# OPTIMAL OBJECT ALLOCATION FOR MULTIMEDIA BROADCAST

*Edwin A. Heredia*

Thomson Consumer Electronics
Corporate Research (INH 700)
P.O. Box 6139, Indianapolis, IN 46206-6139
Email: *heredia@crmail.indy.tce.com*

## ABSTRACT

With the arrival of terrestrial digital TV, a distribution network able to deliver up to 19 Mbits/s in each of the physical transmission channels will become available. Using the adopted data broadcast protocols, simultaneous transmission of multimedia documents to large population segments can be achieved. While these protocols describe methods for recognizing files in data streams, no method is known yet on how to distribute large collections of files in one or more data streams. This paper addresses this problem. The method proposed in the paper allocates objects in multiple streams according to their sizes and access probabilities, in such a way that average access latency is minimized. We show that the minimization problem can be described as a particular form of the NP hard quadratic allocation model for which an algorithmic solution for finding local minima exists.

## 1. INTRODUCTION

Terrestrial digital TV will offer speeds of up to 19 Mbits/s per physical channel, enough to support not only 3 or 4 SDTV programs but also data broadcasting. Multimedia broadcast services are being considered as the obvious generalization towards which the traditional TV receiver will evolve. In fact, with a capacity to reach billions of people all around the world, multimedia broadcast has the potential of becoming the first truly massive deployment of this type of technologies.

An important part for the development of multimedia transmission services is the standardization of data broadcasting. Both of the major players, DVB in Europe and ATSC in the United States, have agreed to use DSM-CC [7] as the base encapsulation protocol for such services. Files are modularized according to DSM-CC rules, and later packetized following the MPEG-2 Systems protocol. The resulting 188-byte packets are multiplexed with others sharing the same Transport Stream, and after channel encoding, they modulate pilot carriers using 8-VSB (See Fig. 1).

While DSM-CC and the upper-shell data broadcasting protocols indicate how to modularize and identify files within the data streams embedded in the Transport Stream, they do not provide methods to efficiently distribute objects in the streams. If a small number of objects is transmitted per document, then such an organization may not be required. However, multimedia documents such as existing WWW newspapers and magazines are typically composed of thousands of files with different sizes and access requirements. For large documents, intelligent file distribution over multiple streams is necessary to reduce file access delay and save bandwidth.

In this paper we address such a problem. Based on object sizes and their access probabilities, we develop a method to distribute objects in multiple streams in such a way that the average file access time is minimized. We show that the resulting optimization problem can be transformed into a particular form of the quadratic allocation model for which an algorithmic solution has been developed. With bandwidth being very likely the most important parameter in the determination of broadcast costs, methods as the one introduced here are needed for its efficient use in the deployment of future multimedia broadcast services.

## 2. DATA BROADCAST STREAMS

The Transport Stream defined in the ATSC specifications [1] derived from the MPEG-2 Systems protocol [5, 6] is a digital channel supporting a throughput of 19.2 Mbits/s embedded in a 6 MHz band. The channel can itself be partitioned into multiple streams each of which uses a guaranteed portion of the total bandwidth. We define the term *stream* as a collection of packets identified separately by a given packet identifier (PID) and transmitted at approximately constant bit rate (CBR).

Figure 1 illustrates the data multiplexing process that leads to the Transport Stream. Buffers with occupancy feedback are used to control and guarantee the rate of individual streams. Data streams can therefore be transmitted in CBR mode. However, due to video traffic and multiplexing priorities, data packets can be dispersed when inserted in the Transport Stream. Such dispersion causes the actual bit rate to fluctuate but the specifications under development in the ATSC ensure that the fluctuation shall be kept small.

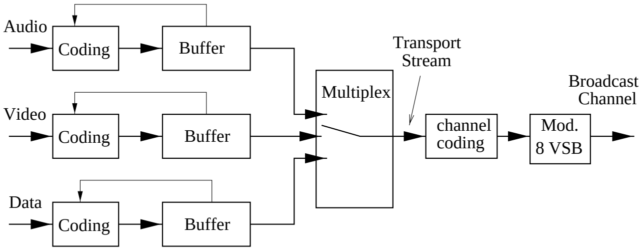

Figure 1: Encoding system architecture for digital TV.

A sequence of files transmitted one after the other and repeated periodically as part of a particular stream is called a data carousel. Periodic transmission allows users to access files randomly. Figure 2 shows a time-bandwidth diagram for a single stream showing the sequence of objects transmitted in the stream. Two carousels groups distinguished by their colors are shown in the figure. The time segment occupied by the object represents the interval required for transmission of all the packets. Objects are transmitted sequentially with no free slots in between. Also in the same figure, the repetition time for object $q_1$ is explicitly indicated. The longer the repetition time is for individual objects, the longer the access delay will be during downloading.

For large documents with hundreds or thousands of files, avoiding long access delays may require the use of multiple streams. For example, high priority files should be placed in streams where small repetition times can be guaranteed, whereas low priority ones could be queued all together in separate streams. This is the problem addressed in this paper. Based on a priority measurement such as the object access probability, we develop an optimal allocation method to place objects in the proper streams in such a way that the average access time per object is minimized.

## 3. OBJECT ACCESS TIMES

A collection of objects (files, pages, or directories) $Q = \{q_1, q_2, \ldots, q_M\}$ is sequentially streamed at a constant bit rate of $b$ bits per second as shown in Fig. 2. The access time for object $k$ is defined as the time required to make the object available to any type of application software. Fig. 2 illustrates that object access times have two components. The first one is the *wait time*, $w$, from the instant when the object is requested to the instant when the object appears in the stream. The second component is the *download time* and represents the time required to recover all the object packets from the stream. If $S_k$ is the size in bits for object $q_k$ and if $b$ is the stream bandwidth, then the download time is $S_k/b$. Consequently, the access time for object $q_k$ is given by

$$t_k = \frac{S_k}{b} + w \quad . \tag{1}$$

If the object request happens to be just before the object appearance in the stream, then the wait time will be null. However, if the request instant is slightly after the first object header packet, then $w$ will be maximum and equal to the time needed for the object to circulate and reappear in the stream. Figure 2 shows three examples of possible wait times when accessing object $q_7$.

Consequently, the wait time $w$ is a random variable uniformly distributed between 0 and $W_{max}$, with

$$W_{max} = \frac{\sum_{m=1}^{M} S_m}{b} \quad . \tag{2}$$

If $E\{\ \}$ denotes the expectation operator, then $E\{w\} = W_{max}/2$, and therefore:

$$E\{t_k\} = \frac{S_k}{b} + \frac{\sum_{m=1}^{M} S_m}{2b} \quad . \tag{3}$$

A WWW server with Internet access can register the number of times each page is accessed over a period of time. Based on this information, an empirical measure of probability can be found for the pages and by extension for the files that compose the pages. Let $p(k)$ be the access probability corresponding to the k-th object of the document object collection $Q$, then the overall average access time is given by:

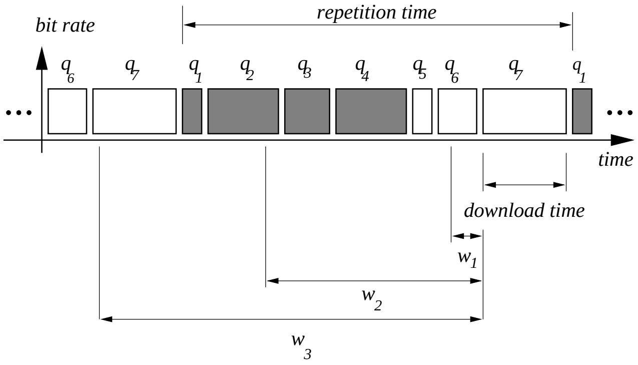

Figure 2: Access time components when retrieving objects from a broadcast stream.

$$t_A = \sum_{k=1}^{M} E\{t_k\}\, p(k) \quad . \tag{4}$$

For the single stream case described in the previous section, $E\{t_k\}$ is defined in Eqn. (3). In the next section, we compute $E\{t_k\}$ for the multi-stream case and use $t_A$ to develop an optimization problem whose solution gives the object allocation for minimal access time.

## 4. DISTRIBUTIONS FOR MULTIPLE STREAMS

For documents with a small object collection, a single stream is enough to carry the entire collection. In fact, multiple documents can be streamed together through the use of data carousel grouping such as provided by DSM-CC. However, for large documents the delivery may require multiple streams to guarantee small access delays. Streams are recognized by their packet identifier (PID) that labels the MPEG-2 Transport Stream packets. Current technology enables stream tuning by PID filtering in transport processing chips. Stream tuning is fast and easy while the overhead comes from the wait and download times described in the previous section.

Audiovisual streams, program and system information (PSIP), and other MPEG-2 functions utilize large number of PIDs. Therefore, while multi-streaming is important for reducing access times, the number of defined streams should simultaneously be kept as small as possible due to hardware limitations.

Once more consider the set of $M$ objects $Q = \{q_1, q_2, \ldots, q_M\}$ which for transmission purpose will be broadcast using multiple streams. Each of the objects has a size $S_k$ and an access probability $p(k)$ for $k = 1, 2, \ldots, M$. Let $N$ represent the number of available streams and let $c_j$ designate the j-th stream with bandwidth $b_j$. An example of the distribution of 11 objects over three stream channels with different bandwidths is illustrated in Fig. 3.

An assignment matrix $\mathbf{X} = [x_{ij}]$ of size $M \times N$ is defined, whose elements indicate whether an object belongs or not to a certain stream, that is

$$x_{ij} = \begin{cases} 1 & \text{if} \quad q_i \in c_j \\ 0 & \text{if} \quad q_i \notin c_j \end{cases} \tag{5}$$

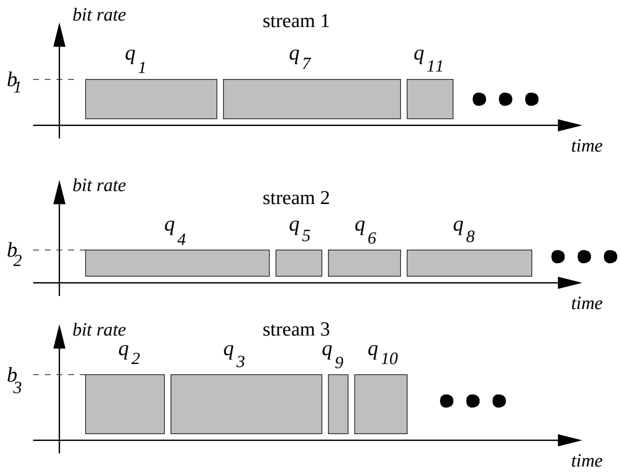

Figure 3: An example of object distribution over multiple streams.

Assuming in principle, that object $q_k$ belongs to the arbitrary streaming channel $c_j$, then equation (3) can be invoked to compute $E\{t_k\}$. This gives

$$E\{t_k\} \;=\; \frac{S_k}{b_j} \;+\; \frac{\sum_{i=1}^{M} S_i\, x_{ij}}{2b_j} \qquad \text{if} \;\; q_k \;\in\; c_j \tag{6}$$

The assignment matrix can be used once more to remove the "if" clause of the previous expression, which gives

$$E\{t_k\} \;=\; \sum_{j=1}^{N} \left[ \frac{S_k}{b_j} \;+\; \frac{\sum_{i=1}^{M} S_i\, x_{ij}}{2b_j} \right] x_{kj} \quad . \tag{7}$$

The assignment matrix $\mathbf{X}$ is precisely the term we would like to find from an optimal point of view.

Substituting equation (7) into (4) and through some algebraic manipulation, it is possible to demonstrate that the overall average access time for a multi-stream object allocation $t_A(\mathbf{X})$ can be written as

$$t_A(\mathbf{X}) \;=\; \sum_{k=1}^{M} \sum_{j=1}^{N} \alpha_{kj}\, x_{kj} \;+\; \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{M} \beta_{ijk}\, x_{kj}\, x_{ij} \tag{8}$$

where the equation constants are defined as

$$\alpha_{kj} \;=\; \frac{p(k)\, S_k}{b_j} \quad , \qquad \beta_{ijk} \;=\; \frac{p(k)\, S_i}{b_j} \quad . \tag{9}$$

Therefore, the optimization problem can be stated as follows. *Find the assignment matrix $\mathbf{X} = [x_{ij}]_{M \times N}$ such that the overall average access time $t_A(\mathbf{X})$ is minimized, subject to the following constraints:*

- constraint 1: $x_{ij} \;\in\; \{0, 1\}$, for all values of $i, j$.
- constraint 2: $\sum_{j=1}^{N} x_{ij} \;=\; 1$

The first constraint implies that the solution space is binary, whereas the second constraint reveals that an arbitrary object can be assigned to one and only one stream. Because of the quadratic form of the cost functional $t_A(\mathbf{X})$ and since the solution space is binary, the problem can be classified as non-linear integer programming of the zero-one kind.

In conventional data broadcasting applications, the stream bandwidths are negotiated *a priori*, and once the broadcast server admits such bandwidths in guaranteed mode, the rates are maintained at the defined levels. When all the selected streams have the same bandwidth, the optimization problem can be further simplified. In this case, after defining the terms

$$C \;=\; \frac{1}{b} \sum_{k=1}^{M} S_k\, p(k) \quad , \qquad a_{ik} \;=\; S_i\, p(k) \tag{10}$$

the overall average access time becomes

$$t_A(\mathbf{X}) \;=\; C \;+\; \frac{1}{2b} \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{M} a_{ik}\, x_{kj}\, x_{ij} \tag{11}$$

From their definitions, it is clear that $C$ and $b$ are positive numbers, consequently, for the equal bandwidth case, a simplified cost function results:

$$J(\mathbf{X}) \;=\; \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k=1}^{M} a_{ik}\, x_{kj}\, x_{ij} \tag{12}$$

subject to the same constraints as the previous non-uniform bandwidth case, that is: $x_{ij} \in \{0, 1\}$, and $\sum_{j=1}^{N} x_{ij} \;=\; 1$.

## 5. OPTIMIZATION ALGORITHM

The optimization model defined by equation (12) is similar to a particular form of the generalized quadratic assignment problem. This form is normally obtained when studying classroom scheduling problems (CSP) [8] or the allocation of interacting activities to facilities [2, 3, 4]. As with most of the known quadratic assignment problems, the existence of nonlinear interaction terms makes these, otherwise simple problems, NP hard.

Carlson and Nemhauser have proposed an optimization algorithm for the CSP problem applicable when the coefficient matrix is symmetric with null diagonal [2]. Under these conditions, local minima can be found through a recursive process. However, for the stream allocation case, it is evident from Eqn. (10) that $\mathbf{A}$ is not symmetric and has, in general, a diagonal which is not null. We show next that a reformulation of the problem is possible to meet the constraints imposed by Carlson and Nemhauser.

Let $y_{ik} \;=\; \sum_{j=1}^{N} x_{ij}\, x_{kj}$. By inspection, it is evident that the matrix $[y_{ik}]$ is: (1) symmetric, (2) has ones as its diagonal elements, and (3) has only ones or zeros as elements. Equation (12) can be written in terms of $y_{ik}$ as $J(\mathbf{X}) \;=\; \sum_{i=1}^{M} \sum_{k=1}^{M} a_{ik}\, y_{ik}$. Then, due to the aforesaid properties of $y_{ik}$, this expression is equivalent to

$$J(\mathbf{X}) \;=\; \sum_{i=1}^{M} S_i\, p(i) \;+\; \sum_{i=1}^{M} \sum_{k=1}^{M} \sum_{j=1}^{N} \tilde{a}_{ik}\, y_{ik} \tag{13}$$

where

$$\tilde{a}_{ik} \;=\; \begin{cases} (a_{ik} + a_{ki})/2 & k \neq i \\ 0 & k = i \end{cases} \tag{14}$$

Having the first summation of Eqn. (13) as a constant, the optimization problem can be restated as:

$$\min \quad z = \sum_{i=1}^{M} \sum_{k=1}^{M} \sum_{j=1}^{N} \tilde{a}_{ik} x_{ij}\, x_{kj} \tag{15}$$

subject to the same constraints as before. The matrix $\tilde{\mathbf{A}} = [\tilde{a}_{ik}]$ is symmetric with null diagonal, and therefore satisfies the conditions required for using the Carlson-Nemhauser algorithm. This algorithm was implemented as described in [2] with the only difference being the use of two starting feasible solutions. The algorithm calculated two optimal allocation matrices and selected the best of both.

## 6. APPLICATION EXAMPLE

Consider the problem of finding the proper number of streams to use when broadcasting a large WWW document. For this example, the document is composed of 150 pages with sizes as shown in Fig. 4. The sizes were obtained using a uniform random number generator. The total document size is about 3.8 MB. From the total of 150 pages, 20 are considered highly likely to be accessed while the remaining 130 have low access probabilities (see Fig. 4) Different probability and size distributions have no impact on the method, since the optimization is carried out without assumptions regarding these distributions.

Assuming an available bandwidth of 250 Kbits/s, we want to determine an efficient way to use the bandwidth for over-the-air broadcast. One option is to allocate the pages uniformly (page 1 placed in stream 1, page 2 to stream 2, and so on). The second option uses the optimization procedure described in the paper. Figure 5 shows the results when the process is repeated for cases with number of streams ranging from 1 to 10, comparing the average access time for each of the cases.

When one stream is used the available bandwidth is entirely dedicated to that stream. When $N$ streams are used, each receives its corresponding fraction of bandwidth. It is evident in this case that blind multi-streaming (that is, without using optimization) produces no beneffits and should be avoided. When using optimization, the situation changes drastically. Multi-streaming helps reducing the average access time from 62 (one stream) to 48 seconds (three streams). Without optimization, such a reduction can only be accomplished by increasing the bandwidth from 250 Kbits/s to 320 Kbits/s. Since bandwith will be a costly commodity in multimedia broadcast services, optimization methods like the one presented in the paper are required to maximize efficiency.

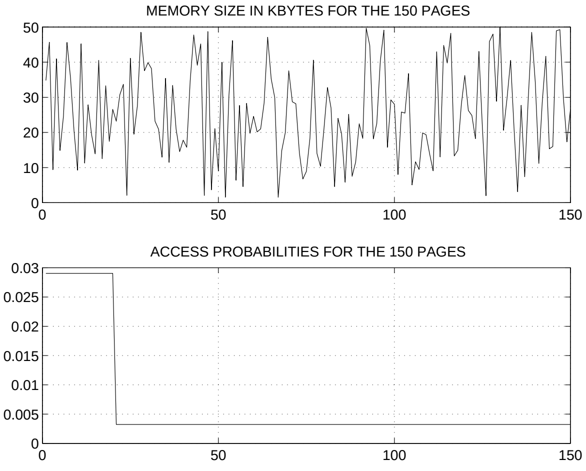

Figure 4: Top: page sizes for each of the 150 pages. Bottom: access probabilities assumed for the multimedia document pages.

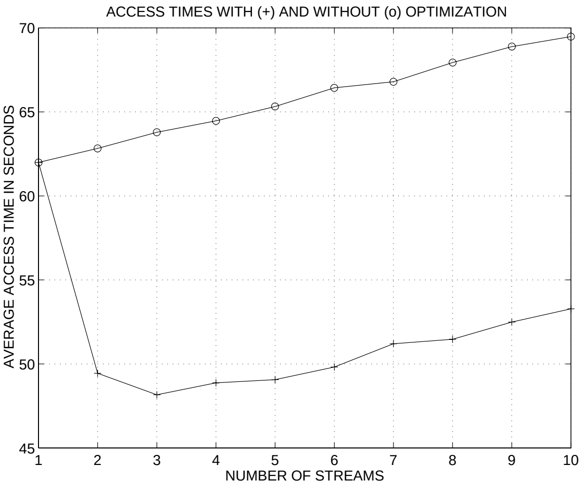

Figure 5: Access times as a function of number of streams for uniform (o) and optimal (+) allocation

## 7. CONCLUSIONS

The problem of distributing multimedia objects among multiple broadcast streams is studied in this paper. Based on a priority measurement such as the object access probability, a nonlinear integer optimization method for stream allocation is proposed. The method minimizes the average delay time when objects are accessed and downloaded by users. For a typical multimedia document composed of 150 pages with sizes ranging from 1KB to 50 KB, we show that the optimization method reduces the access delay from 62 (one stream) to 48 seconds (three streams). A reduction that otherwise can only be achieved by increasing the transmission bandwidth from 250 to 320 Kbits/s. With bandwidth becoming a costly resource, optimization methods like the one introduced in the paper become more and more necessary.

## 8. REFERENCES

[1] ATSC, *Digital Television Standard (A/53)*, 1995.

[2] R. C. Carlson, and G. L. Nemhauser, ”Scheduling to Minimize Interaction Cost”, *Operations Research*, Vol. 14, No. 1, pages 52-58, Jan - Feb 1966.

[3] H. Greenberg, ”A Quadratic Assignment Problem Without Column Constraints”, *Naval Research Logistics Quarterly*, Vol. 16, No 3, pages 417-421, Sept 1969.

[4] H. Greenberg, *Integer Programming*, Academic Press, 1971.

[5] B. G. Haskell, A. Puri, and A. N. Netravali, Digital Video: An Introduction to MPEG-2, Chapman and Hall, 1997.

[6] ISO, *13818-1 Coding of Moving Pictures and Associated Audio - Part 1: Systems*, 1996.

[7] ISO, *13818-6 Coding of Moving Pictures and Associated Audio - Part 6 : Extensions for DSM-CC*, 1996.

[8] M. Padberg, and M. Rijal, *Location, Scheduling, Design and Integer Programming*, Kluwer Academic, 1996.