# Energy-Efficient Scheduling with Individual Delay Constraints over a Fading Channel

Wanshi Chen
Qualcomm Inc.
San Diego, CA 92121
wanshic@qualcomm.com

Urbashi Mitra and Michael J. Neely
Dept. of Electrical Engineering
University of Southern California
Los Angeles, USA
{ubli, mjneely}@usc.edu

*Abstract*— **This paper focuses on energy-efficient packet transmission with individual packet delay constraints over a fading channel. The problem of optimal offline scheduling (vis-à-vis total transmission energy), assuming information of all packet arrivals and channel states before scheduling, is formulated as a convex optimization problem with linear constraints. The optimality conditions are analyzed. From the analysis, a recursive algorithm is developed to search for the optimal offline scheduling. The optimal offline scheduler tries to equalize the energy-rate derivative function as much as possible subject to the causality and delay constraints. The properties of the optimal transmission rates are analyzed, from which upper and lower bounds of the average packet delay are derived. In addition, a heuristic online scheduling algorithm, using causal traffic and channel information, is proposed and shown via simulations to achieve comparable energy and delay performance to the optimal offline scheduler in a wide range of scenarios.**

## I. Introduction

The fundamental tradeoff between packet delay and transmission power/energy has been extensively studied in the literature (for instance, see [1] and references therein). Two delay models are often considered, namely, the *single transmission deadline* model [19][17][8] where a set of packets are subject to a common deadline, and the *individual delay constraint* model [11][20][5] where each packet is subject to its own delay constraint.

For *dynamic packet arrivals under a time-invariant link*, the optimal energy-efficient schedule in minimizing the total transmission energy for the single transmission deadline model was developed in [19]. This optimal algorithm assumed knowledge of the total number of packets and the inter-arrival times of these packets before packet scheduling. As a result, it is an *offline* scheduling algorithm. An *online* algorithm, which assumed information of the current scheduling backlog and a maximum packet arrival rate, was also developed in [19]. Online scheduling for the single deadline model was also treated in [20] in which a stochastic optimal control algorithm was developed. Since all packets observe a single transmission deadline, the scheduling algorithm may result in large per packet delays, especially when the total number of packets to be transmitted is very large, as shown in [5]. The individual delay constraint model provides a flexible tradeoff between transmission energy and packet delay [11][20][6]. It was proven in [11] that all online scheduling can be expressed as a time-varying low-pass linear filter. Offline scheduling with general arrivals and QoS constraints was considered in [20], where an optimal scheduling procedure was developed. It was shown in [6] that the optimal offline scheduling yields a symmetry property in the optimal packet transmission durations from which a simple and exact solution of the average packet delay (including queuing and transmission delays) can be obtained.

For *fading channels without traffic variations*, a lot of work has been devoted to maximizing channel capacity under various power constraints. This includes the *delay-unconstrained* long-term capacity [9], and the *delay-limited* capacity [3][10]. For a finite coding delay with causal channel feedback, it was shown that in the limit of high signal-to-noise ratio (SNR), the optimal power control converges to a constant power transmission [12], while in the low SNR region, threshold-based power control appears to be optimal [14].

Power efficient scheduling utilizing *both queue and channel information* has also been studied, for instance, under average transmit power and average delay constraint [7][1][16][13] and under the single deadline model [17][8]. In [17], the optimal energy-efficient offline scheduling under a single deadline was extended to a fading channel with dynamic packet arrivals. A heuristic online scheduling algorithm named *look-ahead water-filling*, which exploits both backlog information and channel fading state, was studied and shown to be more energy efficient than the water-filling scheme purely based on channel states.

In this paper, we will focus on energy-efficient packet transmission with *individual packet delay constraints* over a *fading channel*. This can be viewed as an extension of the work in [20] and [5][6] for a continuous-time arrival model and time-invariant channels. The optimal offline scheduling (*vis-à-vis* total transmission energy) over a fading channel is analyzed. It is shown that the optimal scheduling tries to equalize the derivatives of the energy-rate function as much as possible, subject to the causality and delay constraints, in contrast to the equalization of transmission rates in static channels. From the analysis, a recursive algorithm called 'Constrained FlowRight', based on the 'FlowRight' algorithm for the single deadline model in [17], is developed to search for the optimal offline scheduling. The symmetry property of the optimal transmission rate vector still holds under the *i.i.d.* assumption of packet sizes and channel coefficients.

The average packet delay (including queuing and transmission delays) is characterized, in consideration of the symmetry property and potentially idling periods. Motivated by the properties of optimal offline scheduling, a heuristic online scheduling algorithm, which assumes both causal traffic and channel information, is developed. Both offline and online schedulers yield significant energy savings compared with simply clearing the buffer, regardless of traffic and channel states.

This paper is organized as follows. In Section II, the system model is described. The optimal offline scheduling over a fading channel and its properties are presented in Section III. Online schedulers are investigated in Section IV. Numerical results are given in Section V. Finally, some concluding remarks are drawn in Section VI.

## II. System Model

We consider a slow-varying flat fading channel. Such a channel is often modeled as a discrete-time block-fading additive white Gaussian noise (BF-AWGN) channel [15]. In the BF-AWGN model, the channel gain is fixed during a whole block and varies independently from block to block, where the block duration represents the channel's coherence time. This is illustrated in Figure 1, where the channel gains are represented by $g_i, i \in [1, \cdots, M+D-1]$, and $M+D-1$ is the total number of slots. Packets arrive only at slot boundaries, and can be served immediately ($i.e.$, the minimum possible queuing delay is 0). Each packet is subject to an individual delay constraint, which is assumed to be an integer $D > 1$ [1](in units of slots). The same individual delay constraint is assumed for all packets, although it can be extended to distinct individual delay constraints per packet. The total number of slots is fixed to be $M+D-1$, and the slot duration is denoted as $\tau_s$. Without loss of generality, a single packet is assumed to arrive at each slot with random packet sizes $B_i > 0, i \in [1, \cdots, M]$, and $B_i = 0$ for $i \in [M+1, M+D-1]$. This one-packet-per-slot assumption facilitates analysis, but is not restrictive, as the solution when some slots $i$, $i \in [1, \cdots, M]$, do not receive any packets can be obtained as a limiting case when the $B_i$ for those slots are made arbitrarily small. Likewise, the case when multiple packets arrive during the same slot can be treated by viewing them as a single bulk packet with a size given by the sum of the individual sizes. A fluid packet departure model is assumed. That is, a transmitted packet is not necessarily an integer number of arrived packets, but may be assembled using fragmented arrival packets up to an arbitrary precision.

Each packet transmission consumes some energy. The goal of the optimal offline schedule is to choose the number of transmitted bits $x_i$, or equivalently, the optimal transmission rate $r_i$, for each slot such that the total transmission energy of these $M$ packets is minimized while the underlying delay constraints are satisfied. The energy-rate function $f(r, g)$ is assumed to be strictly convex and monotonically increasing in transmission rate $r$ for each channel state $g$. For instance, from Shannon capacity, the energy-rate function is given by $f(r, g) = N_0(2^{2r} - 1)/g$, where $N_0$ denotes thermal noise.

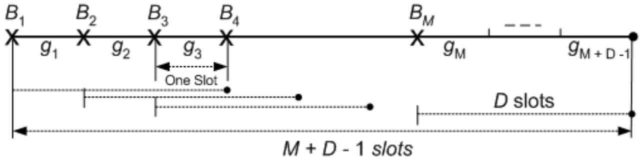

Fig. 1. The slotted individual delay constraint model.

The *offline* schedule assumes perfect knowledge of packet sizes and channel states for the entire duration $[0, \cdots, M+D-1]$ before scheduling. *Online* schedulers, on the other hand, only assume information of the current scheduling backlog. As in [19][6], all schedulers are assumed to follow the first-in-first-out (FIFO) service rule and the causality constraint. The FIFO rule implies that packets are transmitted in the same order in which they arrive. The causality constraint ensures that a packet cannot be transmitted before it arrives.

## III. Optimal Offline Scheduling over a Fading Channel

Here we formulate the optimal offline schedule over a fading channel as a convex optimization problem. We start by analyzing the optimality conditions, and deriving a recursive algorithm to search for the optimal schedule, before presenting the symmetry property of the optimal transmission rate vector and the resulting average packet delay performance.

The objective of the optimal offline scheduling is to solve the following problem

$$\min \sum_{m=1}^{M+D-1} f(r_m, g_m) \tag{1}$$

subject to

1) $\sum_{i=1}^{M+D-1} r_i \tau_s = \sum_{i=1}^{M+D-1} B_i$,
2) $\sum_{i=1}^{m} r_i \tau_s \leq \sum_{i=1}^{m} B_i, \qquad m = 1, \cdots, M$,
3) $\sum_{i=1}^{m} r_i \tau_s \geq \sum_{i=1}^{m-D+1} B_i, m = D, \cdots, M+D-2$,
4) $r_m \geq 0, \qquad m = 1, \cdots, M+D-1$,

where conditions 1 and 2 are due to the causality constraint, $i.e.$, the total number of delivered bits is no more than the total number of arrived bits so far, while condition 3 is due to the individual delay constraints, $i.e.$, the total number of delivered bits so far should be no less than the total number of arrived bits accumulated up to $D-1$ slots earlier. The non-negative transmission rate constraint by condition 4 is natural. Note that the causality condition 2) is also implicitly valid for $m = M+1, \cdots, M+D-1$. This can be easily checked by comparing condition 1) and condition 2) when $m = M$, and by realizing that $B_m = 0$, $m = M+1, \cdots, M+D-1$.

First, we have the uniqueness of optimal offline schedule:

*Theorem 1:* The optimal offline schedule for the individual delay constraint model over a fading channel is unique.

In addition, the properties of the per-bit transmission duration with respect to slot boundaries can be characterized by the

[1] The case when $D = 1$ is trivial and hence not considered.

following Lemma (A similar observation was also obtained in [18] for a continuous-time model):

*Lemma 1:* Under the optimal offline schedule, all bits transmitted over a slot have the same transmission duration. In other words, the transmission rate is constant during any slot.

The above two results can be proved via contradiction, using the strict convexity assumption of the energy-rate function, and the proofs are omitted here (see [4] for details).

### *A. The Optimality Conditions*

The problem in (1) is a convex problem with linear inequality constraints. Thus, the Karush-Kuhn-Tucker (KKT) conditions are sufficient for optimality [2]. The Lagrangian function is defined by

$$\begin{aligned} &L(\vec{r}, \lambda_r, \vec{\mu}, \vec{g}) = \textstyle\sum_{m=1}^{M+D-1} f(r_m, g_m) \\ &+\lambda_r \left(\textstyle\sum_{i=1}^{M+D-1} r_i \tau_s - \textstyle\sum_{i=1}^{M+D-1} B_i\right) + \textstyle\sum_{m=1}^{M} \mu_{1,m} h_{1,m}(\vec{r}) \\ &+ \textstyle\sum_{m=D}^{M+D-2} \mu_{2,m} h_{2,m}(\vec{r}) + \textstyle\sum_{m=1}^{M+D-1} \mu_{3,m} h_{3,m}(\vec{r}) \end{aligned} \tag{2}$$

where $\lambda_r$, $\mu_{1,m}$, $\mu_{2,m}$ and $\mu_{3,m}$ are Lagrange multipliers for conditions 1, 2, 3, and 4, respectively, and

$$\begin{cases} h_{1,m}(\vec{r}) \triangleq (\sum_{i=1}^{m} r_i \tau_s - \sum_{i=1}^{m} B_i) \leq 0 \text{ (causality)}, \\ h_{2,m}(\vec{r}) \triangleq (\sum_{i=1}^{m-D+1} B_i - \sum_{i=1}^{m} r_i \tau_s) \leq 0 \text{ (delay)}, \\ h_{3,m}(\vec{r}) \triangleq - r_m \leq 0 \text{ (non-negativity)}. \end{cases}$$

Note that when the above three constraints are tight, slot $m$ will be empty-ending, delay-critical, and idle, respectively. A slot is said to be *delay-critical* if, under a scheduling algorithm, it ends with an active delay constraint condition, $i.e.$, $h_{2,m}(\vec{r}) = 0$. If the slot is non-idle, the last transmitted packet in a delay-critical slot is called a *delay-critical packet*.

Denote

$$f^{'}(r_m, g_m)$$

as the derivative of $f(r_m, g_m)$ with respect to $r_m$, then there exist unique Lagrange multipliers $\lambda_r^*$, $\mu_{1,m}^* \geq 0$, $\mu_{2,m}^* \geq 0$ and $\mu_{3,m}^* \geq 0$, such that the following optimality conditions hold:

$$f^{'}(r_m^*, g_m) + \lambda_r^* \tau_s - \mu_{3,m}^* =$$

$$\begin{cases} \tau_s[-\sum_{i=m}^{M} \mu_{1,i}^* + \sum_{i=D}^{M+D-2} \mu_{2,i}^*], m = 1, \cdots, D \\ \tau_s[-\sum_{i=m}^{M} \mu_{1,i}^* + \sum_{i=m}^{M+D-2} \mu_{2,i}^*], \\ \qquad\qquad\qquad\qquad m = D+1, \cdots, M \\ \tau_s \sum_{i=m}^{M+D-2} \mu_{2,i}^*, \qquad m = M+1, \cdots, M+D-2 \\ 0, \qquad\qquad\qquad\qquad m = M+D-1 \end{cases} \tag{3}$$

Note that the complementary slackness condition [2] holds

$$\mu_{l,m}^* h_{l,m}(\vec{r}^*) = 0, 1 \leq l \leq 3, \text{ for all feasible } m \tag{4}$$

That is, for each $l$ and $m$, whenever the constraint $h_{l,m}(\vec{r}^*) \leq 0$ is slack, $i.e.$, $h_{l,m}(\vec{r}^*) < 0$, we must have $\mu_{l,m}^* = 0$. Similarly, when $\mu_{l,m}^* > 0$, we must have $h_{l,m}(\vec{r}^*) = 0$.

Under a *time-invariant* channel, the optimal offline scheduling observes no idling periods [6]. This comes from the assumption that the energy function is an increasing function of the transmission rate, and subsequently, the total transmission energy can always be reduced by increasing the transmission duration (hence reducing the transmission rate) for one or more packets. However, over a *time-varying* channel, even though there is data in the queue, the optimal offline schedule may choose to not transmit in a slot when future channel states are more energy-efficient. Hence, idle slots may become necessary. For instance, consider a simple example of scheduling two packets of finite sizes with $D = 2$, such that there are three slots. Assume the channel gains are given by $[0^+, \infty^-, \infty^-]$, $i.e.$, the channel gain is arbitrarily small for the first slot, and arbitrarily large for the second and the third slots. Also, assume the energy-rate derivative function $f^{'}(r, g)$ is positive at $r = 0$. Obviously, it is optimal not to transmit any data over the first slot, which results in an idle slot.

Besides the standard complementary slackness condition given by (4), we also notice three additional complementary slackness conditions for this particular problem:

$$\begin{cases} 1): \mu_{1,m}^* \mu_{3,m}^* = 0, \qquad m = 1, \cdots, M \\ 2): \mu_{1,m}^* \mu_{2,m}^* = 0, \qquad m = D, \cdots, M \\ 3): \mu_{2,m}^* \mu_{3,m+1}^* = 0, \qquad m = D, \cdots, M+D-2 \end{cases} \tag{5}$$

That is: 1) *any slot* $m \in [1, \cdots, M]$ *can not be idle and end with an empty buffer at the same time* (as the slot will have at least one packet, the one that arrived in that slot); 2) *any slot* $m \in [D, \cdots, M]$ *can not be a delay-critical slot and end with an empty buffer at the same time* (due to the FIFO constraint and the fact $D \geq 2$); 3) and *if slot* $m+1$ *is idle, slot* $m$ *can not be delay-critical* (otherwise, a delay constraint violation will occur). The detailed proof of (5) can be found in Appendix A. It is worth emphasizing that since it is assumed that data arrives at the beginning of each slot $m \in [1, \cdots, M]$, an empty-ending slot means that all data is cleared at the end of that slot (while new data will arrive immediately after, $i.e.$, at the beginning of the next slot, except for slot $M$.).

We will now characterize the properties of these optimal Lagrangian multipliers, which are crucial to developing a recursive algorithm to obtain the optimal offline scheduling. This is done via comparison of the derivatives of the energy-rate function over adjacent slots. As we will see, the properties of the optimal derivatives herein are similar to the properties of the optimal transmission durations or optimal transmission rates for the static channel case [20][5]. First, consider the difference between the $m$-th equation and the $(m+1)$-th equation in (3), $i.e.$, $\Delta f_{m,1}^{'*} \triangleq f^{'}(r_m^*, g_m) - f^{'}(r_{m+1}^*, g_{m+1})$. We have

$$\Delta f_{m,1}^{'*} = \mu_{3,m}^* - \mu_{3,m+1}^* +$$

$$\begin{cases} -\tau_s \mu_{1,m}^*, \qquad m = 1, \cdots, D-1 \\ -\tau_s \mu_{1,m}^* + \tau_s \mu_{2,m}^*, \qquad m = D, \cdots, M \\ \tau_s \mu_{2,m}^*, \qquad m = M+1, \cdots, M+D-2 \end{cases} \tag{6}$$

Combining (4), (5), and (6), the following Lemma is straightforward:

*Lemma 2:* When the optimal transmission rates during two adjacent slots $m$ and $m+1$ are strictly positive, $i.e.$, both slots are non-idle, we have:

1) $\Delta f^{'*}_{m,1} = -\tau_s \mu^*_{1,m} \le 0$, if $m \in [1, \cdots, M-1]$ and slot $m$ ends with an empty buffer[2];
2) $\Delta f^{'*}_{m,1} = \tau_s \mu^*_{2,m} \ge 0$, if $m \in [D, \cdots, M+D-2]$ and slot $m$ ends with a delay critical packet;
3) $\Delta f^{'*}_{m,1} = 0$, if $m = 1, \cdots, M+D-2$, and slot $m$ ends neither with an empty buffer nor with a delay-critical packet.

*Proof:* See Appendix B. ■

Thus, this Lemma completely characterizes the Lagrangian multipliers $\mu^*_{1,m}$ and $\mu^*_{2,m}$ via $\Delta f^{'*}_{m,1}$ when both slots $m$ and $m+1$ are not idle.

*Remarks*: These properties in optimal transmission rates can be interpreted as follows. While an active causality constraint may result in a lower transmission rate than an otherwise more energy-efficient rate (non-increasing derivatives between two adjacent non-zero rate slots), an active individual delay constraint may require a higher transmission rate than an otherwise more energy-efficient one (non-decreasing derivatives between two adjacent non-zero rate slots). When both constraints are inactive, the optimal transmission rate is chosen to achieve the best possible energy efficiency purely depending on the channel states (zero derivative difference, and hence a constant transmission rate in case of a time-invariant channel).

Now let us consider the case when only one slot in the pair $\{m, m+1\}$ is idle. The following Lemma characterizes the properties of $\Delta f^{'*}_{m,1}$ in such scenarios:

*Lemma 3:* When one and only one of the two slots in the pair $\{m, m+1\}$ is idle (*i.e.*, zero transmission rate), we have:

1) $\Delta f^{'*}_{m,1} = \mu^*_{3,m} + \tau_s \mu^*_{2,m} \mathbf{1}_{m-D} \ge 0$, if $m \in [1, M+D-2]$ and slot $m$ is idle but slot $m+1$ is not;
2) $\Delta f^{'*}_{m,1} = -\mu^*_{3,m+1} - \tau_s \mu^*_{1,m} \mathbf{1}_{M-m} \le 0$, if $m \in [1, M+D-2]$ and slot $m+1$ is idle, but slot $m$ is not.

*Proof:* See Appendix B. ■

where the indicator function $\mathbf{1}_n$ is 1 when $n \ge 0$ and 0 otherwise. In other words, even with a zero transmission rate, idle slots still have derivatives no less than those of the non-idling neighboring slots.

In case of two or more consecutive idle slots, the comparison of the derivatives between two idle slots would no longer provide much information, as $\Delta f^{'*}_{m,1}$ can be either non-negative or negative. However, we can still compare any idle slot with its two closest non-zero rate slots. Note that the number of consecutive idle slots has to be no more than $D-1$ to avoid delay violations. Define $\Delta f^{'*}_{m,l} \overset{\Delta}{=} f^{'}(r^*_m, g_m) - f^{'}(r^*_{m+l}, g_{m+l})$, $D > l \ge 1$ and $m + l < M + D$. Note that

$$\Delta f^{'*}_{m,l} = \sum_{j=m}^{m+l-1} \Delta f^{'*}_{j,1}.$$

Using this and (6), we obtain

$$\begin{aligned} \Delta f^{'*}_{m,l} &= \mu^*_{3,m} - \mu^*_{3,m+l} \\ &- \tau_s \textstyle\sum_{i=0}^{l-1} \mu^*_{1,m+i} \mathbf{1}_{M-m-i} \\ &+ \tau_s \textstyle\sum_{i=0}^{l-1} \mu^*_{2,m+i} \mathbf{1}_{m+i-D}, \end{aligned} \tag{7}$$

for $m = 1, \cdots, M+D-2$, where we have assumed slots $m+1$ to $m+l-1$ are idle. Now we can have results similar to Lemma 3:

*Lemma 4:* When there are two or more consecutive idle slots (*i.e.*, $l \ge 2$), we have:

1) $\Delta f^{'*}_{m,l} = \mu^*_{3,m} + \tau_s \sum_{i=0}^{l-1} \mu^*_{2,m+i} \mathbf{1}_{m+i-D} \ge 0$, if $m \in [1, \cdots, M+D-3]$ and slots $m, m+1, \cdots, m+l-1$ are idle but slot $m+l$ is not;
2) $\Delta f^{'*}_{m,l} = -\mu^*_{3,m+l} - \tau_s \sum_{i=0}^{l-1} \mu^*_{1,m+i} \mathbf{1}_{M-m-i} \le 0$, if $m \in [1, \cdots, M+D-3]$ and slots $m+1, m+2, \cdots, m+l$ are idle, but slot $m$ is not.

*Proof:* See Appendix B. ■

That is, the comparison of an idle slot to its closest non-zero rate slots provides the same information as in the case of $\{m, m+1\}$ when one of them is idle.

We can now generalize Lemma 2 to any pair of slots $\{m, m+l\}$, $l \ge 2$, which are separated by one or more idle slots:

*Lemma 5:* When the optimal transmission rates during two slots $m$ and $m+l$, $l \ge 2$, $m+l < M+D$, are strictly positive (*i.e.*, $r^*_m > 0$ and $r^*_{m+l} > 0$), but all slots in between are idle (*i.e.*, $r^*_{m+i} = 0$ for $i = 1, \cdots, l-1$), we have:

1) $\Delta f^{'*}_{m,l} = -\tau_s \sum_{i=0}^{l-1} \mu^*_{1,m+i} \mathbf{1}_{M-m-i} \le 0$, if $m \in [1, \cdots, M-1]$ and slot $m+l-1$ is not a delay-critical slot;
2) $\Delta f^{'*}_{m,l} = \tau_s \sum_{i=0}^{l-1} \mu^*_{2,m+i} \mathbf{1}_{m+i-D} \ge 0$, if $m \in [D, \cdots, M+D-3]$ and slot $m$ does not end with an empty buffer;
3) $\Delta f^{'*}_{m,l} = 0$, if $m = 1, \cdots, M+D-3$, and slot $m$ does not end with an empty buffer and slot $m+l-1$ is not a delay-critical slot.

*Proof:* See Appendix B. ■

That is, it is as if all the idle slots in between can be ignored when determining the difference of the derivatives of two non-idle slots.

*Remarks*: Lemma 5 has a constrained *'water-filling'* interpretation. The optimal offline scheduling tries to *equalize the derivative $f^{'}(r_m, g_m)$ as much as possible* (*i.e.*, $\Delta f^{'*}_{m,l} = 0$), subject to the causality and individual delay constraints. In slots where equalization is not justified from an energy-efficiency perspective, these slots should be idle instead. Thus, the derivative $f^{'}(r_m, g_m)$ can be viewed as a measure of *relative cost* of choosing a transmission rate $r_m$ given a channel state $g_m$. In static channels, the energy-rate derivative function degenerates to $f^{'}(r_m)$, in which case a constant relative cost translates into a constant transmission rate. This is consistent with the efforts of equalizing transmission rates or transmission durations by the optimal offline scheduling over time-invariant channels in [5][6][20]. This important observation, indeed, motivates a simple online scheduler design, as will be discussed in Section IV.

Lemmas 2, 3, 4, and 5 provide us with a complete characterization of the optimal Lagrangian multipliers $\lambda^*_r$, $\vec{\mu}^*_1$, $\vec{\mu}^*_2$, and $\vec{\mu}^*_3$ in all possible scenarios. Next, we will exploit these optimality conditions to develop a recursive optimal offline scheduling search algorithm.

[2] Note that if slot $m \ge M$ ends with an empty buffer, all the subsequent slots will be idle since no arrivals are assumed after slot $M$.

### B. A Recursive Search Algorithm

We will first briefly review the *FlowRight* algorithm developed in [17] for the optimal offline scheduling algorithm under a *single transmission deadline model*, before presenting the recursive algorithm for the individual delay constraint model. The FlowRight algorithm in [17] is based on optimizing rates over two adjacent slots $\{m, m+1\}$. The per-pair optimization is done for all pairs $m = 1, \cdots, M+D-2$, from left to right, and a *pass* is completed when $m = M+D-2$. The recursive algorithm is performed with as many passes as needed until a desirable accuracy has been reached. It was proven that such an algorithm converges, and it converges to the optimal offline scheduler [17]. The name of 'FlowRight' comes from the fact that the derivatives of two neighboring non-zero rate slots are monotonically non-decreasing, as can be seen from Lemma 2 and Lemma 5 when the impact of individual delay constraints is eliminated ($i.e.$, no delay critical packets in any slot $m \in [1, \cdots, M+D-2]$). This means that in the per-pair optimization, the bits are always moved from left to right such that the transmission of earlier arrived bits can be postponed for better channel states. This information flow is limited by channel states and the causality constraint, but never from right to left.

Here we develop a recursive algorithm for the optimal offline schedule for the individual delay constraint model. The algorithm is referred as a *Constrained FlowRight* algorithm and it differs from the original FlowRight algorithm by incorporating:

- Idling slots (which was not explicitly addressed in [17]), and
- Individual delay constraints.

To achieve this, we can re-write the delay constraint condition in (1) (Condition 3) as: (i) $r_m \geq 0$ for $m = 1, \cdots, D-1$, and (ii)

$$r_m \geq \Big(\sum_{i=1}^{m-D+1} B_i/\tau_s - \sum_{i=1}^{m-1} r_i\Big)^+, \text{ for } m = D, \cdots, M+D-1. \tag{8}$$

where $(x)^+ \triangleq \max\{0, x\}$. Now the Constrained FlowRight algorithm can be developed as follows:

1) Initialize the rates for each slot as $\vec{r}^0 = \vec{B}/\tau_s$. Set $k = 0$.
2) In increasing order, for each $m \in [1, \cdots, M+D-2]$, do the following:
   i) Minimize $f(r) + f(r^{total}_{m,m+1} - r)$ subject to the following constraints: a) $r^{total}_{m,m+1} = r^k_m + r^k_{m+1}$; b) $r \geq 0$ (non-negativity); c) $r$ is upper-bounded by some value that can be computed by the causality constraint (see Condition 2 in (1)) using the rate vector $[r^k_1, r^k_2, \cdots, r^k_{m-1}]$[3]; and d) $r$ is lower-bounded by some value that can be computed by the delay constraint (8) using the rate vector $[r^k_1, r^k_2, \cdots, r^k_{m-1}]$. Denote the above optimal solution as $r^*_m$ and let $r^*_{m+1} \triangleq r^{total}_{m,m+1} - r^*_m$. Update $\{r^k_m, r^k_{m+1}\}$ in $\vec{r}^k$ to $\{r^*_m, r^*_{m+1}\}$.
   ii) If $r^*_{m+1} = 0$, repeat the same per-pair rate optimization for slots $\{m, m+l\}$[4], $l \geq 2$, which yields $\{r^*_m, r^*_{m+l}\}$, where $r^*_{m+l} \triangleq r^k_m + r^k_{m+l} - r^*_m$, until $r^*_{m+l} > 0$, or until $l = D-1$ (no more than $D-1$ consecutive idle slots), or until $m+l = M+D-1$ is reached. Update $\{r^k_m, r^k_{m+l}\}$ in $\vec{r}^k$ to $\{r^*_m, r^*_{m+l}\}$.
3) After one pass ($i.e.$, when $m = M+D-2$ in the above step), set $\vec{r}^{k+1} = \vec{r}^k$ and $k = k+1$. Repeat the same procedure until $k = K$, such that $\max|\vec{r}^{K-1} - \vec{r}^K| < \epsilon$, where $\epsilon \ll 1$ is arbitrarily small.

*Theorem 2:* The following statements hold:

1) For each step in the above Constrained FlowRight algorithm, information always flows right ($i.e.$, $\sum_{m=1}^{j} r^k_m \leq \sum_{m=1}^{j} r^{k-1}_m, \forall j \geq 1$) without violating the individual delay constraints;
2) The Constrained FlowRight algorithm converges to $\vec{r}^*$.

*Proof:* To prove the first statement, we need to look at the per-pair rate optimization in the above recursive algorithm. Following a procedure similar to that in [17] for the *single deadline model*, we can prove that the information always flows right in the per-pair rate optimization. The satisfaction of the delay constraints comes from the explicit rate constraint condition of (8).

Now for the second statement, the convergence part is due to the always right-flow of information, as also shown in [17] for the *single deadline model*. The optimality part can be easily verified by checking the properties of the optimal Lagrangian multipliers for various scenarios as characterized by Lemmas 2, 3, 4, and 5. ■

Note that the convergence speed of this recursive Constrained FlowRight algorithm depends on the energy-rate function $f(r, g)$, the packet sizes $B_i$, the channel gains $g_i$, $i \in [1, \cdots, M+D-1]$, and the delay constraint $D$. In Section V, this algorithm will be used to search for the optimal offline scheduling to investigate its performance and properties.

### C. Symmetry Property and Packet Delay Performance

Similar to the static channel case, we also have the symmetry property of the optimal transmission rates $\vec{r}$. This important property not only provides an insight to the optimal scheduling algorithm, but also makes it possible to analyze the average packet delay performance into a very compact closed form for continuous-time static channels [6]. For time-slotted fading channels, due to the potential existence of idling slots, we can no longer obtain a closed form solution of the average packet delay performance. However, as we will show, the symmetry property leads to a lower bound and an upper bound of the average packet delay performance.

*Theorem 3:* For any $M \geq 1$, when the joint probability distribution for the random packet vector $[B_1, \cdots, B_M]$ is identically distributed to the reversed packet vector $[B_M, \cdots, B_1]$

[3] Effectively, $r \leq r^k_m$, since the information always flows right as shown in [17].

[4] Note that the delay constraint has to updated by replacing $m$ by $m+l-1$ in (8). In addition, $r^*_{m+1} = \cdots = r^*_{m+l-1} = 0$ (idle slots).

(such a property clearly holds, *e.g.*, when $B_i$ are *i.i.d.*), independent of the *i.i.d* channel gains $g_m, 1 \leq m \leq M-1$ [5], then under the optimal offline scheduling, the optimal transmission rates $r_m$ and $r_{M-m+1}$ are identically distributed. In particular, $E\{r_m\} = E\{r_{M-m+1}\}$.

*Proof:* Here we provide a sketch of the proof. The detailed proof follows the same procedure as in [6] for the continuous-time arrival model over a static channel, compared with a time-slotted fading channel discussed herein. The proof essentially relies on a *time reversal* argument, where a sample path trajectory of the forward running system, characterized by

$$\vec{B}^{(f)} = [B_1, \cdots, B_M],$$

and,

$$\vec{g}^{(f)} = [g_1, \cdots, g_{M+D-1}],$$

is compared to a corresponding time reversed system, characterized by

$$\vec{B}^{(r)} = [B_M, \cdots, B_1],$$

and,

$$\vec{g}^{(r)} = [g_{M+D-1}, \cdots, g_1].$$

The unique optimal transmission rate vector for the forward running system can be shown also feasible and uniquely optimal for the corresponding time reversed system. ■

We define the average packet delay as:

$$\bar{q}(M) \triangleq E\{\frac{1}{M}\sum_{m=1}^{M} q_m\}$$

where $q_m$ is the delay (including queuing delay and transmission delay[6]) experienced by packet $m$ under the optimal offline schedule with a particular realization of the channel gain vector and the packet size vector, and the expectation is taken over all channel and packet size realizations. By exploiting the symmetry property, the following bounds can be obtained:

*Theorem 4:* For any $M \geq 1$, when the joint probability distribution for the random packet vector $[B_1, \cdots, B_M]$ is identically distributed to the reversed packet vector $[B_M, \cdots, B_1]$, independent of the *i.i.d* channel gains $g_m, 1 \leq m \leq M+D-1$, then under the optimal offline scheduling, the average packet delay is bounded by

$$\tau_s \left[\frac{D}{2} + \frac{1}{2D}\right] \leq \bar{q}(M) \leq \tau_s \left[1 + \frac{M+1}{2M}(D-1)\right] \quad (9)$$

When $M \to \infty$, $\tau_s(D + 1/D)/2 \leq \bar{q}(\infty) \leq \tau_s(D+1)/2$. In the case of static channels where the channel gains are fixed $g_m = c$, $\forall m$, where $c$ is a constant, the average packet delay is given by

$$\bar{q}(M)^{static} = \tau_s \left[1 + \frac{M+1}{2M}(D-1),\right] \quad (10)$$

and converges to $\bar{q}(\infty)^{static} = \tau_s(D+1)/2$.

*Proof:* See Appendix C. ■

Note that the upper and the lower bounds differ only by the value $\tau_s[(1-1/D)+(D-1)/M]/2$, which converges to $\tau_s(1-1/D)/2$ as $M$ approaches infinity. In addition, the average packet delay performance bounds are *not* a function of the packet size vector. Indeed, as shown in Appendix C, due to the symmetry property, the delay performance bounds only depend on the total transmission duration of these $M$ packets (*i.e.*, $M + D - 1$ slots), and the total number of potentially idling slots.

## IV. Online Scheduling over a Fading Channel

Here we describe a heuristic online scheduler motivated by the optimality conditions for the optimal offline scheduling discussed in Section III. This online scheduler, referred as the *Derivative Directed* or *DD* online scheduler, tries to keep a constant derivative value as much as possible. The derivative variations come from traffic and channel variations.

Traffic variations can be smoothed out by lower-bounding the transmission rate by

$$r^{DD}_{m,min} = \max_{i \in [1,\cdots,D]} \left(\frac{\sum_{i=1}^{D} U_{m,D-i}}{i\tau_s}\right)^{\alpha^{i-1}}, \quad (11)$$

an extension of the optimal static buffer flushing algorithm discussed in [20] and [6], where $\alpha \geq 0$, $U'_m = U_{m,0} + U_{m,1} + U_{m,2} + \cdots + U_{m,D-1}$ is the total buffered bits, and $U_{m,i}$ is the buffered bits with a delay constraint of $D - i$.

Channel variations can be accommodated by choosing a transmission rate as a function of the latest derivative value $f'_{m-1}$ and current channel state $g_m$, *i.e.*,

$$r^{DD}_{m,channel} = v(f'_{m-1}, g_m).$$

where $r = v(f', g)$ as the inverse of $f'(r, g)$. Finally, the transmission rate at slot $m$ given by:

$$r^{DD}_m = \max\{\min\{r^{DD}_{m,channel}, U'_m/\tau_s\}, r^{DD}_{m,min}\}, \quad (12)$$

where $U'_m/\tau_s$ is due to the causality constraint. The derivative can be updated by, *e.g.*,

$$f'_m = \quad \beta f'_{m-1} + (1-\beta) f'(r^{DD}_m, g_m), \qquad \text{if } r^{DD}_m > 0, \quad (13)$$

where $0 < \beta \leq 1$ is the forgetting factor, and remains unchanged in case of idling slots. This is because an idling slot is due to a bad channel state and its energy-rate derivative value may be too large.

Without loss of generality, assume the average channel gain is 1. The following summarizes the online DD scheduling algorithm:

[5] The result also holds for any channel gains when the joint probability distribution for the random packet vector $[g_1, \cdots, g_{M+D-1}]$ is identically distributed to the reversed packet vector $[g_{M+D-1}, \cdots, g_1]$. However, for practical interest, we will only focus on the *i.i.d.* channel gains in this paper.

[6] That is, the time interval from when packet $B_m$ arrives till when its last bit's transmission is completed, which is not necessarily aligned with slot boundaries.

1) Initialize $\hat{f}_0^{'} = f^{'}(r_{avg}, 1)$, where $r_{avg}$ is the known or estimated average arrival rate.
2) For each slot $m = 1, \cdots, M + D - 1$, do the following:
   i) Determine $r_m^{DD}$ by (12).
   ii) Update $\hat{f}_m^{'}$ by (13).

We have the following Lemma:

*Lemma 6:* The above DD online scheduler guarantees the satisfaction of the causality and the individual delay constraints.

*Proof:* The causality constraint is explicitly guaranteed by $r_{m,max}^{DD}$. The satisfaction of the individual delay constraints is guaranteed by the FIFO assumption and the fact that $r_{m,min}^{DD} \geq U_{m,D-1}$ in (11) for $\alpha \geq 0$, where $U_{m,D-1}$ denotes the buffered packets that must be delivered by the end of slot $m$. ■

Note that zero-rate transmissions occur when $v(\hat{f}_{m-1}^{'}, g_m) \leq 0$ and $r_{m,min}^{DD} = 0$. For instance, if $f(r,g) = N_0(2^{2r} - 1)\ln 2/g$, the channel threshold under or at which an idle slot occurs is proportional to $1/f^{'}$. Generally, we expect $f^{'} \propto 2^{2r_{avg}\tau_s}$. Therefore, when the average transmission rate is small, $f^{'}$ is close to 1. Thus the channel threshold is relatively high and we may choose to transmit only in good channel conditions. On the other hand, when high transmission rates are necessary, $f^{'}$ becomes very large, and the channel threshold is close to 0. In this case, the DD online scheduler tends to transmit over all channel states. Such a phenomenon is similar to that of the optimal offline scheduler, as will be shown via simulations.

## V. Numerical Results

The energy-rate function is assumed to be $f(r) = (2^{2r} - 1)/g$ [17], resulting from Shannon capacity. The packet sizes are normalized based on frequency bandwidth and slot duration and can be interpreted as the number of bits per channel use. The channel coefficients are assumed to be Rayleigh distributed.

Figure 2 illustrates one example run of the optimal offline schedule for the individual delay constraint model. The normalized packet size is very small and fixed at $B = 0.1$ such that the energy-rate function approaches a linear relationship. It can be observed that the optimal schedule chooses to transmit only in good channel conditions, and stays idle in bad channels. Indeed, it is not difficult to see that, under a strict linear energy-rate function, the optimal schedule would choose to transmit packet $m$ only in the best slot $m_{opt} \in [m, \cdots, m+D-1]$ which has the largest channel gain. Such a threshold-like scheduling was also observed in other different delay/power tradeoff settings, *e.g.*, [7],[8]. The derivatives of the energy-rate function for the non-idling slots are also shown. The derivatives under the optimal offline scheduling exhibit a stair-case property, which tends to remain unchanged while in response to active causality and delay constraints. The idling slots have larger derivatives (not shown), such that the 'water-filling' rule prohibits transmissions in these slots.

In contrast, when the normalized packet size is very large (fixed at $B = 5$), the optimal offline scheduling tends to transmit over all slots (with rates proportional to channel gains), and approaches a constant transmission rate in the limit. This is illustrated in Figure 3. This is not surprising as when $B$ increases, the energy-rate function $f(r) = (2^{2r} - 1)/g$ is increasingly dominant by $r$, while the channel gain has an decreasing impact. Thus, the idling constraint becomes more slack, and it is more justified to transmit packets over all slots, with transmission rates proportional to the channel gains. In this particular example, the optimal derivatives of the energy-rate function are the same for all slots, except the last one, which has a smaller value due to an active delay constraint at slot $M + D - 2$ (the second slot to the last).

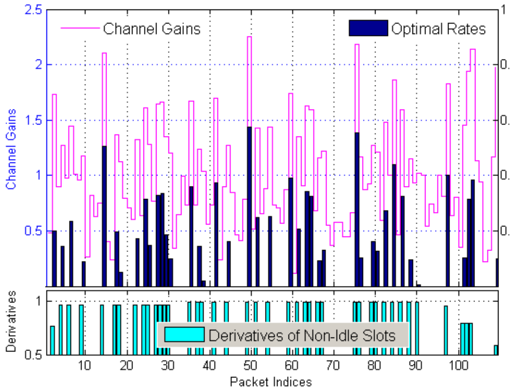

Fig. 2. An example run of the optimal offline scheduling for the individual delay constraint model, $M = 100$, $D = 10$, and $B = 0.1$.

Fig. 4 demonstrates the symmetry property of the average (over independent runs) optimal transmission rates. The single transmission deadline model, which assumes the same transmission duration, packet sizes and channel gains as the individual delay constraint model, is shown for comparison. For the single deadline model, the optimal transmission rate increases, on average, with packet indices, as earlier packets can exploit more future arrivals and potentially postpone some transmissions for better channel states. However, this may lead to significant *individual packet delays*, as analytically shown in [5] for static channels. In contrast, the individual delay model always yields finite individual packet delays, and its average delay performance is governed by Theorem 4 in Section III.

Fig. 5 and Fig. 6 show the average transmission energy and packet delay performance, respectively, for the offline and online schedulers when $M = 100$ and $\bar{B} = 1$. It can be seen that both the offline and the online schedulers require significantly less energy than the greedy buffer clearing scheduler. As the individual delay constraint $D$ increases, the energy required by the schedulers for the individual delay constraint model approaches that of the single deadline model[7]. The DD online scheduler achieves comparable energy performance to the optimal offline scheduler, which demonstrates an effective exploitation of the properties of the optimal offline scheduling for online scheduling design. The average packet delay of the

[7] Note that for fair comparison, the total transmission duration for the single deadline model is equal to $(M + D - 1)$ slots as well.

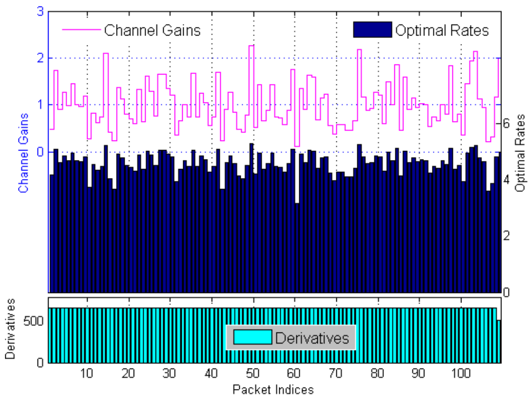

Fig. 3. An example run of the optimal offline scheduling for the individual delay constraint model, $M = 100$, $D = 10$, and $B = 5$.

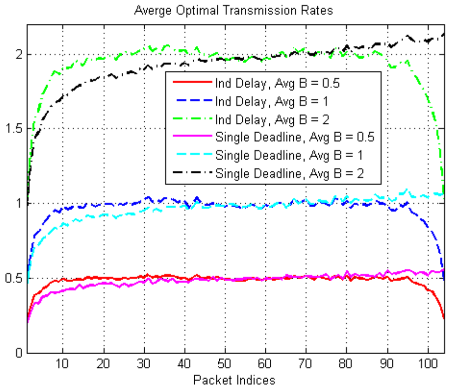

Fig. 4. Average transmission rates under optimal offline scheduling, $M = 100$, $D = 5$.

DD online scheduler is slightly less than that of the optimal offline scheduler for the individual delay constraint model [8], but significantly less than that of the single deadline model.

## VI. Conclusions

This paper focuses on energy-efficient packet transmission with individual packet delay constraints over a fading channel. This can be viewed as an extension of the work by in [20] and [5][6] for a continuous-time arrival model and static channels. The problem of optimal offline scheduling (vis-à-vis total transmission energy), assuming information of all packet arrivals and channel states before scheduling, is formulated as a convex optimization problem with linear constraints. The optimality conditions are analyzed, from which a recursive search algorithm is developed. The symmetry property of the optimal transmission rate vector (or, equivalently, the transmission duration vector) still holds under the *i.i.d.* assumption of packet sizes and channel coefficients. Combining the symmetry property with the potential idling periods, upper and lower bounds of the average packet delay (including queuing and transmission delays) are derived. The properties of the optimal offline scheduling and the impact of packet sizes, individual delay constraints, and channel variations are demonstrated via simulations. A heuristic online scheduling algorithm, which assumes both causal traffic and channel information, is also proposed and compared with the optimal offline schedule via simulations.

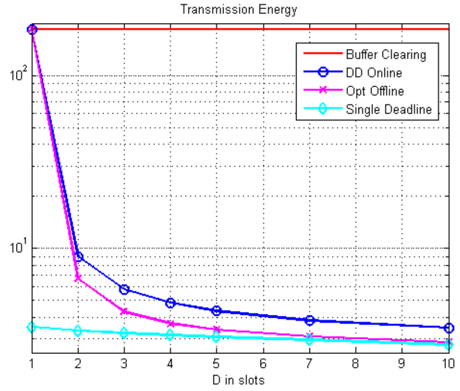

Fig. 5. Average transmission energy for the offline and online schedulers vs. $D$, $M = 100$, $\bar{B} = 1$.

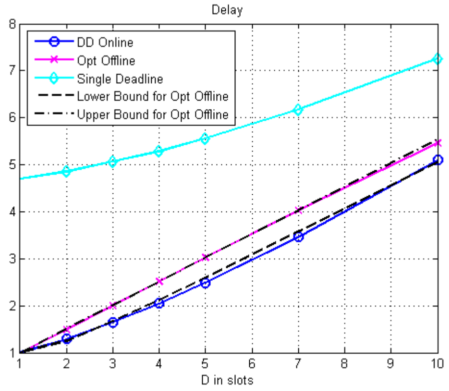

Fig. 6. Average packet delay (right) for the offline and online schedulers vs. $D$, $M = 100$, $\bar{B} = 1$.

## Acknowledgments

This research has been funded in part by one or more of the following grants or organizations: NSF ANI-0087761, NSF Special Projects ANI-0137091, NSF ITR CCF-0313392, NSF Grant OCE-0520324, and the DARPA IT-MANET program, Grant W911NF-07-1-0028.

## Appendix A
## Additional Complementary Slackness Conditions

This is to prove the additional complementary slackness conditions in (5) for the optimal offline schedule. Recall that:

1) $h_{l,m}(\vec{r}^*) < 0$ implies that $\mu^*_{l,m} = 0$;
2) $\mu^*_{l,m} > 0$ implies that $h_{l,m}(\vec{r}^*) = 0$,

for any $l$ and $m$, due to the conventional complementary slack condition by (4).

[8] It is worth noting that our goal is to optimize transmission energy instead of delay. Thus, it is not surprising that the DD online scheduler consumes more energy than the optimal offline scheduler, but yields less packet delay.

The first condition, $\mu_{1,m}^* \mu_{3,m}^* = 0, m = 1, \cdots, M$, is due to the fact that given $B_m > 0, m = 1, \cdots, M$, if $\mu_{3,m}^* > 0$ (hence $h_{3,m}(\vec{r}^*) = -r_m^* = 0$ and so slot $m$ is an idle slot), we have

$$\begin{aligned} h_{1,m}(\vec{r}^*) &= \sum_{i=1}^{m} r_i^* \tau_s - \sum_{i=1}^{m} B_i \\ &= \sum_{i=1}^{m-1} r_i^* \tau_s - \sum_{i=1}^{m} B_i \\ &< \sum_{i=1}^{m-1} r_i^* \tau_s - \sum_{i=1}^{m-1} B_i \\ &\leq 0 \end{aligned}$$

Thus, $\mu_{1,m}^* = 0$ or slot $m$ does not end with an empty buffer. Similarly, the reverse case, *i.e.*, $\mu_{1,m}^* > 0$ implying $\mu_{3,m}^* = 0$, can also be derived. In other words, *any slot* $m \in [1, \cdots, M]$ *can not be idle and end with an empty buffer at the same time*. The second condition, $\mu_{1,m}^* \mu_{2,m}^* = 0, m = D, \cdots, M$, is due to the fact that if $\mu_{1,m}^* > 0$, slot $m$ must satisfy $h_{1,m}(\vec{r}^*) = 0$, *i.e.*, it must end with an empty buffer,

$$\begin{aligned} h_{2,m}(\vec{r}^*) &= \sum_{i=1}^{m-D+1} B_i - \sum_{i=1}^{m} r_i^* \tau_s \\ &= -\sum_{i=m-D+2}^{m} B_i - h_{1,m}(\vec{r}^*) \\ &= -\sum_{i=m-D+2}^{m} B_i \\ &< 0 \end{aligned}$$

Hence, $\mu_{2,m}^* = 0$. Similarly, if $\mu_{2,m}^* > 0$, we have $\mu_{1,m}^* = 0$. In other words, *any slot* $m \in [D, \cdots, M]$ *can not be a delay-critical slot and ends with an empty buffer at the same time.*

For the third condition, first note that an idle slot may possibly be a delay-critical slot, *i.e.*, $h_{2,m}(\vec{r}^*) = \sum_{i=1}^{m-D+1} B_i - \sum_{i=1}^{m} r_i^* \tau_s$ and $r_m^* = 0$ may be satisfied simultaneously, such that

$$\mu_{2,m}^* \mu_{3,m}^* > 0.$$

However, if slot $m+1$ is idle, slot $m$ can not be delay-critical, and vice versa, as indicated by condition 3 ($\mu_{2,m}^* \mu_{3,m+1}^* = 0, m = D, \cdots, M+D-2$). Consider if $\mu_{2,m}^* > 0$, we have $h_{2,m}^* = 0$ and hence slot $m$ is delay-critical. That is, slot $m$ ends with completing transmission of packets arrived at slot $m-D+1$. Since $B_{m-D+1} > 0$, for $m = D, \cdots, M+D-1$, slot $m+1$ has to at least serve packets arrived at slot $m-D+2$ and thus can not be idle, or $\mu_{3,m+1}^* = 0$. Similarly, one can also show that if $\mu_{3,m+1}^* > 0$, we must have $\mu_{2,m}^* = 0$.

## Appendix B
## Proof of Lemmas 2, 3, 4, and 5

For Lemma 2, since $r_m^* > 0$ and $r_{m+1}^* > 0$, we have $\mu_{3,m}^* = \mu_{3,m+1}^* = 0$. Case 1 is further due to an empty buffer at the end of slot $m$, and hence a non delay-critical slot such that $\mu_{2,m}^* = 0$ in (6). Case 2 is further due to $\mu_{1,m}^* = 0$ as slot $m$ is delay-critical and hence non-empty ending. Case 3 is further due to $\mu_{1,m}^* = 0$ and $\mu_{2,m}^* = 0$ as slot $m$ is neither empty-ending nor delay-critical.

For Lemma 3, Case 1 is due to $\mu_{3,m+1}^* = 0$ (non-idling slot) and $\mu_{1,m}^* = 0$ (non-empty ending slot) in (6). Case 2 is due to $\mu_{3,m}^* = 0$ (non-idling slot) and the last condition in (5), *i.e.*, slot $m$ can not be delay-critical if slot $m+1$ is idle.

For Lemma 4, Case 1 is due to $\mu_{3,m+l}^* = 0$ (non-idling slot) and $\mu_{1,i}^* = 0, i = m, \cdots, m+l-1$ (idling slots can not be empty-ending), while Case 2 is due to $\mu_{3,m}^* = 0$ (non-idling slot) and $\mu_{2,i}^* = 0, i = m, \cdots, m+l-1$ (an idling slot can not be preceded by a delay-critical slot) in (7).

For Lemma 5, since $r_m^* > 0$ and $r_{m+l}^* > 0$, we have $\mu_{3,m}^* = \mu_{3,m+l}^* = 0$. Case 1 is further due to $\mu_{2,i}^* = 0, i = m, \cdots, m+l-2$ (as slots $i = m+1, \cdots, m+l-1$ are idle), and $\mu_{2,m+l-1}^* = 0$ (non delay-critical) in (7). Case 2 is further due to $\mu_{1,m}^* = 0$ (non empty-ending), and $\mu_{1,i}^* = 0, i = m+1, \cdots, m+l-1$ (as slots $i = m+1, \cdots, m+l-1$ are idle) in (7). The last case is further due to the combination of Case 1 and Case 2 (*i.e.*, $\mu_{1,i}^* = 0$ and $\mu_{2,i}^* = 0$, $i = m, \cdots, m+l-1$).

## Appendix C
## Packet Delay Lower and Upper Bounds

First, denote $t_{start,m}$ as the time when the first bit of packet $m$ is transmitted. Specially, let $t_{start,1} = 0$ because from a delay perspective, if there are any idling periods before the first packet transmission, the first packet is effectively delayed starting from time 0. Similarly, denote $t_{end,m}$ as the departure time of the last bit of packet $m$'s transmission. Note that both $t_{start,m}$ and $t_{end,m}$ are not necessarily aligned with slot boundaries. Also, $t_{start,m} \geq t_{end,m-1}$, and the equality holds only if there is no idling period between the departure time of packet $m-1$ and the start time of packet $m$'s transmission. By defining the *inter-departure time* of packet $m, m \in [1, \cdots, M]$, as $\phi_m \triangleq t_{end,m} - t_{end,m-1}$, with $t_{end,0} \triangleq 0$, the delay for packet $m$ can thus be computed as

$$q_m = \sum_{l=1}^{m} \phi_l - (m-1)\tau_s,$$

where $\sum_{l=1}^{m} \phi_l = t_{end,m}$ and $(m-1)\tau_s$ are the departure time and the arrival time of packet $m$, respectively.

Now, define the *virtual start time* of packet $m \in [1, \cdots, M]$ as $t_{start,m}^v \triangleq (t_{end,m-1} + t_{start,m})/2$, and the *virtual departure time* of packet $m \in [1, \cdots, M-1]$ as $t_{end,m}^v \triangleq (t_{end,m} + t_{start,m+1})/2$. Let $t_{end,M}^v \triangleq (M+D-1)\tau_s$, regardless of any potential idling slots after $t_{end,M}$. Subsequently, define the *virtual inter-departure time* of packet $m$ as

$$\phi_m^v \triangleq t_{end,m}^v - t_{start,m}^v.$$

Note that if there are no idling slots between $t_{end}(m-1)$ and $t_{start}(m)$, and between $t_{end}(m)$ and $t_{start}(m+1)$, $\phi_m^v$ corresponds to $\phi_m$. Otherwise, $\phi_m^v$ also incorporates half of the idling period between $t_{end}(m-1)$ and $t_{start}(m)$, and half of the idling period between $t_{end}(m)$ and $t_{start}(m+1)$, for $2 \leq m \leq M-1$. For the first packet, $\phi_1^v$ includes the entire idling period, if any, before its transmission, while for the last packet, $\phi_M^v$ includes the entire remaining idling period after $t_{end}(M)$, if any. Denote

$$q_m^v = \sum_{l=1}^{m} \phi_l^v - (m-1)\tau_s.$$

Note that since $t_{end,m}^v = t_{start,m+1}^v$ and $t_{end,m}^v \geq t_{end,m}$, we have $q_m^v \geq q_m$.

Due to the symmetry property of $\vec{B}$ and hence $\vec{r}^*$, as in Theorem 3, it is not difficult to show that, using a sample path trajectory of the forward running system and the corresponding time reversed system, the same symmetry property holds for the virtual inter-departure time vector $\vec{\phi}^v$ as well, *i.e.*,

$$E\{\phi_m^v\} = E\{\phi_{M+1-m}^v\}, \forall m.$$

Therefore,

$$\begin{aligned}
&\bar{q}(M) \overset{\Delta}{=} \tfrac{1}{M} \textstyle\sum_{m=1}^M E\{q_m\} \\
&\leq \tfrac{1}{M} \textstyle\sum_{m=1}^M E\{q_m^v\} \\
&= \tfrac{1}{M} \textstyle\sum_{m=1}^M [\sum_{l=1}^m E\{\phi_l^v\} - (m-1)\tau_s] \\
&\overset{(a)}{=} \tfrac{1}{M} \textstyle\sum_{m=1}^M (M-m+1) E\{\phi_m^v\} - \tfrac{\tau_s}{M} \sum_{m=1}^M (m-1) \\
&\overset{(b)}{=} \tfrac{1}{M} \textstyle\sum_{m=1}^M \tfrac{M+1}{2} E\{\phi_m^v\} - \tfrac{1}{M} \tfrac{M(M-1)}{2} \tau_s \\
&\overset{(c)}{=} \tfrac{M+1}{2M}(M+D-1)\tau_s - \tfrac{M-1}{2}\tau_s \\
&= \tau_s + \tfrac{M+1}{2M}(D-1)\tau_s,
\end{aligned} \tag{14}$$

where (a) holds by counting the number of occurrences of each item, the first term of (b) comes from the symmetry property $E\{\phi_m^v\} = E\{\phi_{M-m+1}^v\}, \forall m \in [1, \cdots, M]$, and equivalently, there are $(M+1)/2$ copies of each $E\{\phi_m^v\}$, and the first term of (c) is due to the fact that $\sum_{m=1}^M E\{\phi_m^v\} = (M+D-1)\tau_s$.

On the other hand, re-writing the average delay computation as

$$\begin{aligned}
\bar{q}(M) &\overset{\Delta}{=} \tfrac{1}{M} \textstyle\sum_{m=1}^M E\{q_m\} \\
&= E\{\tfrac{1}{M} \textstyle\sum_{m=1}^M \sum_{l=1}^m \phi_l\} - \tfrac{1}{M} \sum_{m=1}^M (m-1)\tau_s
\end{aligned}$$

we can see that $\sum_{l=1}^m \phi_l = t_{end,m}$ is used only once for each $m \in [1, \cdots, M]$ before taking the average $(1/M)$ and the expectation. Similar statement also holds true for $\sum_{l=1}^m \phi_l^v = t_{end,m}^v$ as in (14). To quantify the difference between $t_{end,m}^v$ and $t_{end,m}$, denote

$$\Delta t_m^v \overset{\Delta}{=} t_{end,m}^v - t_{end,m} = (t_{start}(m+1) - t_{end}(m))/2 \geq 0.$$

First note that any idling periods between time 0 and $t_{start}(1)$, and in between a packet transmission have no impact on $\Delta t_m^v$, $\forall m$. For a particular $m \in [1, \cdots, M-1]$, if there is one or more idle slot between $t_{end}(m)$ and $t_{start}(m+1)$, $\Delta t_m^v > 0$ and it equals to *half* of the idling period. However, any other idling periods between $t_{end}(j)$ and $t_{start}(j+1)$, $j \neq m$, have no impact on $\Delta t_m^v$. In the special case of $m = M$, if there are idling slots after packet $M$ transmission, we have $\Delta t_M^v = (M+D-1)\tau_s - t_{end,M}$. However, due to the symmetry property, there must exist a realization of the same length of an idling period between time 0 and $t_{start}(1)$ (which does not impact $\Delta t_1^v$). Effectively, the idling period after $t_{end,M}$, if any, only contributes half of its duration to $\Delta t_M^v$. Therefore, we have:

*Lemma 7:* Any idling period effectively at most contributes once to the difference in the average delay computation using $\vec{\phi}^v$ and $\vec{\phi}$, and the contribution is at most half of its duration.

This leaves us to count the *total idling duration* within $M + D - 1$ slots. Due to the delay constraint $D$, there are at least $\lfloor (M+D-1)/D \rfloor \geq M/D$ non-idling slots between slots 1 and $M + D - 1$. Therefore, we have,

$$\sum_{m=1}^M \left\{ \sum_{i=1}^m (\phi_i^v - \phi_i) \right\} \leq \tau_s[(M+D-1-M/D)/2]$$

Defining

$$\begin{aligned}
\delta &\overset{\Delta}{=} \tau_s[(M+D-1-M/D)/2]/M \\
&= \tau_s[(1-1/D) + (D-1)/M]/2,
\end{aligned}$$

we finally have

$$\bar{q}(M) \geq \tau_s \left[ 1 + \frac{M+1}{2M}(D-1) \right] - \delta = \tau_s \left[ \frac{D}{2} + \frac{1}{2D} \right].$$

Note that $\delta$ converges to $0.5\tau_s(1-1/D)$ when $M$ approaches infinity.

In case of static channels, there will be no idling periods under the optimal offline schedule. Thus, $\phi_m = \phi_m^v$, and the equality holds in (14).

## References

[1] R. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Info. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[2] D. P. Bertsekas, *Nonlinear programming*. Boston, MA: Athena Scientific, 1995.

[3] G. Caire, G. Taricco, and E. Biglieri, "Optimal power control over fading channels," *IEEE Trans. Info. Theory*, vol. 45, pp. 1468–1489, July 1999.

[4] W. Chen, "Energy-efficient packet transmissions with delay constraints for wireless communications," *Ph.D. dissertation, University of Southern California, Los Angeles, CA*, May 2007.

[5] W. Chen and U. Mitra, "Energy efficient scheduling with individual packet delay constraints," *IEEE INFOCOM, Barcelona*, 2006.

[6] W. Chen, M. J. Neely, and U. Mitra, "Energy efficient scheduling with individual packet delay constraints: Offline and online results," *IEEE INFOCOM 2007, Anchorage, AL, accepted*.

[7] B. Collins and R. Cruz, "Transmission policies for time varying channels with average delay constraints," *Proceedings of Allerton Conf. on Comm. Control, and Comp.*, pp. 709–717, 1999.

[8] A. Fu, E. Modiano, and J. Tsitsiklis, "Optimal transmission scheduling over a fading channel with energy and deadline constraints," *IEEE Trans. Wireless Communications*, vol. 5, no. 3, pp. 630–641, March 2006.

[9] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Info. Theory*, vol. 43, pp. 1986–1992, Nov. 1997.

[10] S. V. Hanly and D. N. C. Tse, "Multiaccess fading channels - Part II: Delay-limited capacities," *IEEE Trans. Info. Theory*, vol. 44, no. 7, pp. 2816–2831, Nov 1998.

[11] M. A. Khojastepour and A. Sabharwal, "Delay-constrained scheduling: power efficiency, filter design, and bounds," *IEEE INFOCOM, Hong Kong*, pp. 1939–1950, March 2004.

[12] X. Liu and A. Goldsmith, "Optimal power allocation over fading channels with stringent delay constraint," *Proc. ICC'02, New York, NY*, pp. 1416–1418, May 2002.

[13] M. J. Neely, "Optimal energy and delay tradeoffs for multi-user wireless downlinks," *IEEE INFOCOM, Barcelona*, 2006.

[14] R. Negi and J. Cioffi, "Delay-constrained capacity with causal feedback," *IEEE Trans. Info. Theory*, vol. 48, pp. 2478–2494, Sep 2002.

[15] L. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Tech.*, vol. 43, pp. 359–378, May 1994.

[16] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay-bounded packet scheduling of bursty traffic over wireless channels," *IEEE Trans. Info. Theory*, vol. 50, no. 1, pp. 125–144, Jan 2004.

[17] E. Uysal-Biyikoglu and A. E. Gamal, "On adaptive transmission for energy efficiency in wireless data networks," *IEEE Trans. Info. Theory*, vol. 50, no. 12, pp. 3081–3094, Dec 2004.

[18] E. Uysal-Biyikoglu, A. E. Gamal, and B. Prabhakar, "Adaptive transmission of variable-rate data over a fading channel for energy efficiency," *Proc. IEEE GLOBECOM 2002, Taipei, Taiwan*, Nov. 2002.

[19] E. Uysal-Biyikoglu, B. Prabhakar, and A. E. Gamal, "Energy-efficient packet transmission over a wireless link," *IEEE/ACM Trans. Networking*, vol. 10, no. 4, pp. 487–499, Aug 2002.

[20] M. A. Zafer and E. Modiano, "A calculus approach to minimum energy transmission policies with quality of service gurantees," *IEEE INFOCOM, Miami*, pp. 548–559, March 2005.