Design and Analysis of Downlink Utility-Based Schedulers *

Randall A. Berry, Peijuan Liu and Michael L. Honig

Dept. of Electrical and Computer Engineering
Northwestern University
Evanston, IL 60208
rberry,peijuan,mh@ece.northwestern.edu

Abstract

We consider scheduling data traffic for the downlink of a wireless network. A draining problem is formulated where the goal is to transmit a given set of packets. Each packet is assigned a utility that depends on the delay incurred. We propose a simple gradient-based scheduling rule which attempts to maximize the average utility per packet. A deterministic analysis of this rule is given by considering an asymptotic fluid limit where the number of packets becomes large while the packet-size decreases to zero. In this limiting regime, we formulate an optimal control problem which corresponds to finding the best scheduling policy. Using Pontryagin's minimum principle, we prove that in a special case, the gradient-based algorithm is optimal. Simulations are presented to illustrate these results.

1 Introduction

Transmission scheduling is an important component of an efficient wireless data service. Protocols which make scheduling decisions based in part on the channel quality of each user have recently attracted much interest, see for example [2, 3, 4]. These protocols seek to exploit variations in channel quality across the user population to improve overall performance. Such channel-aware schedulers are part of several recent standards such as 1xEV-DO (HDR) [1].

A key issue in the design of channel-aware scheduling algorithms is balancing the total throughput with other performance metrics of interest. For example, in many cases, the total throughput can be maximized by scheduling only the users with the best channel quality. However this approach can be unfair and lead to long delays for other users. To capture such considerations, we consider a utility-based scheduling framework. In particular, we assume that a utility function is associated with each packet. This function indicates the benefit of receiving that packet after a specific delay. The goal of the scheduler is to maximize the total utility summed over all packets. To accomplish this goal, we present a simple gradient-based scheduling algorithm, which we call the "$\dot{U}R$ scheduler".

*This research is supported in part by the Motorola-Northwestern Center for Telecommunications, and by NSF under grant CCR-9903055.

We study the performance of the $\dot{U}R$ scheduler for a draining problem, where the goal is to transmit an initial set of packets while maximizing the total utility. To analyze this problem, we consider a type of fluid limit. In this limit, the dynamics of the scheduling algorithm are given by a set of deterministic differential equations. Simulations are presented which show that the performance of the limiting system accurately predicts that of a finite system. For the limiting system, we then show that the problem of finding the optimal scheduling algorithm can be formulated as a continuous-time optimal control problem. For a special case of this problem, it is shown that the gradient-based algorithm is optimal.

2 System Model

We formulate a simple model for downlink scheduling from a single transmitter, such as a base station in a cellular network or an access point in a wireless LAN. Our focus is on a system where the transmitter sends to one user at a time, as in the HDR standard. However much of the following can be easily extended to the case where multiple users may be scheduled at a time. To simplify our discussion we consider a system with 2 classes of packets; each class corresponds to a different feasible transmission rate.[1] Specifically, for $i = 1, 2$, the base station can transmit class i packets with transmission rate R_i, where $R_1 > R_2$. We also assume that each packet's class is fixed over the time-scale of interest; this assumption is reasonable in a slow fading environment and serves to highlight the possible disparity between classes of users.[2]

Each packet is assumed to contain L bits including any overhead. In the draining problem we consider, there is an initial set of packets given for each class and no new arrivals occur. The system is to be emptied by transmitting all of the packets. For simplicity we assume that there are N packets of each type. The total time required to drain the system is given by

$$T_f = \frac{NL}{R_1} + \frac{NL}{R_2}. \tag{1}$$

This is independent of the order in which packets are served and only requires that the transmitter be non-idling, i.e. that it always transmits a packet if one is available. However, the order in which packets are served does influence the delay incurred by the individual packets. We assume that the delay preferences associated with each packet are indicated by a utility function. The goal is then to schedule the packets to drain the system and maximize the utility per packet.

We assume that each packet has an initial delay at time $t = 0$. This reflects the delay experienced by the packets prior to time $t = 0$ and could include, for example, the delay incurred in forwarding the packet by other nodes in an *ad hoc* network. For $k = 1, \ldots, N$, we denote the initial delay of the kth packet of class i by $W_{i,k}(0)$. If this packet is transmitted after t seconds, then the total delay incurred by the packet is given by

$$D_{i,k} = W_{i,k}(0) + t + \frac{L}{R_i},$$

[1] For the problem considered here all of the packets in a given class can be directed to one user or several users with similar channels.

[2] Note in this setting issues of "opportunistic" scheduling do not arise [5].

where t accounts for the aging of the packet with time and the last term is the packet's transmission time.

The utility associated with each class i packet served is given by $U_i(D_{i,k})$, where $U_i(\cdot)$ is assumed to be non-increasing. The utility per packet generated by a given schedule is

$$U_{avg} = \frac{1}{2N} \sum_{k=1}^{N} \left[U_1(D_{1,k}) + U_2(D_{2,k}) \right].$$

Notice that this depends on the initial delays for the packets in each class.

For a given initial delay distribution, a schedule of packet transmissions is defined to be optimal if it maximizes U_{avg}. Consider the special case where $U_i(x) = -x$ for $i = 1, 2$, and thus maximizing U_{avg} corresponds to minimizing the average delay per packet. In this case, the optimal schedule is to transmit all packets of class 1 before any packets in class 2; within each class, the order in which packets are transmitted does not effect U_{avg}. This can be shown using a simple interchange argument. Next suppose that the utility $U_i(\cdot)$ is strictly concave for each i. Then it can be shown that the optimal schedule must have the property that packets within each class are transmitted in a longest delay first order, i.e. if $W_{i,k}(0) > W_{i,\tilde{k}}(0)$ then packet k will be transmitted before packet $\tilde{k}$. Even with this characterization, there are still more than $O(2^N)$ possible schedules.

Instead of finding the optimal schedule, we consider a simple gradient-based scheduling policy. This policy attempts to schedule a packet from the class which results in the largest first order change in the total utility rate. For $i = 1, 2$, let D_i denote the longest delay of the remaining class i packets at a given scheduling time. If the scheduler transmits a packet in class 1, followed by a packet in class 2, the derived utility can be written as

$$\Delta U_{1,2} = U_1(D_1 + \tfrac{L}{R_1}) + U_2(D_2 + \tfrac{L}{R_1} + \tfrac{L}{R_2}).$$

Approximating $U_i(x)$ by a first order Taylor series around D_i we have

$$\Delta U_{1,2} \approx U_1(D_1) + \dot{U}_1(D_1)\tfrac{L}{R_1} + U_2(D_2) + \dot{U}_2(D_2)\left(\tfrac{L}{R_1} + \tfrac{L}{R_2}\right).$$

Likewise, transmitting in the reverse order yields

$$\Delta U_{2,1} \approx U_1(D_1) + \dot{U}_1(D_1)\left(\tfrac{L}{R_1} + \tfrac{L}{R_2}\right) + U_2(D_2) + \dot{U}_2(D_2)\tfrac{L}{R_2}.$$

Based on these expressions, we define the "$\dot{U}R$ scheduling policy" to be a policy which selects a packet from class i if $\Delta U_{i,j} > \Delta U_{j,i}$ for $j \neq i$. This scheduling rule can be written compactly as follows:

$\dot{\mathbf{U}}$R scheduling rule: *schedule user i^* such that*

$$i^* = \arg\max_i |\dot{U}_i(D_i)| R_i,$$

where ties can be broken arbitrarily.

In the special case of linear utilities we have:

Proposition 1 *If $U_i(x) = -\beta_i x$ for $i = 1, 2$ and $\beta_i > 0$, then the $\dot{U}R$ scheduling rule maximizes the utility per packet.*

Several other scheduling policies proposed in the literature can be viewed as special cases of this policy, corresponding to particular choices of utilities. For example, if $U_i(D) = -\beta_i D^2$ for all i, then the $\dot{U}R$ scheduler is equivalent to the Modified Largest Weighted Delay First rule proposed in [2].

3 Fluid Limit

To analyze the performance of a scheduling policy for the draining problem, we consider a type of fluid limit for the system. In this section, we describe this limit for an arbitrary scheduling rule; in the next section, we consider the limiting behavior of the $\dot{U}R$ scheduling rule.

We scale up the number of packets and decrease the packet size, while keeping a fixed load (in bits). Formally, we consider a sequence of systems indexed by $N = 1, 2, \ldots$; in the Nth system there are initially N packets of each type with packet length L normalized so that $NL = 1$.[3] With this scaling, T_f in (1) is given by $\frac{1}{R_1} + \frac{1}{R_2}$ for all N. As noted previously, the performance of a scheduler depends on the initial packet delays. For each class i, we assume that $\{W_{i,k}(0)\}_{k=1}^{\infty}$ is a sequence of i.i.d. random variables, with complementary distribution function $F_i(w) = \Pr(W_{i,k}(0) \leq w)$. The first N components of this sequence are the initial delays in the Nth system.

Let $N_i^N(t)$ denote the number of type i packets remaining at time t in the Nth system (for a given scheduling policy). Let

$$f_i^N(t) = \frac{N_i^N(t)}{N}$$

be the fraction of the initial type i packets remaining at time t. Likewise, let $\tau_i^N(t)$ denote the amount of time in $[0, t)$ during which the transmitter serves packets from class i. Between times t and $t + \delta t$, the change in $f_i^N(t)$ can be bounded as

$$\frac{-(\tau_i^N(t+\delta t) - \tau_i^N(t))\frac{R_i}{L}}{N\delta t} \leq \frac{f_i^N(t+\delta t) - f_i^N(t)}{\delta t} \leq \frac{-(\tau_i^N(t+\delta t) - \tau_i^N(t))\frac{R_i}{L} + 1}{N\delta t}. \quad (2)$$

For a finite N, the preceding quantities depend on the initial delay and hence are random. For the scheduling policies of interest, we assume that as $N \to \infty$, $\tau_i^N(t)$ converges almost surely to a deterministic limit $\tau_i(t)$.

As $N \to \infty$, $L \to 0$; therefore, from (2) it follows that $f_i(t) = \lim_{N\to\infty} f_i^N(t)$ exists and satisfies

$$\frac{f_i(t+\delta t) - f_i(t)}{\delta t} = \frac{-(\tau_i(t+\delta t) - \tau_i(t))R_i}{\delta t}.$$

Next, letting $\delta t \to 0$, we have

$$\dot{f}_i(t) = -\alpha_i(t) R_i,$$

where $\alpha_i(t) = \dot{\tau}_i(t)$. We note that both $f_i(t)$ and $\tau_i(t)$ are monotonic functions of t and hence the preceding derivatives exist except possibly on a set of measure zero [7].

In the limit, the base station can transmit arbitrarily many packets in any time interval $[t, t+\delta t)$, but only a finite fraction of the initial packets, given by

$$\int_{[t,t+\delta t)} -\dot{f}_i(t)\, dt = \int_{[t,t+\delta t)} \alpha_i(t) R_i\, dt.$$

The quantity $\alpha_i(t)$ can be interpreted as the fraction of the base station's resources spent on class i packets at time t. If $\alpha_i(t) = 1$, then only class i packets are served. In general, $\alpha_i(t)$ can take on any value in $[0, 1]$ and must satisfy $\sum_i \alpha_i(t) \leq 1$ for each time t. For a non-idling system, this inequality is met with equality for all $t < T_f$.

[3] There is no loss in generality in assuming that the product NL is normalized to 1.

As an example of the preceding scaling, consider a round robin scheduler that alternates between scheduling a type 1 packet and a type 2 packet. In this case, for the Nth system we have

$$\left(\frac{t}{\frac{L}{R_1}+\frac{L}{R_2}}-1\right)\frac{L}{R_1} \leq \tau_1^N(t) \leq \left(\frac{t}{\frac{L}{R_1}+\frac{L}{R_2}}+1\right)\frac{L}{R_1}.$$

Hence, as $N \to \infty$, $\tau_1^N(t)$ converges to $\tau_1(t)$ given by

$$\tau_1(t) = \frac{R_2 t}{R_1 + R_2},$$

so that $\alpha_1(t) = \frac{R_2}{R_1+R_2}$ and $\alpha_2(t) = \frac{R_1}{R_1+R_2}$.

Next, we turn to the packet delays in the limiting system. For a given realization of $\{W_{i,k}(0)\}_{k=1}^{\infty}$, let $F_i^N(w)$ denote the empirical (complementary) distribution of the initial delays for type i packets in the Nth system, i.e.

$$F_i^N(w) = \frac{|\{k \leq N : W_{i,k}(0) \leq w\}|}{N},$$

where $|\mathcal{X}|$ denotes the cardinality of the set $\mathcal{X}$. As $N \to \infty$, the Glivenko-Cantelli theorem [6] implies that almost surely, $F_i^N(w) \to F_i(w)$ uniformly in w.

Let $D_i^N(t)$ denote the maximum delay of the type i packets in the Nth system at time t. We assume that under all scheduling policies of interest, packets of a given class are served in the order of longest delay first. In this case,

$$D_i^N(t) = G_i^N(f_i^N(t)) + t, \tag{3}$$

where $G_i^N(f) = \max\{w : F_i^N(w) \leq f\}$. The first term in (3) corresponds to the maximum initial delay of the remaining packets; the second term corresponds to the aging of the packets with time. It follows that in the limiting system, almost surely we have

$$D_i(t) = G_i(f_i(t)) + t,$$

where $D_i(t)$ denotes the maximum initial delay of the remaining packets in the limiting system and $G_i(f) = \max\{w : F_i(w) \leq f\}$.

Note that in a finite system, the functions $G_i^N(f)$ and $D_i^N(t)$ are random quantities that depend on the initial delay distribution. However, in the limiting system, these quantities are deterministic.

In the following, to simplify our analysis we focus on the special case where the initial delays are uniform on $[0, 1]$ (for both classes), i.e.,

$$F_i(w) = \begin{cases} w & 0 \leq w \leq 1, \\ 1 & w \geq 1. \end{cases}$$

In this case

$$D_i(t) = f_i(t) + t, \tag{4}$$

and therefore

$$\dot{D}_i(t) = -\alpha_i(t) R_i + 1,$$

with $D_i(0) = 1$.

In the Nth system, if a packet of class i is served at time t, then it receives a utility $U_i(D_i^N(t) + \frac{L}{R})$. The average utility per packet can be written as

$$U_{avg}^N = \sum_{k=1}^{N} \frac{1}{2N} U_1(D_1^N(t) + \frac{L}{R_1}) + \sum_{k=1}^{N} \frac{1}{2N} U_2(D_2^N(t) + \frac{L}{R_2})$$

As $N \to \infty$ we have $U_{avg}^N \to U_{avg}$, where

$$U_{avg} = \frac{1}{2} \int_0^{T_f} \left[\alpha_1(t) R_1 U_1(D_1(t)) + \alpha_2(t) R_2 U_2(D_2(t))\right] dt.$$

4 Limiting Behavior of $\dot{U}R$ Scheduler

In this section we consider the limiting performance of the $\dot{U}R$ scheduler with two classes of packets. We also make the following simplifying assumptions: (1) The initial delay for each class has a uniform delay distribution; and (2) Both classes have the same utility function $U(D)$.

In the limiting system, it is straightforward to see that the $\dot{U}R$ rule sets

$$\alpha_1(t) = \begin{cases} 1 & \text{if } |\dot{U}(D_1(t))|R_1 > |\dot{U}(D_2(t))|R_2 \text{ and } f_1(t) > 0, \\ 0 & \text{if } |\dot{U}(D_1(t))|R_1 < |\dot{U}(D_2(t))|R_2 \text{ or } f_1(t) = 0 \end{cases}$$

and $\alpha_2(t) = 1 - \alpha_1(t)$. This specifies the scheduling rule except at some time t where $f_1(t) > 0$ and $\dot{U}(D_1(t))R_1 = \dot{U}(D_2(t))R_2$. In that case, to determine the behavior of the $\dot{U}R$ rule, we make the additional assumption that $U(D)$ is concave and $\alpha_i(t)$ is right continuous. With these assumptions, the limiting version of the $\dot{U}R$ scheduling rule is given by the following lemma.

Lemma 1 *Assume $U(D)$ is concave and $\alpha_i(t)$ is right continuous. If $\dot{U}(D_1(t))R_1 = \dot{U}(D_2(t))R_2$ for some t such that $f_i(t) > 0$, then as $N \to \infty$, the $\dot{U}R$ rule gives*

$$\alpha_1(t) = \frac{\ddot{U}(D_1(t))R_1 - \ddot{U}(D_2(t))R_2 + \ddot{U}(D_2(t))R_2^2}{\ddot{U}(D_1(t))R_1^2 + \ddot{U}(D_2(t))R_2^2}$$

and $\alpha_2(t) = 1 - \alpha_1(t)$.

Proof: Let $W(t) = \dot{U}(D_1(t))R_1 - \dot{U}(D_2(t))R_2$. Taking the derivative gives

$$\dot{W}(t) = \ddot{U}(D_1(t))R_1(-\alpha_1(t)R_1 + 1) - \ddot{U}(D_2(t))R_2(-\alpha_2(t)R_2 + 1).$$

Setting $\dot{W}(t) = 0$ and using $\alpha_1(t) + \alpha_2(t) = 1$ gives the preceding $\alpha_i(t)$. With this choice, $W(t^+) = 0$ and the corresponding $\alpha_i(t)$ is right continuous. Notice that

$$\frac{d\dot{W}(t)}{d\alpha_1(t)} = -\ddot{U}(D_1(t))R_1^2 - \ddot{U}(D_2(t))R_2^2 > 0.$$

Hence, if $\tilde{\alpha}_1(t) > \alpha_1(t) > 0$, then $\dot{W}(t) > 0$; this implies that $W(t^+) > 0$ and hence $\alpha_1(t^+) = 0$. This violates the right continuity of $\alpha_1(t)$. Likewise, if $\tilde{\alpha}_1(t) < \alpha_1(t) < 1$, then $\alpha_1(t^+) = 1$, which also violates the right continuity assumption. ■

The following two results characterize the behavior of the $\dot{U}R$ scheduler over time. We omit the proofs due to space considerations.

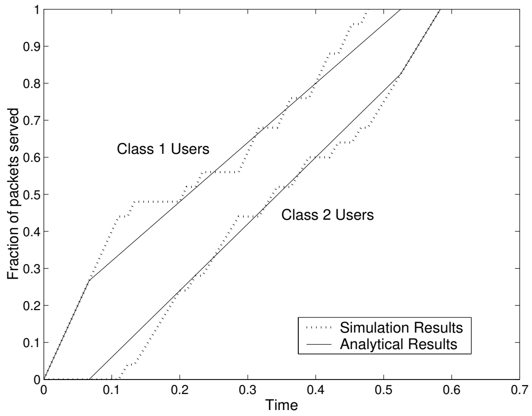

Figure 1: Fraction of packets served vs. time.

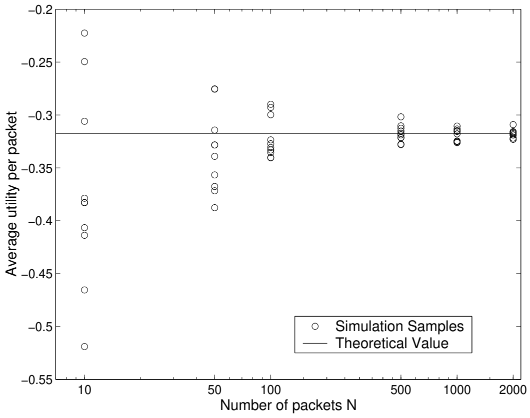

Figure 2: Average utility per packet vs. number of packets, N.

Proposition 2 *If $U(D)$ is concave and $\alpha_i(t)$ is right continuous, then there exists T_1 and T_2 with $0 < T_1 \leq T_2 < T_f$ such that the $\dot{U}R$ gives:*

1. *$\alpha_1(t) = 1$ for all $t \in [0, T_1)$,*
2. *$\alpha_1(t)$ is given by lemma 2 for all $t \in [T_1, T_2)$,*
3. *$\alpha_1(t) = 0$ for all $t \in [T_2, T_f]$,*

and $\alpha_2(t) = 1 - \alpha_1(t)$ for all t.

Corollary 1 *Let $\hat{T}$ be the smallest $t > 0$ such that*

$$\dot{U}(1 + (1 - R_1)t)R_1 = \dot{U}(1 + t)R_2.$$

In Prop. 1, if $\hat{T} \geq 1/R_1$, then $T_1 = T_2 = 1/R_1$, otherwise, $T_2 > T_1 = \hat{T}$.

4.1 Numerical Example

Here we consider the specific utility function $U(D) = -\frac{1}{2}D^2$. In this case, $\alpha_1(t)$ in Lemma 2 is given by

$$\alpha_1(t) = \frac{R_1 - R_2 + R_2^2}{R_1^2 + R_2^2}.$$

Notice that the right-hand side of this expression does not depend on t. Hence, in this case the split of transmitter resources between the two classes is fixed over the interval $[T_1, T_2)$. If $R_1 = 4$ and $R_2 = 3$, then frLemma 3 and Prop. 1, it follows that the scheduler first serves class 1 packets up to time $T_1 = \frac{R_1 - R_2}{R_1^2 - R_1 + R_2} = 1/15$. Then the scheduler drains the two classes simultaneously with $\alpha_1 = 2/5$ and $\alpha_2 = 3/5$. At time $T_2 = \frac{R_2(R_1+R_2)}{R_1(R_1-R_2+R_2^2)} = 21/40$, the scheduler finishes serving all the class 1 packets and starts to serve class 2 solely until $T_f = 7/12$ when all packets are drained. This is illustrated in Fig. 1, which shows the fraction of class 1 and class 2 packets served up to time t. The solid lines correspond to the asymptotic results, where the slope of each line at time t is $\alpha_i(t)R_i$. The dashed lines are from a sample run with $N = 25$ packets.

To study how well the asymptotic results predict the performance of a finite system, we simulated the $\dot{U}R$ scheduler for different numbers of packets, N. The simulation

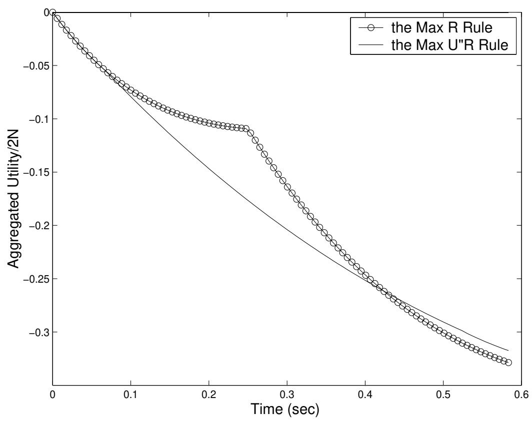

Figure 3: Aggregated utility/$2N$ vs. time for Max R and $\dot{U}R$ rules.

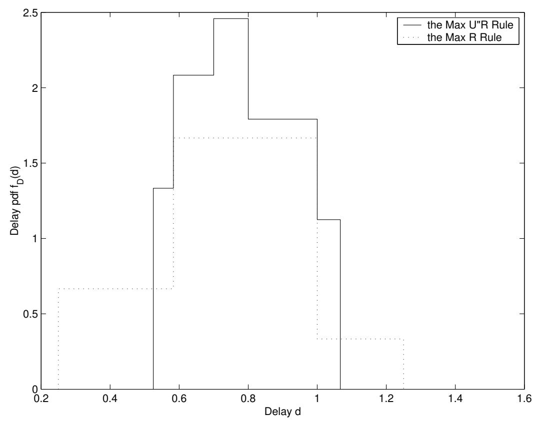

Figure 4: Delay p.d.f.'s for Max R and $\dot{U}R$ rules.

results are shown in Fig. 2, where the x-axis denotes N and each point corresponds to a sample run. As expected, the variance of the utility decreases, and the utilities approach the fluid limit as N increases.

Next we compare the $\dot{U}R$ rule with a policy which gives strict priority to packets in class 1, and within each class still transmits packets in the order of longest delay first. We refer to this as the "Max R" rule. Fig. 3 shows the aggregated utility vs. time under both the Max R and $\dot{U}R$ policies. The $\dot{U}R$ rule generates greater utility over all packets than the Max R rule. Initially the Max R scheduler gives higher utility since it serves only the class 1 packets at the highest rate. The Max R utility then drops below the $\dot{U}R$ utility once the longer delays experienced by class 2 packets dominate.

Fig. 4 shows the asymptotic delay p.d.f. for both the Max R and $\dot{U}R$ rules. These results show that the delay for the $\dot{U}R$ scheduler has a significantly smaller variance than for the Max R scheduler. This may be a desirable feature when considering fairness.

5 Optimal Control Formulation

In this section, we characterize an asymptotically optimal scheduling policy by optimizing $\alpha_i(t)$ for $t \in [0, T_f]$. Formally this problem can be written as

$$\begin{aligned}
\min_{\alpha_1(t),\alpha_2(t)} \quad & \int_0^{T_f} \left[-\alpha_1(t)R_1U(D_1(t)) - \alpha_2(t)R_2U(D_2(t))\right] \, dt && (5)\\
\text{subject to:} \quad & \dot{D}_i(t) = -\alpha_i(t)R_i + 1, \; i = 1, 2, && (6)\\
& D_i(0) = 1 \text{ and } D_i(T_f) = T_f, \; i = 1, 2 && (7)\\
& \alpha_1(t) + \alpha_2(t) = 1, && (8)\\
& \alpha_i(t) \geq 0, \; i = 1, 2. && (9)
\end{aligned}$$

This is a continuous-time optimal control problem [8], where the state is $\mathbf{D}(t) = (D_1(t), D_2(t))$ and $\boldsymbol{\alpha}(t) = (\alpha_1(t), \alpha_2(t))$ is the control variable. Here (6) represents the system dynamics, and (7) gives initial and final boundary conditions for the state.

If all the packets of class i are emptied at time $\hat{t} < T_f$, then for all $t > \hat{t}$, we have that $\alpha_i(t) = 0$ and $D_i(t) = t$. To see that this must hold in the preceding formulation, note that since $f_i(\hat{t}) = 0$, (4) implies

$$D_i(\hat{t}) = \hat{t}.$$

Thus, if $D_i(T_f) = T_f$, then $\alpha_i(t) = 0$ for all $t > \hat{t}$.

The solution to this problem can be characterized using the Pontryagin minimum principle [8]. We first define the Hamiltonian for this problem, which is given by

$$\begin{aligned} &H(\mathbf{D}(t), \boldsymbol{\alpha}(t), \mathbf{p}(t)) \\ &\quad = -\alpha_1(t)R_1\left(U(D_1(t)) + p_1(t)\right) - \alpha_2(t)R_2\left(U(D_2(t)) + p_2(t)\right) + p_1(t) + p_2(t) \end{aligned}$$

where $\mathbf{p}(t) = (p_1(t), p_2(t))$ is the costate or Lagrange multiplier. Let $\boldsymbol{\alpha}^*(t)$ be an optimal control and $\mathbf{D}^*(t)$ the corresponding optimal state trajectory. According to the Pontryagin minimum principle, there exists a $\mathbf{p}^*(t)$ such that

$$\dot{\mathbf{p}}^*(t) = -\nabla_{\mathbf{D}} H(\mathbf{D}^*(t), \boldsymbol{\alpha}^*(t), \mathbf{p}^*(t)) \tag{10}$$

and

$$H(\mathbf{D}^*(t), \boldsymbol{\alpha}^*(t), \mathbf{p}^*(t)) \leq H(\mathbf{D}^*(t), \boldsymbol{\alpha}(t), \mathbf{p}^*(t)) \tag{11}$$

for all admissible controls $\boldsymbol{\alpha}(t)$.

For this problem, (10) yields,

$$\dot{p}_i(t) = \alpha_i(t)R_i\dot{U}(D_i(t)), \ i = 1, 2.$$

Let $A_i(t) = R_i\left(U(D_i(t)) + p_i(t)\right)$ for $i = 1, 2$. Then the Hamiltonian can be written as

$$H(\mathbf{D}(t), \boldsymbol{\alpha}(t), \mathbf{p}(t)) = -A_1(t)\alpha_1(t) - A_2(t)\alpha_2(t) + p_1(t) + p_2(t)$$

which is linear in $\alpha_i(t)$. Hence to satisfy (11), it follows that

$$\alpha_1^*(t) = \begin{cases} 1 & \text{if } A_1(t) > A_2(t) \\ 0 & \text{if } A_1(t) < A_2(t) \end{cases} \tag{12}$$

and $\alpha_2(t) = 1 - \alpha_1(t)$. In the case that $A_1(t) = A_2(t)$, the problem is said to be singular at time t. This means that (11) alone does not specify the optimal control. A singular interval is defined to be an interval $[t_1, t_2]$ such that the problem is singular for all t in this interval; this corresponds to $A_1(t) - A_2(t) = 0$ for all $t \in [t_1, t_2]$. The next lemma characterizes the optimal solution during any singular interval.

Lemma 2 *During any singular interval, the optimal control satisfies the condition given in Lemma 2.*

Proof: Notice that

$$\begin{aligned} \dot{A}_i(t) &= R_i\dot{U}(D_i(t))\dot{D}_i(t) + R_i\dot{p}_i(t) \\ &= R_i\dot{U}(D_i(t)(-\alpha_i(t)R_i + 1) + R_i(\alpha_i(t)R_i\dot{U}(D_i(t))) \\ &= R_i\dot{U}(D_i(t)), \end{aligned}$$

which does not depend on $\alpha_i(t)$.

Furthermore, for all $t \in [t_1, t_2]$, it must be that $\dot{A}_1(t) = \dot{A}_2(t)$. Therefore,

$$R_1\dot{U}(D_1(t)) = R_2\dot{U}(D_2(t)).$$

This corresponds to the choice of $\alpha_1(t)$ in Lemma 2. ■

This lemma implies that during any singular interval, the optimal scheduling policy behaves like the $\dot{U}R$ rule.

With the additional assumption $U(x) = -\beta x^2$, we have the following:

Proposition 3 *For $U(D) = -\beta D^2$ with $\beta > 0$, the $\dot{U}R$ rule is optimal.*

We omit the proof of this due to space considerations. The basic idea is to use the optimality conditions in (12) to show the system must behave as in Prop. 1. This implies that the $\dot{U}R$ rule is optimal.

6 Conclusions

We have presented an analysis of a simple utility-based scheduling rule, the $\dot{U}R$ rule. Although our analysis assumes only two classes of packets and a uniform initial delay distribution, the scheduling rule is easily generalized to other situations. For the draining problem considered, this scheduling rule is optimal. Current work involves understanding optimal scheduling rules under more general assumptions. In related work we have also investigated the delay perfomance of this type of scheduling policy in a system with time-varying channels [9].

Acknowledgment: The authors thank Dr. Rajeev Agrawal for helpful discussions concerning this work.

References

[1] P. Bender, *et al.* "CDMA HDR: a bandwidth-efficient high-speed wireless data service for nomadic users", in *IEEE Commun. Mag.*, Pages 70-77, July 2000.

[2] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar and P. Whiting, "CDMA data QoS scheduling on the forward link with variable channel conditions", Bell Laboratories Tech. Memo., 2000.

[3] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR", Bell Labs Tech. Memo., 2000.

[4] R. Agrawal, A. Bedekar, R. La, V. Subramanian, "A Class and Channel-Condition based Weighted Proportionally Fair Scheduler," *Proc. of ITC 2001*, Salvador, Brazil, Sept. 2001.

[5] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic Beam-forming using Dumb Antennas," *IEEE Trans. on Information Theory*, vol. 48, June 2002.

[6] R. Dudley, *Real Analysis and Probability,* Chapman and Hall, New York, 1989.

[7] F. Riesz and B. Nagy, *Functional Analysis,* Ungar, New York, 1955.

[8] M. Athans and P. Falb, *Optimal Control, An Introduction to the Theory and its Applications,* Mc Graw-Hill, New York, 1966.

[9] P. Liu, R. Berry, and M. Honig, "Utility-based Packet Scheduling in Wireless Networks," submitted to WCNC 2003.