Enabling Trade-offs between System Throughput and Fairness in Wireless Data Scheduling Techniques

Debashis Panigrahi* and Farideh Khaleghi+

*Dept of ECE
University of California
San Diego

+Ericsson Wireless Communications, Inc.
6455 Lusk Blvd.
San Diego, CA

Abstract

Scheduling plays an important role in wireless data communication systems where a common medium is shared to achieve a desired Quality of Service (QoS) for the users as well as the system. The new generation of wireless data systems, such as HRPD, deploy a scheduling algorithm to ensure proportional fairness in midst of varying channel environment. While proportionally fair scheduler attempts to allocate radio resources so as to ensure fairness between users under varying channel conditions, it may not lead to maximum system throughput. Similarly, schedulers designed to ensure maximum system throughput, do not ensure any fairness between users. In this paper, we explore techniques that allow for trade-offs between system throughput and user fairness. We present two schedulers which leads to increase in system throughput without harming the user fairness significantly.

Key Words: Scheduling, 1xEV-DV, Quality of Service (QoS), multi-user diversity, system throughput, and proportional fairness

1. Introduction

With the growing popularity of wireless data applications, new generation systems are being developed for wireless data services. HRPD (High Rate Packet Data) (otherwise known as 1xEV-DO) and 1xEV-DV (1x Evolution Data and Voice) are such new systems that are specifically optimized and targeted for data traffic [1,2,3]. An important feature of the data systems is the scheduling algorithm used to share common radio resource between different data communication sessions. In these systems, the dynamic wireless channel conditions pose a daunting task to enable efficient use of radio resource between users. In order to use the dynamic nature of channels effectively, new class of algorithms, called *multi-user diversity* based algorithms, have been proposed that consider the current channel conditions in selecting appropriate users to schedule for data traffic, hence attempt to achieve the desired QoS for elastic data traffic. HRPD and 1xEV-DV systems provide provisions for implementing such algorithms.

In the HRPD and 1xEV-DV data communication systems, users are served in a time-multiplexed manner in the forward link for data communication. In the above systems, each mobile terminal periodically reports to the Radio Base Station (RBS) the current channel conditions that the terminal is experiencing. Using the current channel condition information, a scheduling algorithm at RBS decides which user (and what data rate) to be scheduled at the next scheduling window. In this framework, scheduling of packet transmissions plays a vital part in order to satisfy the following goals: to maximize *throughput of the network* and to ensure *fair allocation* of resources.

Different scheduling algorithms have been proposed in an attempt to use the random nature of wireless channels so as to increase system throughput as well as achieve desired user fairness. In other words, *multi-user diversity* is realized through scheduling by a data scheduler conscious of channel conditions. Proportional fair scheduling, the notion proposed by Kelly in [4], is one of the most popular scheduling algorithms being used in real systems. Kelly proposed a utility-based optimization problem to derive scheduling decisions so as to maximize throughput of the system and maintain fairness amongst the users.

Utility-based optimization techniques have been very popular for allocating resources fairly in a system where radio resource is shared between multiple users. In these methods, the desire of each participating node is represented by a utility curve that defines the returns (utility) for any resource allocation from user/node point of view. Then the goal of any resource control algorithm is to maximize the aggregate utility value leading to optimal allocation of resources. It has been shown that different values of utility functions achieve different goals [4]. For example, one scheduling algorithm called *MAX C/I scheduling* scheme leads to maximum network throughput, however may not be fair to all users. Similarly, other scheduling algorithms consider user fairness but do not consider network throughput. In this work, we attempt to define two new scheduling algorithms based on the use of different utility functions that allows for trade off between system throughput and user fairness, specifically increase system throughput without significantly affecting the fairness. Our results have shown that we can increase the system throughput significantly (10-15%) without harming fairness.

The rest of the paper is organized as follows. We present the formulation of the problem for wireless data scheduling in the next section and discuss the notion of fairness with respect to the scheduling problem. We describe the simulation environment used to evaluate and compare different scheduling schemes in the section 3. We present experimental results in section 4. Finally, we compare our efforts with previous proposed schemes on CDMA data scheduling and conclude the paper.

2. Scheduling Formulation

In this section, we present a formulation of the scheduling problem in the context of new wireless protocol standards and propose our schedulers. Our formulation is similar to the formulation proposed by Kelly, in order to optimize user and system objectives in a shared medium system. Before we present the formulation of the problem, let us first introduce the variables and assumptions of our formulation.

We consider a wireless system with multiple base stations in which each base station serves M simultaneous active users. In this system, time is divided into periods over which the scheduling decisions are made repeatedly. We assume that the data rates that can be supported by each user will vary over time depending on the channel condition for that user. The feasible rates are typically decided by standards depending on the modulation scheme, coding rate, and slot duration. For example, in case of 1xEV-DV data network the feasible rates range from 76.8 Kbps to 3.1Mbps. We denote the set of feasible rates by $C = \{C_1, C_2, \ldots C_k\}$, where k represents the number of feasible data rates. We also assume each user/session has a weight associated with it to enable differentiated access. Let us denote the weight vector by $W = \{W_1, W_2, \ldots W_m\}$.

We assume that the base station has perfect knowledge of the maximum feasible rate for each user at the start of scheduling window. Based on the channel information, the base station decides a user to schedule as well as the data rate of the transmission. These scheduling decisions are made based on a scheduling algorithm that embodies an underlined fairness algorithm. Let us denote $d_m(n)$ to be the scheduled rate for user m in the n^{th} slot, $c_m(n)$ is the maximum feasible rate for each user, and $x_m(n)$ is a 0-1 variable indicating whether or not the n^{th} slot is assigned to user m. Based on this, we can define average rate of user m as

$$R^i_{avg} = \liminf_{k->\infty} \frac{1}{k} \sum_{n=1}^{k} x_m(n) * d_m(n)$$

The goal of the optimization problem from user fairness perspective is to maximize the sum of utilities of all users, otherwise called *User Objective Function (UOF)*. If the utility curve of user i is represented as U_i^{user}, then the goal is to

Maximize

$$UOF = \sum_{i=1}^{M} W_i * U_i^{user}(R^i_{avg}) \qquad \ldots(1)$$

subject to the following constraints

$$\begin{aligned} & d_m(n) \le c_m(n) \quad where \quad d_m(n), c_m(n) \in C \\ & \sum_{i=1}^{M} x_i(n) = 1 \quad where \quad x_i(n) \in \{0,1\} \qquad (2) \\ & \qquad for\ n = 1\ to\ \infty\ and\ m = 1 to M \end{aligned}$$

Similarly, the goal of the system optimization problem is to maximize the utility of weighted throughput, otherwise called *System Objective Function (SOF)*. In this case, M_i relates to the weight related to the revenue earned by user. The goal of the optimization problem is to

$$maximize\ SOF = U^{sys}\left(\sum_{i=1}^{M} M_i * R^i_{avg}\right) \quad ..(3)$$

under the same constraints as above in equation 2.

In [4], Kelly proposed a similar formulation and demonstrates that Lagrange Methods can be used to find the optimal solution. However, it may not be possible to apply the above analytical results in a real scheduling scenario (as pointed out by [5]) for the following reasons: (1) it is infeasible to predict the future channel behavior accurately, (2) the system is dynamic in terms of number of users and application traffic, and (3) the system can not be moved to optimal rate allocation instantly.

We plan to solve the above optimization function using steepest ascent approach. If we select the user which leads to maximum increase in the objective utility function at each slot (or moving along the direction that would lead to maximum increase in utility function), this can be simplified to an optimization problem over a single slot. In order to compute aggregate utility values (UFO and SFO), the system needs to evaluate average throughput achieved by each user at each scheduling window. Let us simplify the notion of average rate that is meaningful to users and can be computed efficiently. Let **$R_m(n)$** denotes the current average rate at the **n^{th}** slot for each user **m**, and the new average rate is denoted by **$R_m(n+1)$**, where α relates to the time period for which an application can be starved. The new average rate is defined as follows

$$R_m(n+1) = \alpha * R_m(n) + (1-\alpha) * x_m * d_m \quad \ldots(4)$$

Using the above definition of user throughput, let us try to find out the user to be scheduled in order to maximize UOF. For simplicity of expression, we will omit the slot numbers in the following expressions. Let us denote ***d_i*** as the data rate that can be allocated to user ***i*** where ***$1 < i < M$***. So change in UOF if the user ***j*** is scheduled, is given by

$$\begin{aligned}\Delta UOF &= \sum_i U(R_i(n+1)) - \sum_i U(R_i(n)) \\ &= \sum_{i \neq j} [U(\alpha * R_i) - U(R_i)] + U(\alpha * R_j + d_j) - U(R_j) \\ &= \sum_i [U(\alpha * R_i) - U(R_i)] + \frac{U(\alpha * R_j + d_j) - U(\alpha * R_{ji})}{(\alpha * R_j + d_j) - (\alpha * R_{ji})} * d_j \\ &= \sum_i [U(\alpha * R_i) - U(R_i)] + \frac{dU(R)}{dR}\Big|_{R=R_j} * d_j\end{aligned}$$

It can be noted from above that the first part of the expression is same for all users. Hence to select the user with maximum increase in user objective function, the user with maximum value of ***$U'(R_j) * d_j$***, where ***$U'(R)$*** is the derivative of ***$U(R)$***. It is important to note that different utility functions lead to different fairness criteria. For example, if all users follow utility curve of ***log(r)***, *i.e.* for all i, ***$U_i^{user}(r) = \log(r)$***, the system schedules the user with maximum value of ***d_j/R_j***. The above utility curve leads to a notion of proportional fairness, as presented by Kelly. A scheduling algorithm is called proportional fair if the aggregate relative change in resource allocation compared to any other allocation scheme is negative. Assuming X and Y are two allocation vectors representing user throughput of all users, then X is called proportional fair if for all Y, the following equation holds,

$$\sum_{i=1}^{M} \frac{Y_i - X_i}{X_i} \leq 0$$

Similarly, if all users have a utility function of ***r***, *i.e.* ***U_i^{user}* (r) = r,** the system selects the user with maximum value of ***d_j***. This scheduling algorithm is otherwise referred as *Max C/I algorithm*, where resource allocation achieves maximum throughput. In order to achieve appropriate trade off between system throughput and user fairness, we attempt to develop a fair scheduler that attempts to combine the two objectives. In order to find out a common solution for user resource allocation fairness and system throughput, we can combine the above optimization problems in different ways. If **t** represents system throughput, and **R_i** represents throughput for user **i**, the goal is to

Maximize

$$U(t,R) = U^{sys}(t) + K * \sum_{i=1}^{M} W_i * U_i^{user}(R_i)$$

The optimum solution for this optimization problem depends on selection of **U^{sys}** and **U^{user}**. Let us assume **$U^{sys}(t) = t$** and **$U^{user} = \log(r)$**, then following the same analysis as above, a small perturbation of resource allocation by **d_j** for user **j**, leads to a change in utility value proportional to

$$\begin{aligned}&U^{sys\,'}(t) * d_j + K * W_j * U^{user\,'} * d_j \\ &\Rightarrow d_j + K * W_j * \frac{d_j}{R_j}\end{aligned}$$

In steepest ascent approach, the user with maximum value for the above expression is scheduled. Note that, the optimization solution would be the same had we chosen a utility function **$U(r) = r + K * \log(r)$** for the UOF optimization problem. As we presented before, utility function of **$U(r) = r$** leads to *Max C/I scheduling*, that maximizes system throughput, and the utility function of **$U(r) = \log(r)$** leads to *proportional fairness*. Hence the **$U(r) = r + K * \log(r)$** strikes a balance between system throughput and proportional fairness. The parameter **K** can be used to configure the proposed scheduler. When **K=0**, the scheduler becomes *Max C/I scheduler,* and for **K >> 1** the scheduler is nothing but a proportionally fair scheduler. The same argument can be extended for any choice of utility functions for **U^{sys}** and **U^{user}**. In this paper, we will evaluate the performance for **$U(r) = r + K * \log(r)$** only.

In addition to the above formulation, we also investigate another formulation for the combined optimization problem, where we try to maximize the product of two optimization functions, as presented below

Maximize

$$U(t,R) = U^{sys}(t) * \sum_{i=1}^{M} W_i * U_i^{user}(R_i)$$

For **$U^{sys}(t) = t$** and **$U^{user}(r) = \log(r)$,** it can be argued that the optimization choice is analogous to choosing an utility function of **$U(r) = r * \log(r)$**. A parameterized version of the above utility function can be **$U(r) = r^K * \log(r)$,** where $K >= 0$. For K=0, the scheduler becomes a proportionally fair scheduler and for large values of **K** it is equivalent to Max C/I scheduler. We believe the proposed schedulers are configurable whose parameters can be set online by service provider depending on their requirements.

2.3 Fairness and Quality-of-service

Fairness is an important criterion in a system where resource is shared between multiple users. Achieving fairness in scheduling ensures that each participating user gets equal allocation of resource in the long run, and prevents any user from starving. Several measures of fairness have been proposed in literature so far, for example, min-max fairness and proportional fairness, *etc.*

We will use the fairness metric used by 3GPP2 to evaluate different scheduling algorithms for our comparative evaluation. In this definition of fairness, a cumulative probability distribution of normalized user throughput (with respect to average throughput or maximum user throughput) is plotted. According to 3GPP2 specifications, a fair scheduler's CDF plot of normalized throughput should lie to the right of a pre-prescribed line of reference. In order words, it ensures that the percentage of the users having very low data rate compared to average data rate should not go above a threshold value. In this definition of fairness, the data rate achieved by each user is used as a QoS measure for each user applications. This notion of fairness metric can be easily extended to other types of QoS metrics depending on application requirements, such as data latency, jitter *etc.*

In addition to the fairness from user’s perspective, the total system throughput achieved can be thought of as a metric of fairness in network provider's perspective. In order to evaluate our proposed schedulers with others, we developed a system simulation environment to simulate wireless data communication very similar to the recent generation of wireless protocol standards. In the next section, we present some salient features of the developed simulation environment.

3. Simulation Environment

Before we present our experimental results, we want to present the simulation environment developed for evaluating different scheduling schemes. We developed a MATLAB based system simulation environment to perform comparative evaluation of the proposed schedulers. First, we present the overall system architecture and mobility patterns used in our simulation environment. Then, we talk about the traffic models used for our evaluation. Finally, we briefly talk about a rudimentary modulation and coding scheme selection procedure used in our comparative study, similar to the one presented in 1xEV-DV standard.

System Layout

The system consists of seven 3-sectored cells. The 21 sectors in total are wrapped around in a hexagonal geometry. The Modified Hata Urban Propagation Model at 1.9GHz (COST231) is used in the simulation with lognormal shadowing for modeling channel in our simulation environment. A minimum separation between MS and BS and a maximum path-loss were applied. We use two different types of channel models, i.e. pedestrian model and vehicular model. Mobiles are moved randomly guided by an average speed and acceleration. For pedestrian scenario, we used an average speed of 3km/hr where as for vehicular scenario we used an average speed of 30km/hr.

Traffic Models

For our experimental evaluation, we used two types of traffic, *i.e.* IP traffic and HTTP traffic for web applications. In IP traffic model, each object size is constant or based on exponential probability distribution function. The inter-arrival packet time is an exponential function. Similarly, in HTTP traffic model, each object size lognormal distributed and inter-arrival time is based on exponential distribution function.

Modulation Coding Scheme Selection

Adaptive modulation and coding is recently being used to respond to varying channel condition in new generation wireless standards. We simulated a modulation and coding scheme selection procedure in our simulation environment which is very similar to the one proposed to be used for 1xEVDV communication standards. The goal of the selection process is to select a transmission format to best suit the current channel conditions, *i.e.* to choose a 4-tuple information (Encoder Packet Size, Modulation Scheme, Number of Slots, No of Codes). There are six information payloads possible (Encoder Packet Size): *i.e.* 384, 768, 1536, 2304, 3072, and 3840 bits. Each encoded payload can be carried over 1, 2 or 4 slots yielding three different data rates. There are three different modulation schemes that are allowed in the current framework, i.e. QPSK, 8-PSK, and 16-QAM. For each possible combination of Encoder Packet Size, Modulation Scheme and Number of Slots, we compute the required Eb/No to achieve a frame error rate of 1% and later use it to select appropriate modulation and coding scheme given the current channel condition. In our simulations, we assume all 28 codes are used for data communication.

Each mobile reports the current received channel quality information for the downlink pilot channel. The channel quality information is used (with a 3-slot delay) for deciding appropriate transmission format. The Channel Quality Information (CQI) needs to be scaled appropriately to correspond to data channels. The received CQI value is translated to Eb/No. Then, the received Eb/No is compared with required Eb/No for each possible configuration as presented above. Then the combination with maximum data rate satisfying the required Eb/No is selected for current transmission.

Next, we present the comparative evaluation of our proposed scheduler with a proportionally fair scheduler.

4. Experimental Results

In this section, we present our experimental results comparing our proposed scheduler with the proportionally fair scheduler. We compare in terms of fairness criteria used by 3GPP and system throughput.

In Figure 1, we plot cumulative probability distribution of normalized user throughput with respect to average user throughput. The red line on the plot is called STRAWMAN plot, which defines the fairness requirement. A scheduler is fair if the entire plot lies to the right of STRAWMAN plot. It can be noted from the plot our proposed schedulers are very close to the proportionally fair scheduler.

Next, we compare the normalized system throughput achieved in the proposed scheduler in comparison to normal scheduler. As it can be seen from the plot (figure 2), our proposed scheduler fairs better than proportionally fair scheduler (an improvement of 13%). The scheduler with utility function of U(r) = r + log(r) fairs marginally better than U(r) = r*log(r).

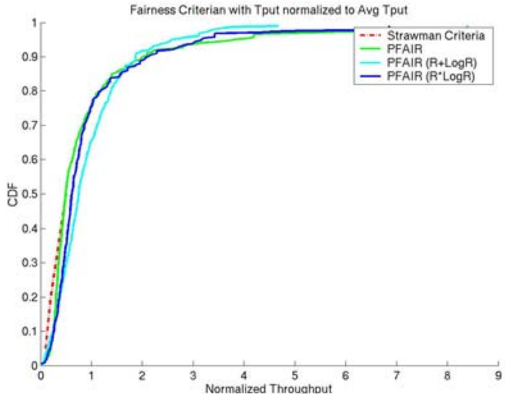

Figure 1 *Cumulative Distribution Function of Normalized Throughput: A Fairness Criteria*

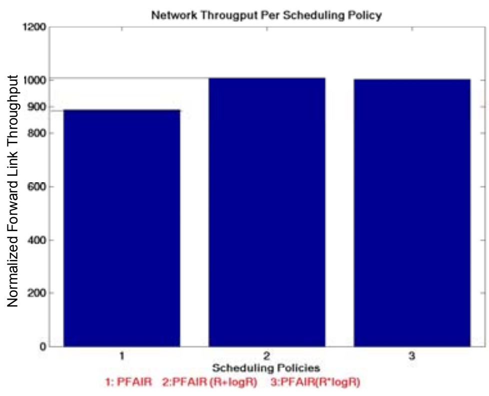

Figure 2 Normalized System Throughput

5. Conclusions and Future Work

In this study we presented a configurable scheduler that performs a trade-off between system throughput and user fairness. In future work, we plan to consider different application QoS requirements apart from user throughput in our scheduling policy. We are planning to explore the possibilities of implementing similar scheduling algorithm on the reverse link also.

6. References

[1]Paul Bender, P. Black, M. Grob N. Sindhushayana, and Andrew Viterbi, "CDMA/HDR: A Bandwidth Efficient High-Speed Wireless Data Service for Nomadic Users", *IEEE Communications Magazine, 70-77*, July 2000.

[2] 1xEV-DO Airlink Overview, Qualcomm Inc., http://www.qualcomm.com/main/whitepapers/1xV_AirlinkOverview_110701.pdf, 2001

[3] TIA/TIA/IS-2002.2-C, *Physical Layer Standard for cdma2000 Spread Spectrum Systems*, June 2002.

[4] Frank Kelly, "Charging and Rate Control for Elastic Traffic", *European Transactions of Tele communications*, vol. 8, pp. 33-37, 1998

[5] Patrick A. Hosein, "A Generalized Scheduling Algorithm for HRPD Wireless Networks", *Proceedings of the 2002 IASTED Conference on Wireless and Optical Communications*, Canada, July 2002.