Response Time in Data Broadcast Systems: Mean, Variance and Trade-Off *

Shu Jiang Nitin H. Vaidya
Department of Computer Science
Texas A&M University
College Station, TX 77843-3112, USA
Email: {jiangs,vaidya}@cs.tamu.edu
URL: http://www.cs.tamu.edu/faculty/vaidya/mobile.html

Abstract

Data broadcast has been suggested as a promising method of information dissemination [1, 15]. In such an environment, the information server cannot afford to serve the requests from a large population of users individually. Instead, the server uses a broadcast channel to deliver information to all users. A single transmission of a data item satisfies all pending requests for that item. The response time of a request depends on the broadcast time of the desired data item, which is scheduled by the server according to the overall demands for various data items. Therefore, the response time may vary in a large range. We argue that, in addition to mean response time, the variance of response time should also be taken into account by the broadcast scheduler.

In this paper, we address the issue of variance optimization in regard to response time. Building on our previous research on mean response time optimization, we propose an algorithm which can minimize the variance of response time. Furthermore, we evaluate an algorithm that facilitates a trade-off between the mean and variance of response time. Numerical examples that illustrate the performance of our algorithms are also presented.

1 Introduction

In any client/server information system, user response time is one of the most important factors to evaluate the system's quality of service. It is even more critical in a broadcast data delivery system [5]. In a broadcast system, server plays an active role and broadcasts data items to the whole user community repeatedly, whereas any user who desires a particular data item listens to the channel until the data is broadcasted.

Needless to say, the schedule of broadcast affects user response time. Several researchers have proposed various scheduling schemes [3, 4, 6, 7, 8, 10, 14]. But almost all of them evaluate the effectiveness of scheduling schemes based on how they reduce the overall mean response time. The variance of response time has long been neglected. In the real world, it is hard to find two users having exactly same demand patterns. Actually, some users' demand patterns may largely deviate from the overall demand pattern and their own mean response time may be much worse than the overall mean. In this paper, we address this problem by introducing variance of response time as a performance metric. Contributions of this paper are as follows:

- The paper determines the relationship between a broadcast schedule and the variance of response time it may achieve. Starting from the analytical results, we developed a condition satisfying which results in minimal variance of response time. Based on this condition, we propose and evaluate an algorithm that can reduce the variance.
- In general, the objective in designing broadcast schedule is likely to be to achieve both low mean and low variance of response time. However, these two goals are often contradictory. We propose and evaluate an algorithm that achieves a trade-off between the mean and variance.

*This research is supported in part by Texas Advanced Technology Program under grants 009741-052-C and 010115-248, and National Science Foundation Grant MIP-9423735.

The rest of the paper is organized as follows. In Section 2, we define the problem and introduce notations. Section 3 reports our analysis results regarding the relationship between broadcast schedule and response time. Those results are then used in Section 4 to propose scheduling schemes which can minimize the mean response time, reduce the variance of response time, or implement a balance between these two metrics. Section 5 discusses our simulations and some numerical results. We summarize our conclusions in Section 6.

2 Problem Definition

The focus of this paper is on a *pure push-based* system [2] in which server broadcasts data items based on a known demand distribution for the various items.[1] We define the demand probability of item i as the probability that item i is requested in a client request and denote it with p_i. Let M be the total number of available items at the server and these items are numbered from 1 to M. It holds that $\sum_{i=1}^{M} p_i = 1$. The size of an item is another important factor to consider when server makes broadcast schedule. We measure the item size (or length) in terms of time taken when being broadcasted. l_i represents length of item i.

Response time of a request is defined as the duration of time from when the request is made until the desired item starts transmission on channel, i.e. the waiting time a user has to spend getting the request satisfied[5]. It is important to minimize the response time (in some literatures it is also called *access time*[6, 9, 13]) so as to reduce the idle time at the users.

The mean of response time has long been the primary performance metric. Several scheduling schemes have been proposed which are able to reduce or even minimize the mean response time [12, 13, 14]. However, minimizing the mean response time most benefits a "virtual" user whose request pattern happens to be coincident with the overall item demand pattern on which the broadcast schedule is based. An individual user whose demand pattern differs from the overall demand pattern may experience a mean response time greatly worse than the optimal value. To accurately evaluate the quality of service experienced by real users, the variance of response time should also be taken into account.

[1] Our ideas can also be applied to a *pull-based* system by replacing demand probability p_i in the discussion below with the number of requests pending for item i.

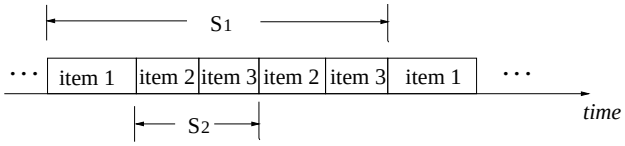

Figure 1: An example broadcast schedule

In the next section, we will determine a condition satisfying which results in minimal variance.

3 Analysis

First, we make an assumption about the user request generation process. As pointed out in [12], when the user population is large enough, we may assume that the aggregate request generation process is *Poisson* with constant rate.

Second, for the theoretical development, we consider broadcast schedules with Equal Spacing property. In such a schedule, the transmissions of any particular item on broadcast channel are equally spaced. Let $s_i, 1 \leq i \leq M$ be the spacing between consecutive instances of item i. We refer to the vector $< s_1, s_2, \ldots, s_M >$ as *schedule vector*. Figure 1 is a snapshot of a broadcast schedule that illustrates the concept of item spacing on broadcast.

Based on the two assumptions above, we can calculate the mean response time μ and the variance of response time σ^2, as follows. Detailed derivations are presented in Appendix I.

$$\mu = \frac{1}{2}\sum_{i=1}^{M} s_i p_i \tag{1}$$

and

$$\sigma^2 = \frac{1}{3}\sum_{i=1}^{M} p_i s_i^2 - (\frac{1}{2}\sum_{i=1}^{M} s_i p_i)^2 \tag{2}$$

or

$$\sigma^2 = \frac{1}{3}\sum_{i=1}^{M} p_i s_i^2 - \mu^2 \tag{3}$$

Based on the above expressions, we have obtained two useful results, of which the first one was presented in an earlier paper [13].

Previously Known Result: Minimizing the Mean Response Time[13]

Note that the expected mean response time and variance of response time are only decided by the schedule vector (i.e., by s_i's). Using the expression for mean response time, [13] derives a property satisfying which results in a schedule that minimizes the mean response time μ. Specifically, if the equality below is satisfied, then the mean response time is minimized.

$$\frac{s_i^2 p_i}{l_i} = constant, \forall i, 1 \leq i \leq M \qquad (4)$$

New Result: Minimizing the Variance of Response Time

Similar to the above property for minimizing the mean, we found the property which, if satisfied, minimizes the variance of response time.

Theorem 1 *Given the demand probability p_i of each item i, the minimal* variance of response time, *σ^2, is achieved when the schedule vector possesses the following property, assuming that transmissions of each item i are equally spaced by s_i.*

$$\frac{p_i s_i^2}{l_i}(\frac{2}{3}s_i - \mu) = constant, \forall i, 1 \leq i \leq M \qquad (5)$$

Appendix II presents the proof.

The above two results provide valuable insight into the relationship between the schedule vector and the quality of service, as well as the theoretical basis for designing the scheduling algorithms. In the next section, we introduce a broadcast scheduling scheme which is based on these observations. We also evaluate an algorithm that can trade the mean with variance.

4 Scheduling Algorithms

The results stated above imply that minimal mean or variance of response time can be achieved if the schedule used by server satisfies the condition in Equation 4 or 5, respectively. Unfortunately, it is intractable to find an optimal schedule. Therefore, we propose a heuristic-based scheduling scheme by which server makes the decision regarding which item to broadcast next. Whenever an item finishes broadcasting, the server calls the algorithm presented below to choose next appropriate item. The algorithm uses a *decision rule* motivated by the above analytical results. Our algorithms attempt to achieve the equality in Equation 4 or 5, depending on whether mean or variance is to be minimized, respectively. We later present another algorithm that can trade the mean response time with the variance of response time.

The first algorithm below, for reducing mean response time, appeared in out previous work [13]. The new algorithms proposed in this paper are based on this algorithm.

Reducing Mean Response Time [13]

Let Q be the current time and R_i be the time when item i was most recently transmitted.(If item i has never been broadcasted, R_i is initialized to -1.) Define F_i as

$$F_i = (Q - R_i)^2 p_i / li \qquad (6)$$

F_i is defined similar to the left hand side of Equation 4. Notice that Q changes continually and R_i is updated whenever item i is transmitted. To keep the values of all F_i's as close to each other as possible, the item j with maximum F value is broadcasted.

Algorithm for reducing mean response time [13]:

Step 1. For each item i, $1 \leq i \leq M$, update the value of F_i.
Step 2. Determine maximum F_i over all items. Let F_{max} denote the maximum value.
Step 3. Choose item j such that $F_j = F_{max}$. If this equality holds for more than one item, choose any one of them arbitrarily.
Step 4. Broadcast item j.
Step 5. $R_j = Q$.

The definition of F_i in this algorithm is inspired by Equation 4. [13] has showed that the above algorithm results in near-optimal mean response time. In the rest of this paper, we will refer to it as *Mean Optimal Algorithm.*

Proposed Algorithm for Reducing Variance of Response Time

We can modify the above algorithm to reduce the variance of response time by replacing the definition of F_i with the following one, motivated by Equations 5 and 1.(Note that we replace s_i in Equations 5 and 1 with

$(Q - R_i)$ to obtain the expression below.)

$$F_i = \frac{p_i(Q-R_i)^2}{l_i}\Big(\frac{2}{3}(Q-R_i) - \frac{1}{2}\sum_{i=1}^{M} p_i(Q-R_i)\Big) \quad (7)$$

With this definition, we are now trying to maintain the equality in Equation 5 to the extent possible. We will refer to the new algorithm as *Variance Optimal Algorithm* in next section. Note that the name *Variance Optimal* may be a misnomer, as the algorithm is not *proved* to achieve near-optimal variance (as we do not know a tight lower bound on variance).

Proposed Algorithm to Achieve a Trade-Off Between Mean and Variance

In general, minimal mean and minimal variance of response time are two contradictory goals. When mean response time is reduced to minimal, the variance may climb to an unacceptable high. If we minimize the variance, mean response time may become too large. To achieve a trade-off between a small mean and a small variance response time, we consider a third algorithm that attempts to achieve the equality below.

$$\frac{s_i^{\alpha} p_i}{l_i} = constant, \forall i, 1 \le i \le M \quad (8)$$

When $\alpha = 2$, the above equation reduces to Equation 4. Also, observe that the dominant exponent of s_i in Equation 5 is 3. Therefore, we expect that a scheduling algorithm that attempts to achieve the above equality, with $\alpha = 3$, will have performance approaching that of the *variance optimal algorithm* presented above. Based on Equation 8, we present a new expression to calculate F_i for each item.

$$F_i = (Q - R_i)^{\alpha} p_i / li, 2 \le \alpha \le 3 \quad (9)$$

The scheduling algorithm that uses the above F_i expression will be referred to as α-*algorithm*. When $\alpha = 2$, the α-algorithm reduces to Mean Optimal Algorithm. The α-algorithm was also evaluated by Su and Tassiulas [12]. They simulated an algorithm, equivalent to the α-algorithm, for various values of α, and empirically showed that $\alpha = 2$ minimizes the mean response time. We obtained the same result analytically in our prior work [13]. Su and Tassiulas, however, did not consider the impact of varying α on the variance of the response time. When α is picked close to 3, it is expected to produce a schedule which can make the variance of response time small(due to the cubic term in Equation 5). Although we cannot provide any analytical evidence for this claim, the simulation results in next section support it indeed.

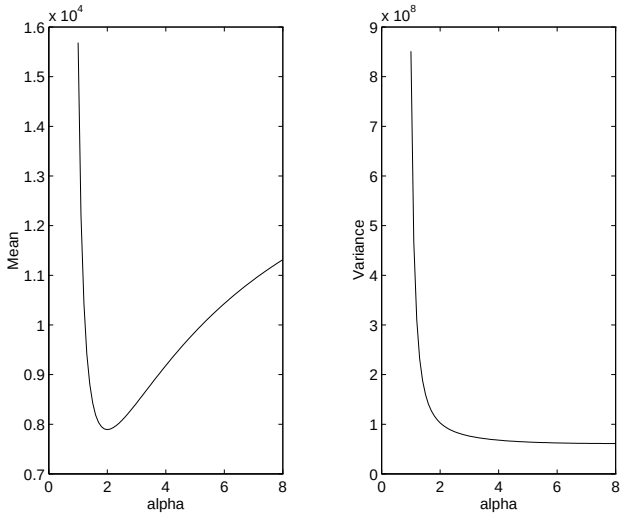

Figure 2: The lower bounds on mean and variance of response time when α-algorithm is used as scheduling algorithm and other system parameter settings are: $M = 250$, $\theta = 0.75$, *Increasing Length distribution*(θ and length distribution are defined in Section 5).

As shown in Appendix III, it is possible to derive good lower bounds on mean and variance of response time achieved by the α-algorithm. The lower bounds for one set of length and demand probability distributions are plotted in Figure 2. Our experience shows that the α-algorithm typically yields mean and variance quite close to the lower bounds. Thus, Figure 2 shows how the value of α affects the mean-variance trade-off.[2]

5 Performance Evaluation

In this section, we present some numerical results from our simulation of a broadcast data delivery system. The server uses various algorithms we presented above to do scheduling. The user requests are generated according to a Poisson process. For each experiment, 1 million requests are generated and served. Other

[2] An alternative approach to achieve a trade-off between the mean and variance would be to define F_i as a linear interpolation between the expressions used for mean optimal and variance optimal algorithms. We have not evaluated this alternative approach as yet.

simulation parameters are described below.

5.1 Demand Probability Distribution Of Items

In our simulation, the demand probabilities of all items follow Zipf distribution, with item 1 being the most frequently requested, and item M being the least frequently requested. The Zipf distribution may be expressed as follows:

$$p_i = c\left(\frac{1}{i}\right)^{\theta}, 1 \leq i \leq M$$

where $c = \frac{1}{\sum_{i=1}^{M}(\frac{1}{i})^{\theta}}$ is a normalizing factor, and θ is a parameter named *access skew coefficient*. When $\theta = 0$, Zipf distribution reduces to a uniform distribution with each item equally likely to be requested. However, the distribution becomes increasingly "skewed" as θ increases(that is, the difference among items with respect to the degree of popularity becomes more significant).

5.2 Length Distribution Of Items

The following three length distributions are considered in our simulations:

1. Equal Length Distribution:
   All items are equally sized and the size is 1, without loss of generality.

2. Increasing Length Distribution:
   In this case, the lengths of M items follow an increasing function, i.e. item 1, the most popular item, is the smallest item whereas item M, the most unpopular item, is the longest item in terms of transmission time. The length distribution function is as follows,

   $$l_i = l_{min} + \frac{(i-1)(l_{max} - l_{min})}{M-1}$$

   with $l_{min} = 1$ and $l_{max} = 250$.

3. Decreasing Length Distribution:
   In this case, the length distribution function is

   $$l_i = l_{max} - \frac{(i-1)(l_{max} - l_{min})}{M-1}$$

   with $l_{min} = 1$ and $l_{max} = 250$.

M	250
θ	0.25,0.5,0.75,1.0,1.25,1.5
l_i	Equal, Increasing, Decreasing
α	2.2, 2.6, 3

Table 1: Parameter Settings

5.3 Simulation Results

Table 1 shows the parameter settings for our simulations. We conducted a number of experiments under different combinations of the parameter settings. In each experiment, the response time of every request is captured. After sampling 1 million requests, we plot the mean and variance values in the following figures.

5.3.1 Validation of algorithms

As we mentioned before, our algorithms are heuristic based. In each algorithm, we define a variable F_i for each item i and obtain a variable group $\{F_1, F_2, \ldots, F_M\}$. The variable values keep changing and depend on the broadcast schedule generated by the algorithm. As we know, an ideal schedule should maintain the equality in Equation 4, 5 or 8 respectively. Since the definition of F_i is derived from one of the equalities, a schedule that can make the variable group "small", i.e. all variable values in the group are close to each other, is desired. To reach the goal, we manipulate the changes of F_i values by choosing the item with maximum value to broadcast and thus "pulling it back". In order to verify that this heuristic does work, we record the values of F_{max}'s in simulation experiments. Figure 3 plots the data we captured in an experiment. Clearly, the F_i variable values are effectively bounded. Similar results are obtained for other algorithms and presented in [11].

5.3.2 Equal Length Case

In the first simulation experiment, we let all items be of size 1 and examine the user response time when demand distribution of items varies. The simulation results are shown in Figure 4. In this figure, the graph whose y-axis is labeled "Mean" plots the mean response time when using different scheduling algorithms. Also, the second graph with y-axis labeled "Variance" plots the variance of response time. The number marked on each curve in these graphs is the value of θ used for that curve. From left to right along

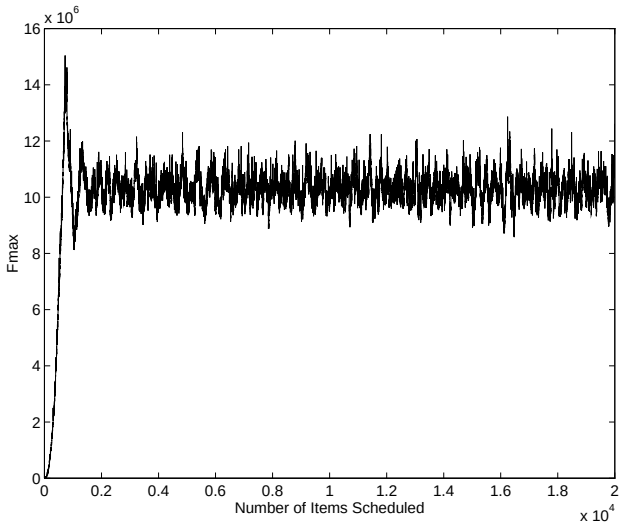

Figure 3: The change of F_{max} (α-algorithm with $\alpha = 2.6$)

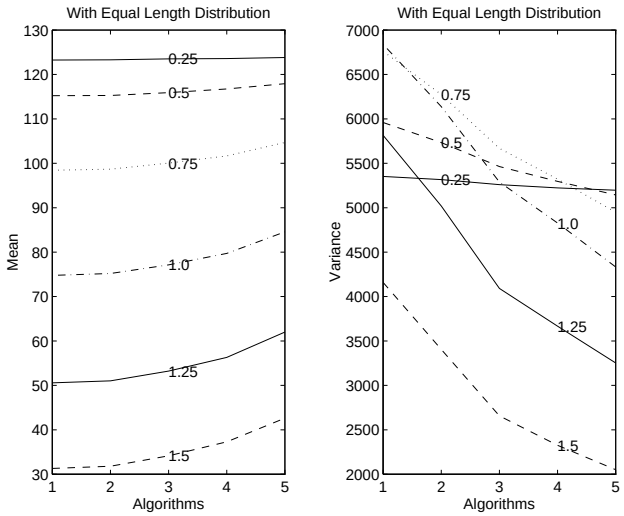

Figure 4: Performance of different algorithms (Equal Length Distribution) (1:Mean Optimal Algorithm, 2:α-algorithm with $\alpha = 2.2$, 3:α-algorithm with $\alpha = 2.6$, 4:α-algorithm with $\alpha = 3$, 5:Variance Optimal Algorithm)

the x-axis, the algorithms we used in the experiment are numbered consecutively from 1 to 5 and they are (1) Mean Optimal Algorithm, (2) α-algorithm with $\alpha = 2.2$, (3) α-algorithm with $\alpha = 2.6$, (4) α-algorithm with $\alpha = 3$, and (5) Variance Optimal Algorithm.

The key observations are as follows. As we expected, the lowest mean response time is achieved when server uses the Mean Optimal Algorithm, and the lowest variance of response time when Variance Optimal Algorithm. The performance of α-algorithms falls between the Mean Optimal Algorithm and Variance Optimal Algorithm. When α changes from 2.2 to 2.6 and then to 3, the measured mean response time is observed to increase gradually while variance is dropping at the same time.

However, the effectiveness of α-algorithm and Variance Optimal Algorithm in reducing variance of response time is challenged when the skew in user demands for items is small. When $\theta = 0.25$, either mean or variance does not show any significant change when different algorithms are adopted. Actually, for $\theta = 0.25$, the mean response time results produced by 5 algorithms are so close with each other that the difference between the maximum value and the minimal value is less than 1 time unit and all are very close to the theoretically minimal value.

When θ increases but is still less than 1.25, the skew in user demands becomes a little large but not too large. For instance, when $\theta = 0.75$ and the number of items is 250, about half user requests are for top 33 items and the remaining 217 items take the burden of serving the other half requests. However, to reduce the overall mean response time, most bandwidth is given to a few items by the Mean Optimal Algorithm. The poor service for the "not most popular" items results in a higher variance of response time. Both α-algorithms and Variance Optimal Algorithm greatly improve the situation. From the curves, the reduction of variance brought by those algorithms is conspicuous, and the unavoidable increase of mean is not very significant.

In general, when θ is increased starting from 0, the variance increases at first, but after a point, it starts to decrease again. For instance, in Figure 4, for the Mean Optimal Algorithm, the variance increases when θ increases from 0.25 to 0.75, but after $\theta = 1.0$, the variance starts decreasing. This phenomenon looks counter-intuitive at first. In fact, when the skew in user demands becomes extremely large, the percentage of requests attracted by a few hot items becomes very large. In our example of 250 items, if $\theta = 1.5$, top 33 items are demanded by 91% user requests. To serve them well means to serve all well. On the other hand, although α-algorithms and Variance Optimal Algorithm reduce the variance of response time drastically as expected, they make a large sacrifice with respect to mean response time. Note the large increase

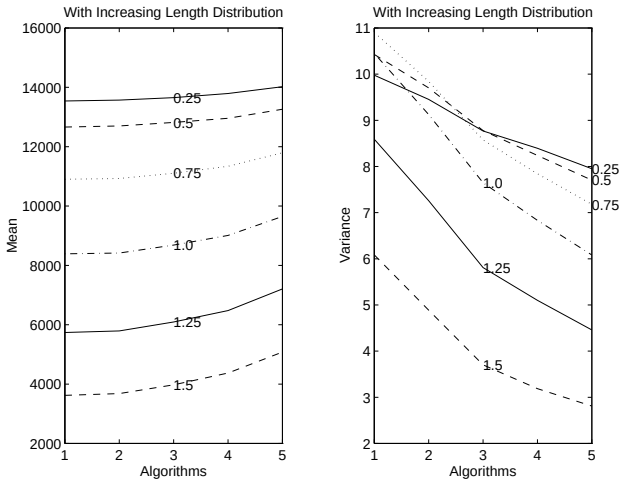

Figure 5: Performance of different algorithms (Increasing Length Distribution) (1:Mean Optimal Algorithm, 2:α-algorithm with $\alpha = 2.2$, 3:α-algorithm with $\alpha = 2.6$, 4:α-algorithm with $\alpha = 3$, 5:Variance Optimal Algorithm)

of mean when $\theta = 1.5$ in Figure 4.

In summary, α-algorithms and Variance Optimal Algorithm perform best in the situation with medium-skewed demand distribution, but when user demands are lightly skewed or severely skewed, Mean Optimal Algorithm is still a good alternative.

5.4 Unequal Length Case

Assuming that the items follow an Increasing Length Distribution and Decreasing Length Distribution respectively, we repeated our experiment. Due to space restriction, only the results for Increasing Length Distribution case are shown here(Figure 5). The results for Decreasing Length Distribution case are similar and can be found in [11].

The curves show that performance of our scheduling algorithms is insensitive to the length distribution of items. Same conclusion as in Equal Length case can be drawn from both Unequal Length cases.

6 Conclusion

In this paper, we argue that from a user's point of view, the variance of response time is also important, not just the mean response time. Variance of response time affects a user's impression about the quality of service a system can provide. In the so-called *pure push-based* data broadcast system, where there is no direct channel for users to send requests explicitly, it is possible for the server to reduce the variance of response time by making appropriate broadcast schedules. In particular, we found a property satisfying which results in a schedule with minimal variance of response time. Based on the property, we proposed a scheduling algorithm that attempts to minimize the variance of response time as well as an algorithm that can trade mean response time with variance of response time.

Simulation was conducted to evaluate the performance of these algorithms. Our α-algorithm performs best when user item demands are medium skewed and effectively implements the trade-off between mean and variance. However, when user demands are lightly or severely skewed, all algorithms present almost same performance and the Mean Optimal Algorithm[13] is still a good alternative.

The evaluation presented in this paper assumed a push-based system. Our algorithms can be easily adapted to achieve low variance in *pull-based* systems[2]. In this case, the number of requests pending for a particular item can be used in place of *demand probability* of the item.

A Appendix I: Mean and Variance of Response Time

In section 2, we define the response time t of any user request as the duration time from when the request is made until the desired item appears on broadcast channel. Based on the assumption that all the users work independently from each other in terms of requesting data items and getting served, we claim that both the generation of requests and the item asked in a request are random events. Two random variables can be defined: T, the issue time of the request, and I, the item required in the request.

I is a discrete random variable taking integer values from 1 to M. The probability of item i being requested, i.e. I taking value i, is just the *demand probability* of item i we defined in section 2, p_i. So

$$Prob[I = i] = p_i$$

Further, if a request for item i comes at time T, its response time t falls in the range $(0, s_i]$ depending on where T resides between two consecutive broadcasts of item i. Since we assume that request arrival

is governed by a Poisson process, the request comes equally likely at any time. In the case of item i being requested, t is uniformly distributed over $(0, s_i]$ and the probability density function of t, $q_i(t)$, is:

$$q_i(t) = \begin{cases} \frac{1}{s_i} & , 0 < t \leq s_i \\ 0 & , \text{ otherwise} \end{cases}$$

Since t is a continuous random variable, cumulative distribution function for t is obtained as:

$$P[t \leq x | I = i] = F_i(x) = \int_{-\infty}^{x} q_i(t)dt, \quad x \quad real$$

where $F_i(x)$ is the cumulative distribution function for t given that $I = i$.

Above is the conditional probability of t. Using the *Multiplication Rule*, we may have the cumulative distribution function $F(x)$ for t.

$$\begin{aligned} P[t \leq x] &= F(x) \\ &= \sum_{i=1}^{M} (Prob[I = i] Prob[t \leq x | I = i]) \\ &= \sum_{i=1}^{M} (p_i F_i(x)) \end{aligned}$$

Let $g(t)$ be the probability density function of random variable t. It follows that,

$$g(t) = \sum_{i=1}^{M} p_i q_i(t) \qquad (10)$$

Now, we will be able to derive the expressions for μ, the mean response time, and σ^2, the variance of response time.

A.1 Expressions for μ

Mean response time μ is the expected value of t. By the definition of expected value, we have

$$\begin{aligned} \mu &= \int_0^{\infty} t g(t) dt \\ &= \int_0^{\infty} (t \sum_{i=1}^{M} (p_i q_i(t))) dt \\ &= \int_0^{\infty} \sum_{i=1}^{M} p_i (t q_i(t)) dt \\ &= \sum_{i=1}^{M} p_i \int_0^{\infty} t q_i(t) dt \\ &= \sum_{i=1}^{M} p_i \int_0^{s_i} \frac{t}{s_i} dt \\ &= \frac{1}{2} \sum_{i=1}^{M} s_i p_i \end{aligned}$$

A.2 Expressions for σ^2

The variance of response time t is the expected value of random variable $(t - \mu)^2$. So, we have

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (t - \mu)^2 g(t) dt \\ &= \int_{-\infty}^{\infty} (t - \mu)^2 \sum_{i=1}^{M} (p_i q_i(t)) dt \\ &= \int_{-\infty}^{\infty} \sum_{i=1}^{M} (p_i (t - \mu)^2 q_i(t)) dt \\ &= \sum_{i=1}^{M} \int_{-\infty}^{\infty} p_i (t - \mu)^2 q_i(t) dt \\ &= \sum_{i=1}^{M} \int_0^{s_i} p_i (t - \mu)^2 \frac{1}{s_i} dt \\ &= \sum_{i=1}^{M} \frac{p_i}{s_i} [\frac{1}{3} (t - \mu)^3 |_0^{s_i}] \\ &= \frac{1}{3} \sum_{i=1}^{M} p_i s_i^2 - (\sum_{i=1}^{M} p_i s_i) \mu + (\sum_{i=1}^{M} p_i) \mu^2 \end{aligned}$$

Since $\mu = \frac{1}{2} \sum_{i=1}^{M} s_i p_i$ and $\sum_{i=1}^{M} p_i = 1$, the above equation can be further simplified as

$$\sigma^2 = \frac{1}{3} \sum_{i=1}^{M} p_i s_i^2 - \mu^2$$

or

$$\sigma^2 = \frac{1}{3} \sum_{i=1}^{M} p_i s_i^2 - (\frac{1}{2} \sum_{i=1}^{M} s_i p_i)^2$$

B Appendix II: Minimizing the Variance

Theorem 1 *Given the demand probability p_i of each item i, the minimal* variance of response time, σ^2, *is*

achieved when the schedule vector possesses the following property, assuming that transmissions of each item i are equally spaced by s_i.

$$\frac{p_i s_i^2}{l_i}(\frac{2}{3}s_i - \mu) = constant, \forall i, 1 \leq i \leq M$$

Proof:

σ^2 is a multi-variable function of $s_1, s_2, \cdots, s_M$. However, only M-1 of the $s_i's$ can be changed independently instead of M. To find this fact, let us define the share of bandwidth that each item occupies. For item i, it is broadcast once every s_i time period and each transmission takes l_i time. So, the percentage of time taken by item i during the broadcast is $\frac{l_i}{s_i}$. To utilize the bandwidth of broadcast channel to its full extent, we should make

$$\frac{l_1}{s_1} + \frac{l_2}{s_2} + \ldots + \frac{l_{M-1}}{s_{M-1}} + \frac{l_M}{s_M} = 1$$

or

$$s_M = l_M(1 - \frac{l_1}{s_1} - \frac{l_2}{s_2} - \ldots - \frac{l_{M-1}}{s_{M-1}})^{-1} \qquad (11)$$

Back to our objective of minimizing the σ^2, we have to find the schedule vector which makes $\frac{\partial \sigma^2}{\partial s_i} = 0, \forall i$. We now solve these equations, beginning with $0 = \frac{\partial \sigma^2}{\partial s_1}$.

$$0 = \frac{\partial \sigma^2}{\partial s_1} = \frac{\partial}{\partial s_1}[\frac{1}{3}\sum_{i=1}^{M} p_i s_i^2 - (\frac{1}{2}\sum_{i=1}^{M} p_i s_i)^2]$$

$$= p_1[\frac{2}{3}s_1 - \frac{1}{2}(\sum_{i=1}^{M} p_i s_i)] + p_M[\frac{2}{3}s_M - \frac{1}{2}(\sum_{i=1}^{M} p_i s_i)]\frac{\partial s_M}{\partial s_1} \qquad (12)$$

From Equation 11, it can be found that

$$\frac{\partial s_M}{\partial s_1} = -\frac{s_M^2}{s_1^2} \cdot \frac{l_1}{l_M}$$

By substitution, Equation 12 becomes

$$0 = p_1[\frac{2}{3}s_1 - \frac{1}{2}(\sum_{i=1}^{M} p_i s_i)] - \frac{p_M s_M^2}{s_1^2} \cdot \frac{l_1}{l_M}[\frac{2}{3}s_M - \frac{1}{2}(\sum_{i=1}^{M} p_i s_i)]$$

which implies that

$$\frac{p_1 s_1^2}{l_1}(\frac{2}{3}s_1 - \frac{1}{2}\sum_{i=1}^{M} p_i s_i) = \frac{p_M s_M^2}{l_M}(\frac{2}{3}s_M - \frac{1}{2}\sum_{i=1}^{M} p_i s_i)$$

Similarly,

$$\frac{p_2 s_2^2}{l_2}(\frac{2}{3}s_2 - \frac{1}{2}\sum_{i=1}^{M} p_i s_i) = \frac{p_M s_M^2}{l_M}(\frac{2}{3}s_M - \frac{1}{2}\sum_{i=1}^{M} p_i s_i)$$

$$\ldots$$

$$\frac{p_{M-1} s_{M-1}^2}{l_{M-1}}(\frac{2}{3}s_{M-1} - \frac{1}{2}\sum_{i=1}^{M} p_i s_i)$$

$$= \frac{p_M s_M^2}{l_M}(\frac{2}{3}s_M - \frac{1}{2}\sum_{i=1}^{M} p_i s_i)$$

In other words,

$$\begin{aligned} \frac{p_1 s_1^2}{l_1}(\frac{2}{3}s_1 - \frac{1}{2}\sum_{i=1}^{M} p_i s_i) &= \frac{p_2 s_2^2}{l_2}(\frac{2}{3}s_2 - \frac{1}{2}\sum_{i=1}^{M} p_i s_i) \\ &= \ldots \\ &= \frac{p_M s_M^2}{l_M}(\frac{2}{3}s_M - \frac{1}{2}\sum_{i=1}^{M} p_i s_i) \end{aligned}$$

This is equivalent to saying that

$$\frac{p_i s_i^2}{l_i}(\frac{2}{3}s_i - \frac{1}{2}\sum_{i=1}^{M} p_i s_i) = constant, \forall i, 1 \leq i \leq M$$

or

$$\frac{p_i s_i^2}{l_i}(\frac{2}{3}s_i - \mu) = constant, \forall i, 1 \leq i \leq M$$

Thus, we have proved Theorem 1.

C Appendix III: Lower Bounds for the α-Algorithm

In the ideal situation, α-algorithm can create a schedule making the equation $\frac{s_i^\alpha p_i}{l_i} = C$ to be true, where C is a constant. In the following, we will derive the value of C, and the values of s_i's when the ideal condition holds. Then, both the mean and variance of response time can be obtained. They serve to be the lower bounds of mean and variance of response time respectively, which can be attained by an α-algorithm.

From the equation $\frac{s_i^\alpha p_i}{l_i} = C, i = 1, 2, \cdots, M$, it follows that

$$s_i = (C \cdot \frac{l_i}{p_i})^{\frac{1}{\alpha}} \qquad (13)$$

Let r_i be the share of bandwidth by item i during broadcast. Since each transmission of item i takes l_i time and item i is transmitted every s_i time period, we have $r_i = \frac{l_i}{s_i}$. As $\sum_{i=1}^{M} r_i = 1$,

$$\sum_{i=1}^{M} \frac{l_i}{s_i} = 1 \quad (14)$$

Substituting the s_i in above equation with the expression in Equation 13, we have

$$\sum_{i=1}^{M} \frac{l_i}{(C \cdot \frac{l_i}{p_i})^{\frac{1}{\alpha}}} = 1$$

or

$$\frac{\sum_{i=1}^{M} (p_i^{\frac{1}{\alpha}} l_i^{1-\frac{1}{\alpha}})}{C^{\frac{1}{\alpha}}} = 1 \quad (15)$$

Solving the equation above, we get

$$C = \Big(\sum_{i=1}^{M} (p_i^{\frac{1}{\alpha}} l_i^{1-\frac{1}{\alpha}}) \Big)^{\alpha} \quad (16)$$

Substituting the value of C into Equation 13, we find the value of s_i for item i as follows,

$$s_i = [\sum_{i=1}^{M} (p_i^{\frac{1}{\alpha}} l_i^{1-\frac{1}{\alpha}})] \ (\frac{l_i}{p_i})^{\frac{1}{\alpha}}, i = 1, 2, \cdots, M$$

Finally, the values of mean μ and variance σ^2 in this case can be derived by substituting the above expression for s_i into Equations 1 and 2.

References

[1] S. Acharya, R. Alonso, M. Franklin, and S. Zdonik, "Broadcast disks - data management for asymmetric communications environment," in *ACM SIGMOD Conference*, May 1995.

[2] S. Acharya, M. Franklin, and S. Zdonik, "Balancing push and pull for data broadcast," in *ACM SIGMOD Conference*, May 1997.

[3] S. Acharya, M. Franklin, and S. Zdonik, "Dissemination-based data delivery using broadcast disks," *IEEE Personal Communication*, pp. 50–60, Dec. 1995.

[4] D. Aksoy and M. Franklin, "Scheduling for large-scale on-demand data broadcasting," in *Proc. of INFOCOM'98*, Apr. 1998.

[5] M. H. Ammar, "Response time in a teletext system: An individual user's perspective," *IEEE Transactions on Communications*, Nov. 1987.

[6] V. Gondhalekar, R. Jain, and J. Werth, "Scheduling on airdisks: Efficient access to personalized information services via periodic wireless data broadcast," in *IEEE Int. Conf. Comm.*, June 1997.

[7] S. Hameed and N. H. Vaidya, "Log-time algorithms for scheduling single and multiple channel data broadcast," in *ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM)*, Sept. 1997.

[8] H.V.Leong and A.Si, "Data broadcasting based on statistical operators," *Electronics Letters*, vol. 32, Oct. 1996.

[9] T. Imielinski, S. Viswanathan, and B. R. Badrinath, "Energy efficient indexing on air," in *International conference on Management of Data*, May 1994.

[10] T. Imielinski, S. Viswanathan, and B. R. Badrinath, "Data on the air - organization and access," *IEEE Transactions of Data and Knowledge Engineering*, July 1996.

[11] S. Jiang and N. H. Vaidya, "Scheduling algorithms for a data broadcast system:minimizing variance of the response time," Tech. Rep. 98-005, Computer Science Department, Texas A&M University, College Station, Feb. 1998.

[12] C.-J. Su and L. Tassiulas, "Broadcast scheduling for information distribution," in *Proc. of INFOCOM'97*, Apr. 1997.

[13] N. H. Vaidya and S. Hameed, "Data broadcast in asymmetric wireless environments," in *Workshop on Satellite Based Information Services (WOSBIS), Rye, NY*, Nov. 1996.

[14] J. W. Wong, "Broadcast delivery," in *Proceedings of IEEE*, pp. 1566–1577, Dec. 1988.

[15] Z. Zdonik, R. Alonso, M. Franklin, and S. Acharya, "Are 'disks in the air' just pie in the sky?," in *IEEE Workshop on Mobile comp. System*, Dec. 1994.