . For real time applications the utility function may be described by

$$U(r)=\begin{cases} U_0-U_o\sqrt{1-\dfrac{r^2}{r_0^2}} & 0<r\le r_0 \\ U_0+(1-U_0)\sqrt{1-\dfrac{(r-1.4r_0)^2}{0.16r_0^2}} & r_0<r\le 1.4r_0 \\ 1 \quad r>1.4r_0 \end{cases}$$

B. *Simulation Assumptions*

We consider a mix of service classes, each with a separately defined utility function. Table 1 describes the mix of service classes used for the results shown here.

Traffic	Type	ro	Uo
VOIP	Real-Time	100 kbps	0.8
Video Streaming Traffic	Real_time	580 kbps	0.8
Low-Level Rate Adaptive Traffic	Rate Adaptive	512 kbps	0.5
High-Level Rate Adaptive Traffic	Rate Adaptive	1.74 Mbps	0.5

Table I: Mix of service classes used for resource allocation.

We also compare the performance of the low-complexity approach with various other commonly used resource allocation schemes:

Scheduling Method	Description
MAX-SINR	Allocates resources to the user with the best SINR condition.
Round-Robin	Resources are allocated amongst users with equal probability.
Proportional-Fair	Resources are assigned to users who are experiencing their best channel conditions. The proportional fair algorithm is a special case of utility based allocation, where the utility function is logarithmic.
Linear Utility	This is the low-complexity utility scheme.
G-Utility (Iterative method)	This is a utility based approach where an iterative algorithm is used to assign resources. Utility is measured as a function of instantaneous data rates [2].

Table 2: Types of resource allocation schemes considered for comparison.

Cellular layout	19 cell, hexagonal cellular lay-out, 3 sector per site
Cell size	2 km
Frequency reuse	1
DL Modulation and coding (MCS)	QPSK ½, (repetition 1, 2, 4, 6), ¾ 16QAM ½, ¾ 64QAM 1/2, 2/3, 3/4
Permutation	AMC
Link to System Mappings	Mean instantaneous capacity
Channel model	Spatially uncorrelated Flat fading and ITU
HARQ (Hybrid Automatic Repeat Request)	None
Antenna Configuration	1x2
Target packet error rate	1%
Feedback delay	none
Traffic Models	None. Full buffer queues.

Table 3: DL 802.16 simulation assumptions for system level simulation (standard system level simulation methodology as described in [7] is assumed)

IV. System performance results

It can be seen from Figure 3-Figure 4 that the utility based approaches perform the best in terms of delivering high utility to a larger proportion of users. The performance of the low-complexity utility based approach performs close to the near-optimal iterative approach described in [2], especially when measuring utility as a function of average throughput. The performance of the low-complexity approach is also good when utility is measured in terms of instantaneous data rates. Finally, lowering the number of users admitted in the system from 15 to 10, significantly improves the QOS experience of the admitted users. Thus tracking the overall utility in the system can indicate a suitable mechanism for admission control. Figure 5 illustrates the probability of assigning a specific throughput across each service class. It can be seen that the utility based approaches concentrate their allocation around the rates required for yielding high satisfaction for that service class. Thus they avoid unnecessary allocation to users who cannot derive any benefit from the extra resources assigned. As a result, the system is able to satisfy the QOS requirements of a larger number of users through efficient allocation of resources. Computing the average utility per user for fixed outage conditions, we are able to observe approximately 28% improvement in system capacity when outage is fixed at 6% (see Figure 6). Comparison in terms of average system metrics is presented in Table 4.

V. SUMMARY AND CONCLUSIONS

We developed a linear complexity, utility based resource allocation framework for OFDMA systems. This technique is based on local linearization of the utility function given that system performance is measured as a function of average throughput. It is shown that the performance of this low complexity scheme is similar to earlier near-optimal approaches with cubic complexity. We further showed that the low-complexity scheme performs well even if performance is measured in terms of instantaneous throughput. System performance results for Downlink 802.16 system show that around 25% gain in system capacity are possible through efficient allocation of resources tailored to the QoS requirements of each service class. Next steps include extending the performance results to consider both traffic and channel statistics and to explicitly consider delay sensitive traffic.

VI. REFERENCES

[1] Guocong Song, and Ye Li, "Cross-layer optimization for OFDM wireless networks-part I: Theoretical Framework," *IEEE Trans. Wireless Commun., vol. 4, no. 2, pp. 614-624*, March 2005.

[2] Guocong Song, and Ye Li, "Cross-layer optimization for OFDM wireless networks-part II: algorithm development," *IEEE Trans. Wireless Commun., vol. 4, no. 2, pp. 625-634*, March 2005.

[3] Z. Jiang, H. Mason, B. J. Kim, N. K. Shankaranarayanan, and P. Henry, A subjective survey of user experience for data applications for future cellular wireless networks," *in Proc. Symp. Applicat. Internet, San Diego, CA. pp. 167-175*, Jan. 2001.

[4] Z. Jiang, Y. Ge, and Y. (G.) Li, "Max-utility wireless resource management for best effort traffic," *IEEE Trans. Wireless Commun., vol. 4, no. 1, pp. 100-111*, Jan. 2005.

[5] M. Krunz and S. Tripathi, "On the characterization of VBR MPEG stream, *ACM*, *pp. 192-202*, 1997.

[6] Guocong Song and Ye Li, "Utility based resource allocation and scheduling in OFDM based wireless cellular networks," *IEEE Communications magazine, pp. 127-134*, December, 2005.

[7] R. Srinivasan et. al., " Downlink spectral efficiency of Mobile WiMax," *IEEE VTC*, Spring 2007

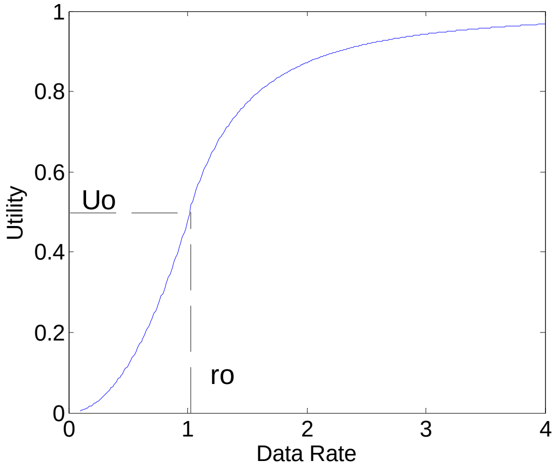

Figure 2: Example utility function for rate adaptive applications. Here r_0=1024 Mbps and U_0=0.

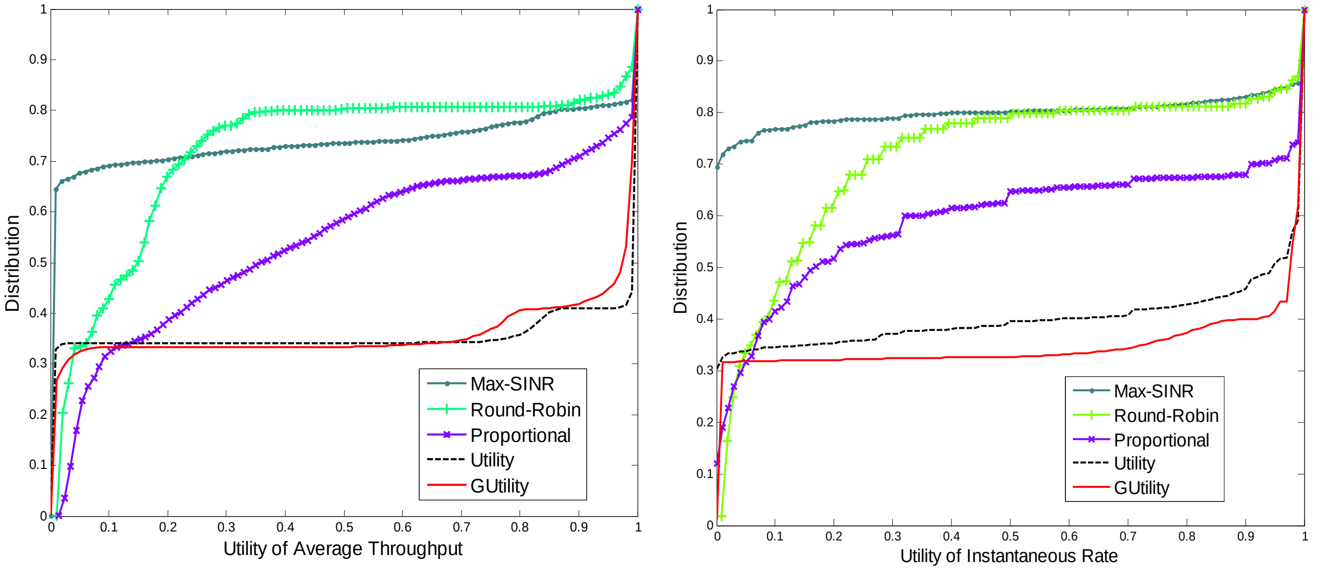

Figure 3: Cumulative distribution function of over-all user utility, when measured as a function of a) average user throughput and b) instantaneous user throughput. 15 users are considered.

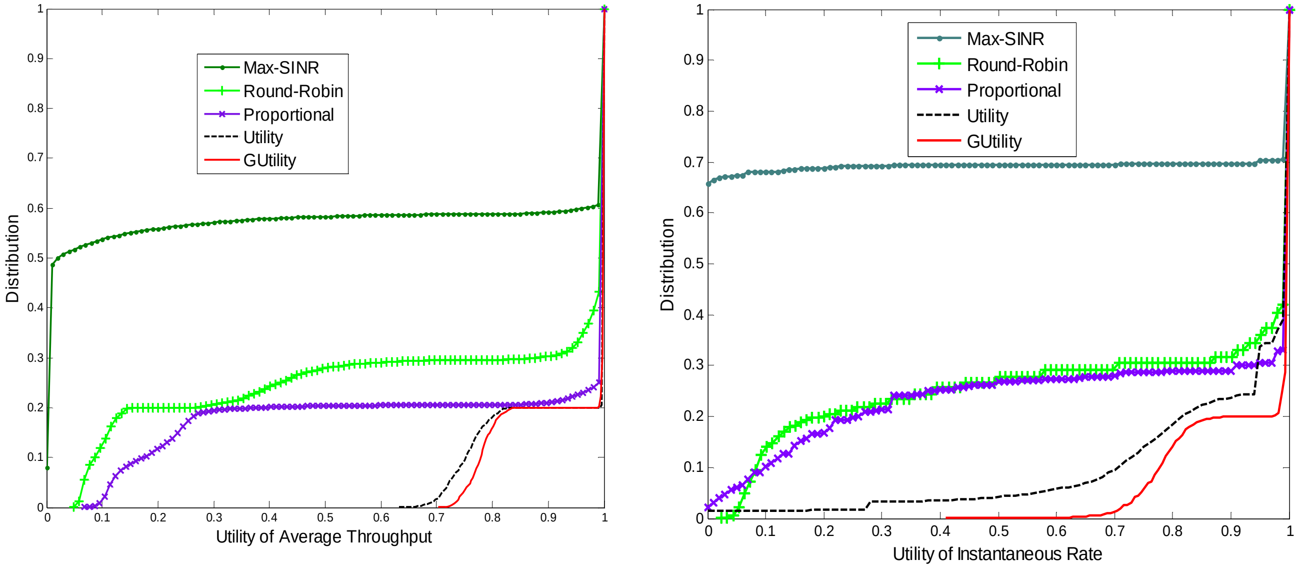

Figure 4: Cumulative distribution function of over-all user utility, when measured as a function of a) average user throughput and b) instantaneous user throughput. 10 users are considered.

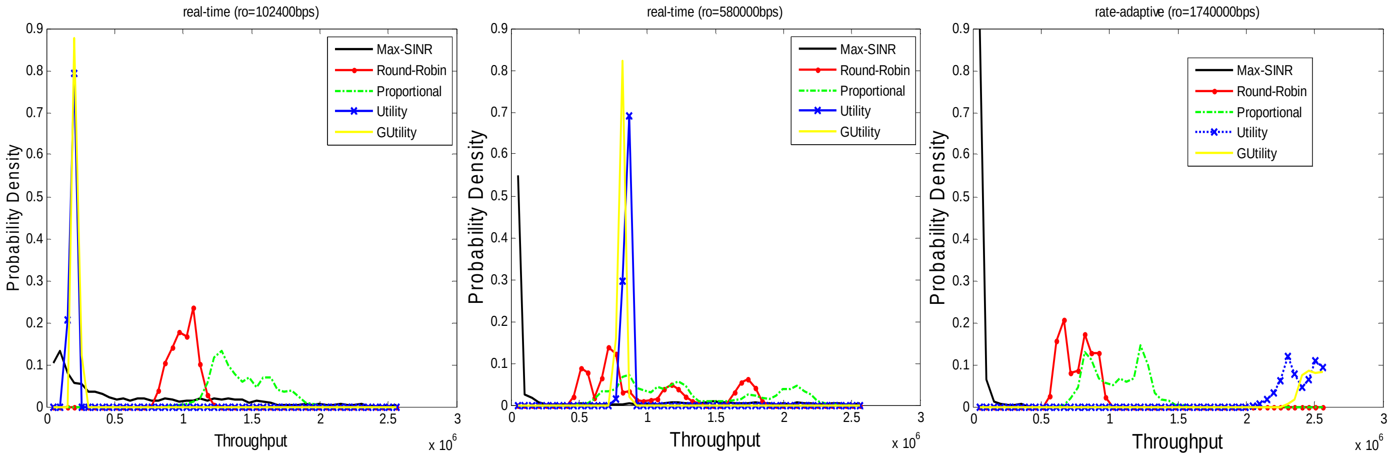

Figure 5: The probability distribution of assigned throughput across various resource allocation schemes a) for real-time service at 100 kbps b) real time services at 580 kbps and c) rate adaptive scheme at 1.74 Mbps. It can be seen that utility based approaches are better at allocating just the required throughput to satisfy the user QOS. Since unnecessary resources are not allocated, these approaches can better satisfy more users when compared to the conventional approaches.

Scheduling algorithm	Average throughput (Mbps)	Average throughput utility	Average rate utility	Outage I (%)	Outage II (%)	Execution Time (ms)
Max-SINR	17.4300	4.2668	3.0788	5.8762	6.9597	0.09
Round Robin	9.1745	7.5734	7.5128	2.9555	3.0598	0.13
Proportional	12.7160	8.3052	7.6676	2.0584	2.7636	0.11
Utility	9.5690	9.4988	9.0843	0	0.4729	0.39
GUtility	9.7408	9.5458	9.5275	0	0.0014	330

Table 4: Comparing average throughput and average utility as a function of average throughput as well as instantaneous throughput across scheduling scheme. Outage measures the percentage of users unable to meet the utility requirements of their QoS class.

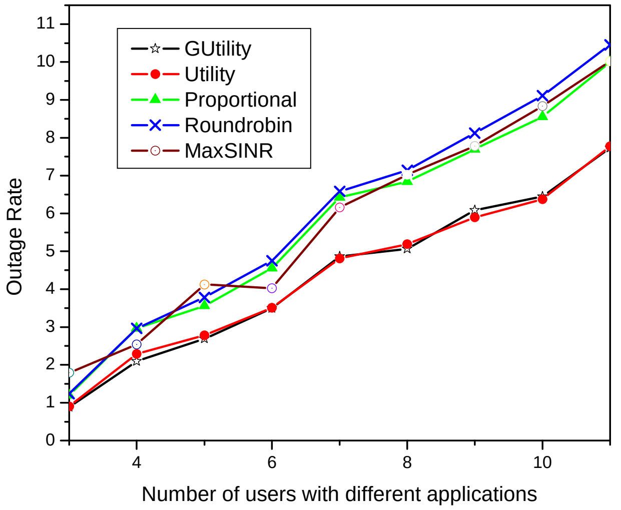

Figure 6: System performance gains with utility based approaches: The outage rate (number of users unable to meet the utility required by their QoS Class) as a function of users. It can be seen that utility based users can admit more users into the system for the same fixed outage level. At the outage of 6%, 28% gain in number of users admitted is observed.