Optimal Universal Schedules for Discrete Broadcast*

Michael Langberg Alexander Sprintson Jehoshua Bruck

California Institute of Technology
Email: {mikel,spalex,bruck}@caltech.edu

Abstract

In this paper we study the scenario in which a server sends dynamic data over a single broadcast channel to a number of passive clients. We consider the data to consist of *discrete* packets, where each update is sent in a separate packet. On demand, each client listens to the channel in order to obtain the most recent data packet. Such scenarios arise in many practical applications such as the distribution of weather and traffic updates to wireless mobile devices and broadcasting stock price information over the Internet.

To satisfy a request, a client must listen to at least one packet from beginning to end. We thus consider the design of a broadcast schedule which minimizes the time that passes between a clients request and the time that it hears a new data packet, i.e., the *waiting time* of the client. Previous studies have addressed this objective, assuming that client requests are distributed uniformly over time. However, in the general setting, the clients behavior is difficult to predict and might not be known to the server. In this work we consider the design of *universal* schedules that guarantee a short waiting time for *any* possible client behavior. We define the model of dynamic broadcasting in the universal setting, and prove various results regarding the waiting time achievable in this framework.

1 Introduction

In this paper we investigate efficient schedules for sending dynamic data over lossless broadcast channels. We consider a system in which the server periodically transmits highly dynamic data to a number of passive clients. We study the case in which the data consists of *discrete* packets, where each update is sent in a separate packet. Each client listens to the channel in order to obtain the most recent data. Such systems have many practical applications such as in the distribution of weather and traffic updates to wireless mobile devices and in broadcasting stock price information over the Internet.

Our goal is to allow for each client to access the most recent data as soon as possible. In particular, we want to minimize the time elapsed since the client started to listen to the channel until it received the information. Designing efficient broadcast schedules (with respect to this objective) attracted a large body of research (see e.g., [1, 2, 4, 5] and references therein). To the best of our knowledge, all previous studies assumed that client requests are distributed uniformly over time. However, in general settings, the clients' behavior is difficult to predict. For example, there might be more requests in the top of the hour, as many clients want to synchronize their internal databases. Alternatively, the distribution of client requests may depend on various *global* events over which the server has no control. Finally, to take this to an extreme, one may consider a situation in which the server has no knowledge whatsoever on the distribution of client requests. The question of whether one can design scheduling strategies that allow a low expected waiting time experienced by the client in such scenario arises naturally. In this paper we concentrate on this question and focus on *universal* schedules; that is to say, schedules which must perform well for *any* possible request sequence.

[1]This work was supported in part by the Caltech Lee Center for Advanced Networking and by NSF grant ANI-0322475.

We consider the basic framework of scheduling packets with continuously updated content over a single broadcast channel. In this framework, packets of equal length (of say one unit) are being broadcasted periodically. The client must listen to at least one packet from beginning to end in order to satisfy a request. Our objective is to design schedules which minimize the time that passes between a clients request and the broadcast of a new item (referred to as the *waiting time* of the client), where we have no assumption on the client behavior. It turns out that this basic framework is interesting and poses major challenges. A schedule example is depicted in Fig. 1(a). In this example if a client request arrives at time t_1 the client will wait until time t_2 which is the arrival of the next packet (number 3); the request will be completed at time t_3. If a request arrives at time t_4, the client must wait until the transmission of packet 5 that begins at time t_5. Note that even through the client may listen to (part of) packet 4, it still needs to wait until the completion of item 5 in order to get a necessary update.

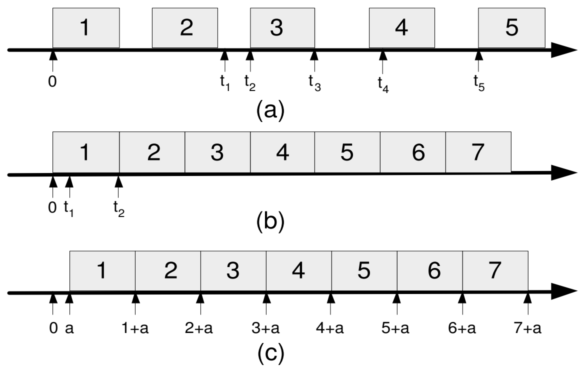

Figure 1: Examples of possible schedules

1.1 Universal schedules

We study the scenario in which the server does not have any knowledge of the clients request distribution. In this setting, our goal is to devise universal schedules whose waiting time is low for every possible distribution of client requests. The universal schedules we present effectively eliminate the need for the server to know the probability distribution of client requests.

Before we rigourously present our model and results, we briefly discuss several aspects of universal broadcasting. Consider the trivial schedule in which the packets are sent immediately one after another (depicted in Fig. 1(b)). What is the waiting time of such a schedule in the universal setting? Recall that we do not assume any distribution on client requests, and so we must analyze the waiting time on the *worst case* request distribution of the client. This corresponds, for example, to the *Quality of Service* (QoS) guaranteed by the presented schedule. That is, a bound on the waiting time of a client, no matter how it acts. In this case it is easily seen that the *quality* of the suggested schedule is not very good. Specifically, let ϵ be a small constant and consider a request given at time $t_1 = \epsilon$ (after a small portion of the first message has been transmitted, see Fig. 1(b)). The client issuing such a request must wait almost an entire unit of time (until t_2) until it will start hearing a *fresh* new message. This implies that the waiting time of this schedule in the universal setting is arbitrarily close to 1.

Can one design a schedule that is any better? It is not hard to verify that for any *deterministic* schedule of the server, the waiting time in the universal setting will always be arbitrarily close to 1 time unit. This follows from the simple fact that once the scheduling is determined there will always be a request time t which appears immediately after a broadcasted packet. We thus turn to consider *random* scheduling strategies of the server. In the random setting, the server has a distribution over schedules instead of a single scheduling. Instead of considering the waiting time of any request given at time t, we consider its *expected* waiting time. Notice that in this case the expectation is taken over the randomness of the server, as apposed to the standard scheduling models in which the expectation is taken over the assumed distribution of the clients.

Consider the following simple random scheduling. The server starts out by uniformly choosing a number between 0 and 1, say a. Afterwards, the server preforms the schedule which broadcasts a unit length packet at time $t = i + a$ for any integer i (see Fig. 1(c)). In other words, a schedule similar to the previous deterministic schedule is preformed, the only difference is the starting time of the first packet (which in this case is a instead of 0). Now consider a client request given at time t. It is no longer the case that we can find a specific time t that will have a long waiting time for a large portion of the possible server schedules. Namely, it can be seen that the expected waiting time of the client is independent of t, and equal to $1/2$ time units. This is a dramatic improvement over the deterministic case, and demonstrates that the introduction of randomness is crucial when considering the universal setting. The question whether this can be further improved now arises. It is not hard to see, that one cannot obtain a random schedule with expected universal waiting time of value less than $1/2$.

So are we done? Not quite. In the above examples we considered the client to be *oblivious* of its view of the schedule so far. In the QoS example previously mentioned, this implies that the quality analysis of the schedule assumes that the clients requests do not depend on the information previously obtained by the client (via previous data requests). This is not necessarily the case in general, where one must consider clients which base their requests on information previously broadcasted. Such scenarios arise naturally, for example, in the broadcast of stock market information where the broadcast of a *special* message may trigger many additional client requests (e.g., with the aim of receiving updates as soon as possible). We would like to compute the worst case (expected) waiting time in the setting in which clients may base their requests on their view of the broadcast channel.

Consider the random scheduling previously presented. This strategy has an expected waiting time of $1/2$ on every oblivious client request at time t. Namely, for any time t, the waiting time of a client request at time t averaged over all possible schedules of the server is exactly $1/2$. We claim that this is not the case when considering *adaptive* clients (i.e., clients who have knowledge of the previous data broadcasted). Consider the case in which each packet is a stock market update, and say a client requested the initial information at time 0. Such a client will wait an expected waiting time of $1/2$ until at time a the initial message will be received. After some time has passed (say 10 time units), our client is interested in an update. Based on the information gained by the first request, the client decides to place an additional request at time $t = 10 + a$ (so that its wait will not be long). However, unfortunately, the clients clock is running a bit slow and the actual request is at time $10 + a + \epsilon$ for some small $\epsilon > 0$. In this case the clients expected waiting time on the second request is $1 - \epsilon$ and not $1/2$! Notice that this expectation is taken only over schedules of the server which are consistent with the view of the client (in this example the server is deterministic after the choice of a - so there is only one such schedule).

Again we ask ourselves, if one can obtain stronger scheduling strategies which hold against any adaptive client. We stress that the adaptivity of a client must be in some sense *limited*, otherwise the question is resolved trivially. For example, if a client may place a request based on any information gathered in previous requests, a *malicious* client may wait until a broadcast is received (on the first request) and immediately (that is after time ϵ) place an additional request. Clearly the (expected) waiting time of the second request will be arbitrarily close to 1.

In our work, we focus on the case in which the adaptivity of the client is of unit length (the length of a single packet). This corresponds to assuming that after a client gains any information via previous requests, at least a unit of time will pass until an additional request is placed. One may also study the case in which the client has stronger or weaker adaptivity. The model and analysis techniques we develop in this paper can be used to analyze these cases as well. We touch upon this briefly in Section 4.

1.2 Rate considerations

Our goal so far was to minimize the waiting time for each client request. However, in some practical settings the *transmission rate*, i.e., the average number of packets sent over a period of time, is also important. Indeed, along with clients that listen to the channel from time to time, there might be clients that monitor the information all the time. Such clients prefer schedules with high transmission rates, which allow to receive as many updates as possible. For example, the schedules depicted in Figures 1(b) and (c) have high transmission rate, while the schedule depicted in Fig. 1(a) has low transmission rate. It turns out that for universal schedules there exists a tradeoff between the transmission rate and minimum worst case waiting time. In this paper we investigate this tradeoff and present universal schedules that provide minimum worst case waiting time subject to rate constraints.

1.3 Our results

In this paper we study universal broadcast scheduling, under the assumption that the client has adaptivity of unit length. We present a scheduling strategy which promises an expected waiting time strictly lower than 1, no matter when the request was placed, or what the viewed history of the channel was before the request. Specifically, our scheduling strategy guarantees a worst case expected waiting time of no more than $1/\sqrt{2} \simeq 0.7$. Moreover, we show that this is the best strategy possible. Namely, we show that no matter what scheduling strategy is used, it is impossible to obtain a universal waiting time of value less than $1/\sqrt{2}$.

The optimal schedule strategy we present is a random strategy. As mentioned above, this cannot be avoided, if a worst case waiting time less than 1 is obtained. This however implies that the transmission rate of our schedule is strictly less than 1. Our introduction of randomness on one hand has enabled us to obtain a worst case expected waiting time of $1/\sqrt{2}$, but on the other has admitted (random) transmission gaps in the channel which reduce the transmission rate below the optimal value of 1. Nevertheless, rather surprisingly, our optimal schedule has a high rate of $r = \frac{2}{1+\sqrt{2}} \simeq 0.82$. Moreover, consider a server which prefers to transmit with a rate higher than r. In this paper, we present for any larger rate r, a schedule of rate r with expected waiting time of no more than $\frac{2-r-\sqrt{2-2r}}{r}$ time units, no matter when the request was placed, or what the viewed history of the channel was before the request (notice that when $r = 1$, this waiting time is exactly 1). We show that this is the best schedule possible under certain restrictions on the server. Roughly speaking, the restriction we impose on the server is that its behavior after each packet is sent is governed by an identically distributed random variable. The tradeoff between the transmission rate and worst case expected waiting time is depicted in Figure 2.

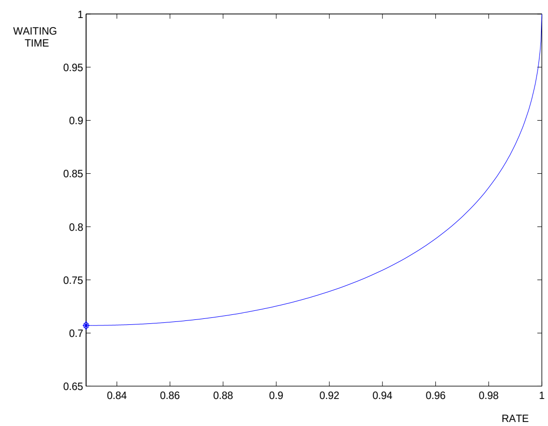

Figure 2: The tradeoff between the transmission rate and worst case expected waiting time. The asterisk at point $\left(\frac{2}{1+\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ represents the overall optimal schedule with respect to the universal setting.

The remainder of this paper is organized as follows. In Section 2 we present our model. In Section 3 we prove our main results on clients with unit adaptivity. Due to space limitations some of our results will appear without detailed proof. Finally, in Section 4 we discuss the case in which clients have various degrees of adaptivity.

2 Model

As mentioned in the Introduction, we are interesting in the design of a (random) schedule strategy for packets of unit length. A schedule can be defined by specifying, for each packet i, the amount of time that passes between the end of the transmission of packet $i-1$ and the beginning of the transmission of packet i (for simplicity, we assume that the transmission of packet 0 ends at time 0). We denote this time as the *interleaving time*.

Definition 1 (Schedule $\mathcal{S}$) *A* schedule *is a sequence of random variables $\{X_1, X_2, \dots\}$ such that X_i is the interleaving time for packet i.*

A schedule $\mathcal{S} = \{X_1, X_2, \dots\}$ can also be defined by its *transmission sequence* $\{T_1, T_2, \dots\}$, where T_n represents the time in which packet n was transmitted. Namely, $T_n = \sum_{i=1}^{n} X_i + n - 1$ for all $n \geq 1$.

Let $\mathcal{S}$ be a schedule, and suppose that a client request is placed at time t. We define the client waiting time as the time between t and beginning of the next packet.

Definition 2 (Waiting Time, $WT(\mathcal{S}, t)$) *The* Waiting Time *for a request at time t using a schedule $\mathcal{S} = \{X_1, X_2, \dots\}$ is defined to be $WT(\mathcal{S}, t) = T_n - t$, where n is the first packet for which it holds that $T_n \geq t$.*

Notice that $WT(\mathcal{S}, t)$ is a random variable. We denote the expectation of $WT(\mathcal{S}, t)$ by $EWT(\mathcal{S}, t) = E[WT(\mathcal{S}, t)]$.

2.1 Adaptive clients

We start by a few definitions. Let $\mathcal{S} = \{X_1, X_2, \dots\}$ be a (random) schedule. A *realization* R of $\mathcal{S}$ is a deterministic schedule $\{x_1, x_2, \dots\}$ that is in the domain of $\mathcal{S}$. We would like to define the notion of a *history* of $\mathcal{S}$ observed at time t. Let $V_t(x_1, \dots, x_\ell)$ be the event in which: (a) $\forall n \leq \ell \quad X_n = x_n$, (b) $\sum_{i=1}^{\ell} X_i + \ell - 1 \leq t$, and (c) $\sum_{i=1}^{\ell+1} X_i + \ell > t$. That is, the event in which (a) for $n \leq \ell$, the random variables X_n are equal to x_n, (b) the number of (partial) packets broadcasted until time t is at least ℓ, and (c) the $\ell + 1$'th package has not been transmitted up to time t. We call such an event a history of $\mathcal{S}$ at time t. Namely, any realization $R \in V_t(x_1, \dots, x_\ell)$ is completely described up to time t by the interleaving times $\{x_1, \dots, x_\ell\}$. Let $\mathcal{V}(\mathcal{S}, t)$ be the set of possible histories of $\mathcal{S}$ at time t. Finally, for any $V \in \mathcal{V}(\mathcal{S}, t)$ let $\mathcal{S}|V$ be the schedule distribution obtained by conditioning $\mathcal{S}$ on the event V.

Our goal is to design schedules that perform well with any behavior of incoming requests no matter what the viewed history of the channel was before the requests. In particular, we consider the case in which the client is *adaptive*, i.e., its behavior on time t depends on the history of the schedule up to time t. The clients might have different degrees of adaptivity.

Definition 3 (Degree of adaptivity, ω) *We say that a client is ω-adaptive if its actions at time t are based on a history $V \in \mathcal{V}(\mathcal{S}, t - \omega)$.*

Consider an ω-adaptive client which places a request at time t based on a history $V \in \mathcal{V}(\mathcal{S}, t - \omega)$ (i.e., viewed at time $t - \omega$). The expected waiting time of the client is defined to be

$$EWT_V(\mathcal{S}, t) = E[WT(\mathcal{S}|V, t)]$$

The worst case expected waiting time of the schedule $\mathcal{S}$ on ω-adaptive clients, $W(\mathcal{S}, \omega)$, is now defined as

$$\max\left(\max_{0 \leq t < \omega} EWT(\mathcal{S}, t),\ \max_{t \geq \omega} \max_{V \in \mathcal{V}(\mathcal{S}, t-\omega)} EWT_V(\mathcal{S}, t)\right).$$

Namely, $W(\mathcal{S}, \omega)$ bounds the waiting time of a client no matter at what time t its request is placed or what the history of the schedule was at time $t - \omega$. Notice that the first expression above addresses the case in which the client placed a request at time $t < \omega$. This implies that the client has not based his request on prior knowledge of the schedule.

3 Universal scheduling for $\omega = 1$

In the following section, we study the design of scheduling strategies in the case in which our clients are ω-adaptive for $\omega = 1$. We present a schedule $\mathcal{S}$ for which $W(\mathcal{S}, 1)$ is strictly less than 1. Namely, our schedule has $W(\mathcal{S}, 1) = 1/\sqrt{2}$. The schedule we present is of a simple nature as the random variables $X_1, X_2, \dots$ that define it are independent and identically distributed (i.e., i.i.d.). We show that our schedule is optimal. That is every other schedule $\mathcal{S}' = \{X'_1, X'_2, \dots\}$ has a corresponding waiting time $W(\mathcal{S}', 1)$ of value at least $1/\sqrt{2}$.

Finally, we show that our optimal schedule has *transmission rate* of value $\frac{2}{1+\sqrt{2}} \simeq 0.82$. For larger values of r, we present a scheduling strategy that has rate r, and worst case expected waiting time which is bounded by $\frac{2-r-\sqrt{2-2r}}{r}$ time units. Our schedule is defined by i.i.d. random variables, and is the best possible under such a construction.

3.1 Optimal schedule

We now turn to define a schedule $\mathcal{S}$ that has an expected waiting time (i.e., $W(\mathcal{S}, 1)$) which is bounded by $1/\sqrt{2}$. Our schedule is defined by a single random variable X. That is, we define $\mathcal{S}$ to be $\{X_1, X_2, \dots\}$, where each random variable X_i is independent and equals X. Recall the definition of $W(\mathcal{S}, 1)$:

$$\max\left(\max_{t\in[0,1)} EWT(\mathcal{S}, t),\ \max_{t\geq 1}\ \max_{V\in\mathcal{V}(\mathcal{S},t-1)} EWT_V(\mathcal{S}, t)\right).$$

The value of $W(\mathcal{S}, 1)$ depends on the random variable X in a complicated manner. Roughly speaking, one may argue that it is in our favor to define X to be as *uniform* as possible. As such a random variable seems to overcome the dependencies implied by conditioning over histories V. However, it can be seen that such a definition will not suffice, and cannot yield a worst case expected waiting time less than 1. Thus, we consider enhancing the uniform random variable. We observe, that the values of t which yield a worst case expected waiting time of 1 when X is uniform, benefit when X is deterministically set to be 0. Hence, we study the random variable X which is 0 with some probability p, and with the remaining probability is uniform. Setting p to be small enough, we show that such a random variable, on one hand, allows a sufficient amount of uniformity to overcome dependencies implied by conditioning over histories V, and on the other, guarantees a worst case expected waiting time strictly less that 1.

In what follows we define the random variable X and the schedule $\mathcal{S}$. Let $\mu > 0$ be a parameter that will be fixed in a later stage of our discussion. Let Z be a "random" variable which obtains the value 0 with probability 1. Let $U[0, s]$ be the uniform distribution on the interval $[0, s]$. Finally let $p = 1 - \sqrt{\frac{2\mu}{\mu+1}}$, and $s = \sqrt{2\mu(\mu + 1)}$.

Consider the schedule $\mathcal{S} = \{X_1, X_2, \dots\}$ in which each random variable X_i is independent and identically distributed. Namely, $X_i = X$ for all i where $X = pZ + (1 - p)U[0, s]$. It is not hard to verify that $E[X] = \mu$, and that the support of X is $[0, s]$. In the following theorem we analyze (for a range of values μ) the worst case expected waiting time of the schedule $\mathcal{S}$. Moreover, we show that by choosing μ to be $\frac{\sqrt{2}-1}{2}$ we obtain $W(\mathcal{S}, 1) = 1/\sqrt{2}$.

Theorem 1 *For any $\mu \in \left[0, \frac{\sqrt{2}-1}{2}\right]$ the worst case expected waiting time of $\mathcal{S}$ is $W(\mathcal{S}, 1) = 1 + 2\mu - \sqrt{2\mu(\mu + 1)}$. Specifically, setting $\mu = \frac{\sqrt{2}-1}{2}$ we obtain $W(\mathcal{S}, 1) = 1/\sqrt{2}$.*

Proof: To bound the value of $W(\mathcal{S}, 1)$ we must bound both the expressions $\max_{t\in[0,1)} EWT(\mathcal{S}, t)$ and $\max_{t\geq 1} \max_{V\in\mathcal{V}(\mathcal{S},t-1)} EWT_V(\mathcal{S}, t)$. We start by studying $EWT(\mathcal{S}, t)$ for $t \in [0, 1)$. The value of $EWT(\mathcal{S}, 0)$ (i.e., the expected waiting time on request at time 0) is exactly μ. It is not hard to verify that for $t \in (0, s]$, $EWT(\mathcal{S}, t)$ is equal to $p(1-t+\mu) + \frac{1-p}{s}\left(\int_0^t (x + 1 - t + \mu)dx + \int_t^s (x - t)dx\right)$, which is equal to $1 + 2\mu - \sqrt{2\mu(\mu + 1)}$ for each t.

For $t \in (s, 1)$ (notice that $s < 1$) we have $EWT(\mathcal{S}, t) = p(1 - t + \mu) + \frac{1-p}{s}\left(\int_0^s (x + 1 - t + \mu)dx\right) \leq EWT(\mathcal{S}, s)$. As $1 + 2\mu - \sqrt{2\mu(\mu + 1)} \geq \mu$ for $\mu \in [0, 1]$, we conclude that $\max_{t\in[0,1)} EWT(\mathcal{S}, t) = 1 + 2\mu - \sqrt{2\mu(\mu + 1)}$.

Now consider any $t \geq 1$, and any history $V \in \mathcal{V}(\mathcal{S}, t - 1)$. We would like to analyze the waiting time $EWT_V(\mathcal{S}, t)$. Let V be the event $V_t(x_1, \dots, x_\ell)$. Let $t_\ell = \sum_{i=1}^{\ell} x_i + \ell - 1$. By the fact that $V \in \mathcal{V}(\mathcal{S}, t - 1)$ we have that $t - t_\ell \geq 1$. Furthermore, as X is bounded by s we conclude that $t - t_\ell < 2 + s$. Let $t^* = t - t_\ell - 1$. Namely, $t^* \in [0, 1+s)$ corresponds to the time which has passed between the end of the transmission of the ℓ'th packet and the request t. We consider two cases:

Case 1 Assume that $t^* \in [0, 1)$. In this case we claim that $EWT_V(\mathcal{S}, t) = EWT(\mathcal{S}, t^*)$. The claim follows from the definition of $\mathcal{S}$. In the case under discussion, $t_\ell + 1 > t - 1$, implying that no knowledge on the value of the interleaving time $X_{\ell+1}$ appears in V. Specifically, $\mathcal{S}$ conditioned on the event V (i.e., $\mathcal{S}|V$) is a random process defined by the interleaving times $\{x_1, x_2, \dots, x_\ell, X_{\ell+1}, X_{\ell+2}, \dots\}$, where $X_i = X$ for $i \geq \ell + 1$. Thus the process $\mathcal{S}|V$ can be viewed as the deterministic finite schedule $\{x_1, \dots, x_\ell\}$ followed by the schedule $\mathcal{S}$ (which now starts at time $t_\ell + 1$). We conclude that for any waiting time z and any $t \geq t_\ell + 1$ the probability that a request placed at t in $\mathcal{S}|V$ will have waiting time z is exactly the probability that a request placed at t^* in $\mathcal{S}$ will have waiting time z. This suffices to prove our claim, and implies that in this case $EWT_V(\mathcal{S}, t) \leq 1 + 2\mu - \sqrt{2\mu(\mu+1)}$.

Case 2 In this case we assume that $t^* \in [1, 1+s)$. Similar to the previous case, $\mathcal{S}|V$ is a random process defined by the interleaving times $\{x_1, x_2, \dots, x_\ell, X_{\ell+1}, X_{\ell+2}, \dots\}$. But in this case $X_{\ell+1}$ is not distributed as X (the variables X_i for $i \geq \ell + 2$ are still equal to X). Rather the history V implies that $X_{\ell+1} \notin [0, t^* - 1]$. Thus the distribution of $X_{\ell+1}$ is that of X conditioned on this event. We denote this distribution as $X|_{>t^*-1}$.

Using an argument similar to that applied in case 1, we now claim that $EWT_V(\mathcal{S}, t) = EWT(\mathcal{S}', 1)$ where $\mathcal{S}' = \{X|_{>t^*-1}, X_2, X_3, \dots\}$; as before $X_i = X$ for $i \geq 2$. As $s < 1$, it is not hard to verify that $EWT(\mathcal{S}', 1)$ is equal to $\mu + \frac{1}{s-t^*+1}\int_0^{s-t^*+1} x\,dx = \mu + \frac{s-t^*+1}{2} \leq \mu + \frac{s}{2} = \mu + \sqrt{\mu(\mu+1)/2} \leq 1 + 2\mu - \sqrt{2\mu(\mu+1)}$. The last inequality follows from the fact that $\mu \leq 2/7$ for our choice of μ. □

3.2 Proof of optimality

We now prove that any schedule $\mathcal{S} = \{X_1, X_2, \dots\}$ has a worst case expected waiting time $W(\mathcal{S}, 1)$ of value greater or equal to $1/\sqrt{2}$. This is done in a few steps. In what follows we prove the assertion when $\mathcal{S}$ is defined by i.i.d. random variables X which, in turn, are defined by a continuous density function. The general case, in which each random variable X_n in $\mathcal{S}$ may be arbitrarily distributed and may depend on X_i for $i < n$, is discussed after the proof mentioned above.

Let X be a random variable and $\mathcal{S} = \{X_1, X_2, \dots\}$ where $X_i = X$ for all i. Let F be the distribution function of X, and as discussed above we assume $F(x) = \int_0^x f(x)dx$ for some continuous density function f. Let $\mu = \int_0^\infty (1 - F(x))dx$ be the expectation of X.

Theorem 2 *There exists a* $t \in [0, 1)$ *such that* $EWT(\mathcal{S}, t) \geq 1/\sqrt{2}$. *Specifically,* $W(\mathcal{S}, 1) \geq 1/\sqrt{2}$.

Proof: We start by considering the case in which $\sqrt{2\mu(\mu+1)} \leq 1$. In this case, we prove our assertion by presenting a schedule $\mathcal{S}' = \{X_1', X_2', \dots\}$ with the following properties. (a) Each X_i' is i.i.d. and equal to a random variable X'. (b) The expectation of X' is μ. (c) $\max_{t\in[0,1)} EWT(\mathcal{S}, t) \geq \max_{t\in[0,1)} EWT(\mathcal{S}', t) \geq 1 + 2\mu - \sqrt{2\mu(\mu+1)}$.

It is not hard to verify that $1 + 2\mu - \sqrt{2\mu(\mu+1)} \geq 1/\sqrt{2}$ for all $\mu \geq 0$. Thus, it is left to define X' and prove the above statements (b) and (c). X' is defined as in Theorem 1, namely $X' = pZ + (1-p)U[0, s]$ for $p = 1 - \sqrt{\frac{2\mu}{\mu+1}}$ and $s = \sqrt{2\mu(\mu+1)}$. It is not hard to verify that $E[X'] = \mu$.

To prove statement (c), notice that it is implicit in the proof of Theorem 1 that $\max_{t\in[0,1)} EWT(\mathcal{S}', t) = 1 + 2\mu - \sqrt{2\mu(\mu+1)}$. We now show that $\max_{t\in[0,1)} EWT(\mathcal{S}, t) \geq \max_{t\in[0,1)} EWT(\mathcal{S}', t)$.

It can be seen, using analysis similar to that appearing in Theorem 1, that for $t \in [0, 1]$

$$EWT(\mathcal{S}, t) = \mu + (\mu + 1)F(t) - t.$$

Assume for sake of contradiction that $\max_{t\in[0,1)} EWT(\mathcal{S},t) < \max_{t\in[0,1)} EWT(\mathcal{S}',t)$. This implies that for each $t \in [0,1)$ it holds that

$$EWT(\mathcal{S},t) = \mu + (\mu+1)F(t) - t < 1 + 2\mu - \sqrt{2\mu(\mu+1)}.$$

Thus, $1 - F(t) > \sqrt{\frac{2\mu}{\mu+1}} - \frac{t}{\mu+1}$. However, this implies that the expectation of X is more than μ:

$$\int_0^\infty (1-F(x))\mathrm{d}x > \int_0^s \left(\sqrt{\frac{2\mu}{\mu+1}} - \frac{x}{\mu+1}\right)\mathrm{d}x = \mu.$$

Which is a contradiction (here we use the fact that $s \le 1$).

We are left with the case $\sqrt{2\mu(\mu+1)} > 1$. If $\mu < 1/\sqrt{2}$, we follow the lines of proof given above with a different random variable X' defined as follows: $X' = p_1 Z + (1-p_1-p_2)U[0,1] + p_2 Z'$, where Z' is a random variable which has the value 1 with probability 1, $p_1 = \frac{1-2\mu^2}{2(1+\mu)}$ and $p_2 = \frac{2\mu^2+2\mu-1}{2(1+\mu)}$. It can be verified that $E[X'] = \mu$, $\max_{t\in[0,1)} EWT[S',t) = \frac{1}{2} + \mu - \mu^2$, and that $\frac{1}{2} + \mu - \mu^2 \ge 1/\sqrt{2}$ for our value of μ.

Assume by way of contradiction that $\max_{t\in[0,1)} EWT(\mathcal{S},t) < \max_{t\in[0,1)} EWT(\mathcal{S}',t)$. This implies for $t \in [0,1)$ that

$$EWT(\mathcal{S},t) = \mu + (\mu+1)F(t) - t < \frac{1}{2} + \mu - \mu^2.$$

Thus,

$$1 - F(t) > \frac{1/2 + \mu + \mu^2}{2(1+\mu)} - \frac{t}{1+\mu}.$$

This implies that the expectation $E[X]$ of X is greater than

$$\int_0^1 (1-F(x))\mathrm{d}x > \int_0^1 \frac{1/2+\mu+\mu^2}{2(1+\mu)} - \frac{t}{1+\mu} = \mu,$$

resulting in a contradiction.

Finally, for $\mu \ge 1/\sqrt{2}$, roughly speaking, it can be seen that a random client has expected waiting time at least $\frac{1+\mu}{2}$ (here the expectation is taken over both the client and the server). This implies a worst case expected waiting time $\ge 1/\sqrt{2}$ (detailed proof omitted). □

We now sketch the main ideas which enable us the prove an analog of Theorem 2 for a general schedule $\mathcal{S} = \{X_1, X_2, \dots\}$, in which each random variable X_n may be arbitrarily distributed and may depend on X_i for $i < n$. To prove the case in which $\mathcal{S}$ is defined by i.i.d random variables X that lack a continuous density function, we approximate X by a random variable defined by a continuous density function. The general case is now reduced to the previous one. Namely, for any schedule $\mathcal{S} = \{X_1, X_2, \dots\}$ we show that one can obtain a schedule $\mathcal{S}'$ defined by a single i.i.d. random variable X' where $W(\mathcal{S},1) \ge W(\mathcal{S}',1) - \epsilon$ for any $\epsilon > 0$.

The variable X' we suggest is defined by its expectation μ as in Theorem 2. It is left to define the value of μ which, roughly speaking, is set to be the infimum over all $n = \{1,2,3,\dots\}$ and any history $V_t(x_1,\dots,x_{n-1}) \in \mathcal{V}(\mathcal{S},t)$ of the expectation of X_n given that $X_i = x_i$ for $i < n$. Detailed proof is omitted.

3.3 Optimal schedules for large rates

The transmission rate of a schedule $\mathcal{S} = \{X_1, X_2, \dots\}$ is defined to be the expected amount of time in which the channel is in use.

Definition 4 (Transmission rate) *Let R_t be the expected number of packets sent in $\mathcal{S} = \{X_1, X_2, \dots\}$ up to time t. The transmission rate r of $\mathcal{S}$ is defined to be $\lim_{t\to\infty} \frac{R_t}{t}$.*

In this section, we consider schedules $\mathcal{S}$ which are defined by a series of i.i.d. random variables X. For such schedules it can be shown (e.g., [3]) that the transmission rate of $\mathcal{S}$ is $\frac{1}{1+\mu}$ where μ is the expectation of X. We study the problem of finding *good* schedules (with respect to the universal objective) which have a prespecified rate r.

The optimal schedule presented in Theorem 1 has rate $r = \frac{2}{1+\sqrt{2}}$. We now present for any rate larger than r, a schedule of rate r with expected waiting time of no more than $\frac{2-r-\sqrt{2-2r}}{r}$ time units, no matter when the request was placed, or what the viewed history of the channel was before the request (notice that when $r = 1$, this waiting time is exactly 1). Our schedules are those defined in Theorem 1. They are defined by i.i.d. random variables, and are the best possible under such a construction. The tradeoff between the transmission rate of our schedules and the worst case waiting time is depicted in Figure 2.

Let $r \geq \frac{2}{1+\sqrt{2}}$ be a prespecified transmission rate, and let $\mu = \frac{1}{r} - 1$. It is not hard to verify that Theorems 1 and 2 imply

Corollary 1 *For any $r \in \left[\frac{2}{1+\sqrt{2}}, 1\right]$ there exists a schedule $\mathcal{S} = \{X_1, X_2, \dots\}$ with rate r and worst case expected waiting time of $W(\mathcal{S}, 1) = \frac{2-r-\sqrt{2-2r}}{r}$. Moreover, any schedule $\mathcal{S}' = \{X'_1, X'_2, \dots\}$ in which X'_n are i.i.d with rate r satisfies $W(\mathcal{S}', 1) \geq W(\mathcal{S}, 1)$.*

4 Concluding remarks

In this paper we have defined and addressed the design of universal broadcast schedules. For clients of unit adaptivity, we have presented a schedule which guarantees a worst case expected waiting time of at most $1/\sqrt{2}$, no matter when the client request is placed or what the history of the broadcast channel is before the request. Moreover, we have shown that this is the best schedule possible. Our optimal schedule has a transmission rate of $r \simeq 0.82$. For larger values of r we have presented a tight (subject to certain restrictions on the server) analysis of the tradeoff between the transmission rate and the minimum worst case expected waiting time.

The question whether such analysis can be given for any adaptivity ω of the client now arises naturally. Preliminary results show that for large values of ω a worst case expected waiting time of $\frac{1}{2} + \frac{O(1)}{\omega}$ is obtainable. Moreover, one cannot obtain a schedule with worst case expected waiting time less that $\frac{1}{2} + \frac{c}{\omega}$ for some small constant $c > 0$ (implying that these results are essentially tight). For small values of ω a worst case expected waiting time of $1 - O(\omega)$ is obtainable and again is essentially the best possible. We are currently studying intermediate degrees of adaptivity in aim to present tight bounds for every value of ω.

Acknowledgements

We would like to thank Kevin Foltz for useful discussions.

References

[1] M.H. Ammar. Response Time in a Teletext System: An Individual User's Perspective. *IEEE Trans. Comm.*, COM-35(11):1159–1170, 1987.

[2] A. Bar-Noy, R. Bhatia, J. Naor, and B. Schieber. Minimizing Service and Operation Costs of Periodic Scheduling. In *Proceedings of SODA*, pages 11–20, 1998.

[3] W. Feller. *An Introduction to Probability Theory and Its Applicatitons*. John Wiley & Sons, Inc., 1966.

[4] K. Foltz and J. Bruck. Splitting Schedules for Internet Broadcast Communications. *IEEE Transactions on Information Theory*, 48(2):345–358, 2002.

[5] C. Kenyon, N. Schabanel, and N.E. Young. Polynomial-Time Approximation Scheme for Data Broadcast. In *Proceedings of STOC*, pages 659–666, May 2000.