# Efficient Algorithms for Multi-File Caching

Ekow J. Otoo[1], Doron Rotem[1], and Sridhar Seshadri[2]

[1] Lawrence Berkeley National Laboratory, 1 Cyclotron Road,
University of California, Berkeley, California 94720
[2] Leonard N. Stern School of Business, New York University,
44 W. 4th St., 7-60 New York, 10012-1126

**Abstract.** Multi-File Caching issues arise in applications where a set of jobs are processed and each job requests one or more input files. A given job can only be started if all its input files are preloaded into a disk cache. Examples of applications where Multi-File caching may be required are scientific data mining, bit-sliced indexes, and analysis of sets of vertically partitioned files. The difference between this type of caching and traditional file caching systems is that in this environment, caching and replacement decisions are made based on "combinations of files (file bundles)," rather than single files. In this work we propose new algorithms for Multi-File caching and analyze their performance. Extensive simulations are presented to establish the effectiveness of the Multi-File caching algorithm in terms of job response time and job queue length.

## 1 Introduction

### 1.1 Overview

Caching techniques are widely used to improve the performance of computing systems whenever computations require frequent transfers of data between storage hierarchies that have different access speeds and/or network bandwidth characteristics. Given a sequence of requests for files from some slow or remote storage media, one may use a cache of a small relative size on a faster storage media that holds the most frequently used files. Retrieval to the slow or distant memory is needed only in the case that the file requested is not found in the cache. This results in improved efficiency and reduced costs even with relatively small cache sizes. When a requested file is not found in the cache the system incurs a "fault". The costs associated with a "fault" consists of costs of resources needed to read the file from slow memory and then transferring the file across the network. Efficient caching algorithms choose which current files in the cache must be replaced with new ones in order to maintain a set of files in the cache that minimizes the expected cost of "faults".

There are many papers [1, 3, 6, 8, 9, 11] describing and analyzing caching and replacement policies. These works distinguish between online and off-line algorithms. In both cases, a sequence of requests for files arrive at a queue and must be served on a First Come First Served (FCFS) basis. A replacement decision

must be made whenever a "fault" occurs. Online algorithms make a replacement decision based only on the current request and previous requests but do not have any information about future requests. On the other hand, off-line algorithms make replacement decisions based on the complete sequence of both past and future requests. Off-line algorithms are not practical and are mainly used for establishing bounds and gaining insights on the performance of online algorithms.

In addition, caching algorithms can be classified based on their sensitivity to file sizes and "fault" costs. The following cases are considered in the literature:

**Paging:** Sizes of all files and their "fault" costs are equal
**Fault Model:** File sizes are arbitrary while "fault" costs are the same for all files
**Weighted caching:** All files sizes are the same but "fault" costs may be arbitrary
**Bit Model:** Files may have arbitrary sizes, "fault" costs are proportional to file size
**General Model:** Both "fault" costs and file sizes may be arbitrary

This work is motivated by file caching problems arising in scientific and other data management applications that involve multi-dimensional data [6, 7, 10]. The caching environment for such applications is different than the works described above in two main aspects:

**Number of files associated with a request:** As explained below due to the nature of the applications a request may need multiple files simultaneously. A request cannot be serviced until all the files it needs are in the cache.
**Order of request service:** In case several requests are waiting in the queue, they may be served in any order and not necessarily in First Come First Serve order (FCFS). Policies that determine the order in which requests are served (admission policies), become important and sometimes must be considered in combination with cache replacement policies [6].

### 1.2 Motivating examples of applications

Scientific applications typically deal with objects that have multiple attributes and often use vertical partitioning to store attribute values in separate files. For example, a simulation program in climate modeling may produce multiple time steps where each time step may have many attributes such as temperature, humidity, three components of wind velocity etc. For each attribute, its values across all time steps are stored in a separate file. Subsequent analysis and visualization of this data requires matching, merging and correlating of attribute values from multiple files. Another example of simultaneous retrieval of multiple files comes from the area of bit-sliced indices for querying high dimensional data [10]. In this case, a collection of $N$ objects (such as physics events) each having multiple attributes is represented using bit maps in the following way. The range of values of each attribute is divided into sub-ranges. A bitmap is constructed for

each sub-range with a $'0'$ or $'1'$ bit indicating whether an attribute value is in the required sub-range. The bitmaps (each consisting of $N$ bits before compression) are stored in multiple files, one file for each sub-range of an attribute. Range queries are then answered by performing Boolean operations among these files. Again, in this case all files containing bit slices relevant to the query must be read simultaneously to answer the query.

### 1.3 Problem description

Our approach for caching multiple files consists of two steps that are applied at each decision point of the algorithm. Given a cache of some fixed size and a collection of requests currently waiting in the admission queue for service:

**Step-1, File removal:** We first remove from the cache a set of "irrelevant" files. Files become "irrelevant" if they are not used by any current request and fall below some threshold in terms of their "desirability" according to some known eviction policy such as "least recently used" or "greedy-dual".

**Step-2, Admission of requests:** After the removal step, we load files into the available space in the cache such that the number of requests in the admission queue that can be serviced is maximized. In the weighted version of the problem, each request may have some value reflecting its popularity or priority and the goal in that case is to maximize the total value of serviced requests.

From these two steps, the more interesting for us is Step-2 since Step-1 can be performed based on any known efficient algorithm for file replacement in the cache. The problem presented by Step-2 is called the *Multi-File Caching (MFC)* problem and is described more precisely in Section 2. We will next illustrate the problem with a small example.

### 1.4 Example

As a small example of the non-weighted version of this problem, consider the bipartite graph shown in Fig. 1. The top set of nodes represents requests and the bottom set of nodes represents files. Each request is connected by an edge to each of the files it needs. Assuming all files are of the same size, each request has value 1, and the cache has available space to hold at most 3 files, it can be shown that the optimal solution of value 3 is achieved by loading the files a, c and e into the cache and servicing the three requests 1,3, and 5. Loading a, b and c has a value of 1 as only request 2 can be served whereas loading b, c and e has a value of 2 as it allows us to serve requests 2 and 5.

### 1.5 Main Results

The main results of this paper are:

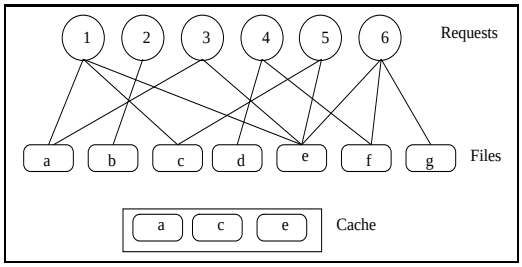


**Fig. 1.** A bipartite graph depiction of a set of jobs and their file requests.

1. Identification of a new caching problem that arises frequently in applications that deal with vertically partitioned files.
2. Derivation of heuristic algorithms that are simple to implement and take into account the dependencies among the files.
3. Analysis of the heuristics and derivation of tight bounds from the optimal solution
4. Extensive simulation results comparing the new algorithm with the traditional first come first serve.

The rest of the paper is organized as follows. In Section 2 we formally describe the MFC problem and discuss its complexity. In Section 3 a heuristic greedy algorithm, called *Greedy Request Value (GRV)* is proposed and its bounds from the optimal solution are shown using Linear Programming (or LP) relaxation. In Section 4 a variation on the *GRV* algorithm is proposed with improved bounds. In Section 5, we present a simulation framework for evaluating the performance of the proposed *GRV* algorithm. Results of the simulation studies, i.e., workload characterization and measurements of performance metrics are presented in Section 6. Conclusions and future work are presented in Section 7.

## 2 Related Problems and Approximation Complexity

The Multi-File Caching (MFC) problem is defined as follows: Given a collection of requests $R = \{r_1, r_2, \ldots, r_n\}$, each with associated value $v(r_i)$, defined over a set of files $F = \{F_1, F_2, \ldots, F_m\}$, each with size $s(F_i)$ and a constant $M$, a bound on the maximum total values, find a subset $R'$ of the requests, $R' \subseteq R$, of maximum total value such that the total size of the files needed by $R'$ is at most $M$. It is easy to show that in the special case that each file is needed by exactly one request the MFC problem is equivalent to the knapsack problem. The MFC problem is NP-hard even if each request has exactly 2 files. This is done by reduction from the Dense $k-$subgraph (DKS) problem [2]. An instance of the DKS problem is defined as follows: Given a graph $G = (V, E)$ and a positive integer $k$, find a subset $V' \subseteq V$ with $|V'| = k$ that maximizes the total

number of edges in the subgraph induced by $V'$. Given an instance of a DKS problem, the reduction to an instance of MFC is done by making each vertex $v \in V$ correspond to a file $f(v)$ of size 1. Each edge $(x, y)$ in $E$ corresponds to a request for two files $f(x)$ and $f(y)$. A solution to the MFC instance with a cache of size $k$ corresponds to a solution to the instance of the DKS where the $k$ files loaded into the cache correspond to vertices of the subgraph $V'$ in the solution of the DKS instance. We also note that any approximation algorithm for the MFC problem can be used to approximate a DKS problem with the same bound from optimality. Currently the best-known approximation for the DKS problem [2] is within a factor of $O(|V|^{1/3-\epsilon})$ from optimum for some $\epsilon > 0$. It is also conjectured in [2] that an approximation to DKS with a factor of $(1+\epsilon)$ is NP-hard. It is also interesting to note that in case each request can start its service when at least one of its files is in the cache (but not necessarily all), the problem becomes equivalent to the Budgeted Maximum Coverage Problem (BMC ) [4, 5]. Using the above terminology, in the BMC problem we are given a budget $L$ (cache size) and the goal is to select a subset of the files $F' \subseteq F$ whose total size is at most $L$ such that the total value of the requests using files from $F'$ is maximized. It turns out that BMC is easier to approximate. In [4], an approximation algorithm is given for the BMC with a factor of $(1 - 1/e)$ from optimal.

## 3 A Greedy Algorithm and Bounds from Optimality

Next, we will describe a simple greedy algorithm called Algorithm GRV (Greedy Request Value) and later prove the relationship between the request value produced by this algorithm and the optimal one. First we need some definitions. For a file $f_i$, let $s(f_i)$ denote its size and let $d(f_i)$ represent the number of requests served by it. The adjusted size of a file $f_i$, denoted by $s'(f_i)$, is defined as its size divided by the number of requests it serves, i.e., $s'(f_i) = s(f_i)/d(f_i)$. For a request $r_i$, let $v(r_i)$ denote its value and $F(r_i)$ represent the set of files requested by it. The adjusted relative value of a request, or simply its relative value, $v'(r_j)$, is its value divided by the sum of adjusted sizes of the files it requests, i.e.

$$v'(r_j) = \frac{v(r_j)}{\sum_{f_i \in F(r_j)} s'(f_i)}$$

Algorithm GRV below attempts to service requests in decreasing order of their adjusted relative values. It skips requests that cannot be serviced due to insufficient space in the cache for their associated files. The final solution is the maximum between the value of requests loaded and the maximum value of any single request.

### 3.1 Linear Programming Relaxation

We now proceed to analyze the quality of the solution produced by this algorithm. The MFC problem can be modeled as a mixed-integer program as follows.

**input** : A set of n requests $R = \{r_1, \ldots r_n\}$, their values $v(r_j$, a set of $n$ files $F$, the sets $F(r_i)$, a cache $C$ of size $s(C)$ and the sizes $s(f_i)$ of all files in F.

**output** : The *solution* - a subset of the requests in $R$ whose files must be loaded into the cache.

**Step 0:** */* Initialize */*
*Solution* $\leftarrow \phi$; //set of requests selected
$s(C') \leftarrow \phi$ ; // $s(C')$ keeps track of unused cache size
**Step 1:** Sort the requests in $R$ in decreasing order of their relative values and renumber from $r_1, \ldots, r_n$ based on the this order
**Step 2:**
**for** $i \leftarrow 1$ ***to*** $n$ **do**
  **if** $s(C') \geq s(F(r_i))$ **then**
    Load the files in $F(r_i)$ to the cache
    $s(C') \leftarrow s(C') - s(F(r_i))$ ; // update unused cache size
    *Solution* $\leftarrow$ *Solution* $\cup\, r_i$ ; // add request $r_i$ to the solution
  **end**
**end**
**Step 2:** Compare the total value of requests in *Solution* and the highest value of any single request and choose the maximum

**Algorithm 1:** GRV

Let

$$z_i = \begin{cases} 1 \text{ if the file} f_i \text{ is in cache} \\ 0 \text{ otherwise} \end{cases}$$

and let

$$y_j = \begin{cases} 1 \text{ if all files used by } r_j \text{ are in cache} \\ 0 \text{ otherwise} \end{cases}$$

Then the mixed integer formulation,$\mathcal{P}$, of MFC can be stated as:

$$\mathcal{P}: \quad \max \sum_{j=1}^{n} v(r_j) y_j$$

*subject to*

$$y_j - z_i \leq 0, \forall i \in F(r_j), \text{ and } \forall j$$

$$\sum_{i=1}^{m} s(f_i) z_i \leq s(C), \quad z_i \in \{0, 1\}$$

The linear relaxation of this problem,$\mathcal{P}_\infty$, and its associated dual problem, $\mathcal{D}$, are not only easier to analyze but also provide a useful bound for a heuristic solution procedure.

$$\mathcal{P}_\infty: \quad \max \sum_{j=1}^{n} v(r_j) y_j$$

*subject to*

$$y_j - z_i \leq 0, \forall i \in F(r_j), \text{ and } \forall j$$

$$\sum_{i=1}^{m} s(f_i) z_i \leq s(C), \quad 0 \leq z_i \leq 1.$$

$$\mathcal{D}: \quad \min s(C)\lambda + \sum_{i=1}^{m} \lambda_i$$

*subject to*

$$\sum_{i \in F(r_j)} \lambda_{ji} = v(r_j) \text{ for } j = 1, 2, \ldots, n \tag{1}$$

$$\lambda s(f_i) + \lambda_i - \sum_{j: f_i \in F(r_j)} \lambda_{ji} \geq 0, \text{ for } i = 1, 2, \ldots, m, \quad \lambda, \lambda_i, \lambda_{ji} \geq 0, \tag{2}$$

where $\lambda_{ji}$ are the dual variables corresponding to the first set of primal constraints, $\lambda$ is the dual variable corresponding to the cache size constraint, and the $\lambda_i$'s correspond to the last set of constraints bounding the $z$'s to be less than one.

To avoid trivialities, we assume that for each request $j : \sum_{i \in F(r_j)} s(f_i) \leq s(C)$, that is, each request can be addressed from the cache, otherwise we can eliminate such requests in the problem formulation.

### 3.2 Primal dual bound from optimal

We shall use the linear programming relaxation to bound the solution produced by GRV. This will be done in two steps. First we shall bound the solution to $\mathcal{P}_1$, by bounding the solution to D and producing a feasible solution to the primal that can be compared to this bound. Then we will bound GRV. Consider an approximation algorithm GRV(LP) for solving $\mathcal{P}_1$ that is similar to GRV except that it allows partial loading of files. It comprises of ranking the requests in descending order of the $v'(r_j)$'s and loading them greedily until the cache is full. Assume that if all the files for a request cannot be fully loaded, they are loaded partially until the cache is full. Let's assume that the collection of requests serviced from the cache without loss of generality are denoted as $r_1, r_2, \ldots, r_p$. Now, we exhibit the feasible dual solution. Let

$$\lambda_{ji} = \frac{v(r_j) s(f_i)/d(f_i)}{\sum_{t \in F(r_j)} s(f_t)/d(f_t)}$$

This assignment to the $\lambda_{ji}$'s satisfies the constraints (1). Let

$$\lambda = \frac{v(r_p)}{\sum_{t \in F(r_p)} s(f_t)/d(f_t)} = v'(r_p).$$

Set $\lambda_j$ to 0 for files not used by the p requests as well as the files used to address only the $p^{th}$ request. Then for this assignment of dual variable values, the left hand side of (2) evaluates to

$$\begin{aligned}
\lambda s(f_i) + \lambda_i - & \sum_{j:f_i \in F(r_j)} \lambda_{ji} \\
&= s(f_i) \frac{v(r_p)}{\sum_{t \in F(r_p)} s(f_t)/d(f_t)} - s(f_i) \sum_{j:f_i \in F(r_j)} \frac{v(r_j)/d(f_i)}{\sum_{t \in F(r_j)} s(f_t)/d(f_t)} \\
&\geq s(f_i) \left( v'(r_p) - \max_{j \geq p} v'(r_j) \right) \geq 0,
\end{aligned}$$

Thus equation 2 is satisfied for such files. Finally, for files used to address the $p-1$ requests, let

$$\lambda_i = \max_{j<p, i \in F(r_j)} \left\{ s(f_i) \left( v'(r_j) - v'(r_p) \right) \right\}.$$

A similar substitution as above reveals that

$$\begin{aligned}
\lambda s(f_i) + \lambda_i - \sum_{j:f_i \in F(r_j)} \lambda_{ji} &= s(f_i) \frac{v(r_p)}{\sum_{t \in F(r_p)} s(f_t)/d(f_t)} \\
&+ \max_{j<p, i \in F(r_j)} \left\{ s(f_i) \left( v'(r_j) - v'(r_p) \right) \right\} - s(f_i) \sum_{j:f_i \in F(r_j)} \frac{v(r_j)/d(f_i)}{\sum_{t \in F(r_j)} s(f_t)/d(f_t)} \\
&\geq s(f_i) \left( v'(r_p) - \max_{j<p} v'(r_j) \right) + \max_{j<p, i \in F(r_j)} \left\{ s(f_i) \left( v'(r_j) - v'(r_p) \right) \right\} \\
&= 0
\end{aligned}$$

Finally, the dual objective function value equals

$$\begin{aligned}
s(C) v'(r_p) + & \sum_{i \in \cup_{j<p} F(r_j)} \max_{j<p, i \in F(r_j)} \left\{ s(f_i) \left( v'(r_j) - v'(r_p) \right) \right\} \\
&\leq \left( s(C) - \sum_{i \in \cup_{j<p} F(r_j)} s(f_i) \right) v'(r_p) + \sum_{i \in \cup_{j<p} F(r_j)} \max_{j<p, i \in F(r_j)} \left\{ s(f_i) \right\} \\
&\leq \left( s(C) - \sum_{i \in \cup_{j<p} F(r_j)} s(f_i) \right) v'(r_p) + \sum_{j<p} v'(r_j) \sum_{i \in F(r_j)} s(f_i) \\
&\leq \max_i d(f_i) \left( \sum_{j<p} v(r_j) + \frac{\left( s(C) - \sum_{i \in \cup_{j<p} F(r_j)} s(f_i) \right)}{\sum_{i \in F(r_j)} s(f_i)} \right). \qquad (3)
\end{aligned}$$

In the second inequality we have used the fact that the maximum of a sum of positive values is less than their sum. The final expression equals the value of the solution produced by the approximation algorithm times the maximum number

of requests that need the same file. However, the objective function value of any feasible solution to the dual is greater than the value of the optimal solution to primal.

**Theorem 1.** *Let $V_{GRV}$ represent the value produced by Algorithm GRV and let $V_{OPT}$ be the optimal value. Let $d^*$ denote the maximum degree of a file, i.e., $d^* = \max_i d(f_i)$ then*

$$\frac{V_{OPt}}{V_{GR}} \leq 2d^*$$

**Proof Outline:** Modify the algorithm $GRV(LP)$ such that it stops with the last request that can only be accommodated partially (or not at all). It then also compares the solution produced to the value of the last request that could not be accommodated and outputs the larger of the two solutions. As one of the two terms within the parentheses on the right hand side of equation 3 is larger than the other the integral solution produced by the modified $GRV(LP)$ is at least $1/2d^*$ times the optimal solution. Algorithm $GRV$ can be adapted to produce equivalent or a better solution then $GRV(LP)\square$

### 3.3 A construction of a case with bound $1.5d^*$

The worst performance ratio of GRV algorithm to the optimal one that we are able to show is 1.5d* as shown below.

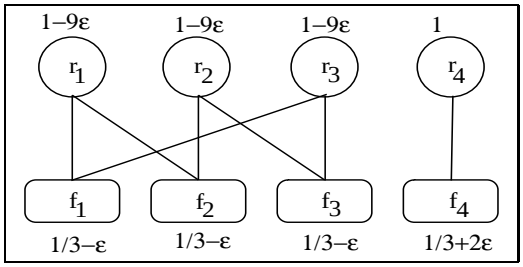


**Fig. 2.** An Example of a case with bound $1.5d^*$

Consider the requests shown in Figure 2. For some small $\epsilon, 0 < \epsilon \leq 1/3$, assume a cache size of $1-3\epsilon$ (measured in some units). The value of each request and the size of each file is shown next to the respective node. The request with the highest relative value is $r_4$ with a value of $3/(1+6\epsilon)$. Each of the other requests has a relative value of $3(1-9\epsilon)/(1-3\epsilon)$ which can be shown to be smaller than that of $r_4$. The $GRV$ (Greedy Relative Value) algorithm will therefore choose to serve $r_4$ and load $f_4$ into the cache using $1/3+2\epsilon$ of the cache capacity. It is then not possible to serve any other requests as each of the remaining requests needs an additional pair of files with space requirements of $2/3-2\epsilon$ exceeding the available cache size of $1-3\epsilon$. The total value of this solution is 1 (the value of

$r_4$). On the other hand, the optimal algorithm can load files $f_1, f_2, f_3$ and serve requests $r_1, r_2$ and $r_3$ for a total value of $3 - 27\epsilon$. Therefore the ratio between the value of the optimal solution to that of the $GRV$ algorithm approaches 3. In this case $d^* = 2$ so the ratio approaches $1.5d^*$.

## 4 An Improved Algorithm

Next, we describe an algorithm called $GRV - k$ that improves on the solution of $GRV$ but comes with a computational cost that is larger by a factor of $O(n^k)$ for a set of $n$ requests. We choose some fixed $k$ integer, and then for each $j$ such that $j \leq k$, we select an initial set of $j$ requests, these requests and their associated files (assuming feasibility) are then removed from the problem and the GRV algorithm is applied to the remaining subproblem consisting of $n - j$ requests and with a cache size reduced by the total size of files needed by the selected $j$ requests. The solution consists of the $j$ initial requests and the requests subsequently selected by the GRV algorithm applied to the remaining subproblem. The output of the $GRV - k$ algorithm is the best solution over all possible choices of $j$ initial requests. (Note that the GRV algorithm can also be called $GRV - 0$ using this notation).

### 4.1 Bounds from Optimality

We now show that the bound on $GRV - k$ is $d^*$ for $k > 1$. First of all, notice that in the GRV algorithm instead of addressing the last set partially (as in $GRV(LP)$), we skip that set and select the next set. Notice that if we discard a set that is not in the optimal solution then no harm is done. In other words, we could restate the original problem without this set and still obtain the optimal solution. If we never have to discard a set from the optimal solution then all sets in the optimal solution are in the solution produced by the approximation algorithm. Thus, it is sufficient to consider the case when the approximation algorithm discards (due to cache restrictions) for the first time a set that is in the optimal solution.

With these observations, we develop the bound using a technique explained in [4]. Clearly, if the optimal solution ($OPT$) addresses only a subset of requests that has cardinality less than or equal to $k$, we have found the optimal solution. Thus, assume that the subset of requests in OPT has a cardinality greater than $k$. As before, the value of the optimal solution is denoted by $V_{OPT}$. Order the requests in $OPT$ in descending order of their value, i.e, $v(r_j)$. Without loss of generality assume that these are the requests, $\{1, 2, \ldots, l\}$, where $l > k$. Consider the solution produced by the modified algorithm when it is started with the requests, $1, 2, \ldots, k$. Let the requests added fully by the approximation be $j = k+1, k+2, \ldots, p-1$. Let the first time a request from $OPT$ can not be added to the solution be when the $t^th$ request in $OPT$ is considered. Let the approximate solution produced comprising only of requests that could be fully addressed from the cache be $V_{approx}$. By the bound developed so far it is clear that

$$V_{approx} \geq \frac{\left(V_{OPT} - \sum_{j=1}^{k} v(r_j)\right)}{d^*} - v(r_t)$$

Thus,

$$\sum_{j=1}^{k} v(r_j) + V_{approx} \geq \frac{V_{OPT}}{d^*} + \frac{d^* - 1}{d^*} \sum_{j=1}^{k} v(r_j) - v(r_t) \tag{4}$$

**Theorem 2.** *If $d^* > 1$ as well as $k = 2$, then, $v_{GRV-k}$, the value produced by the $GRV - k$ algorithm satisfies $V_{OPT}/V_{GRV-k} \leq d^*$.*

**Proof:**

Observe that $v(r_t)$ is not larger than $v(r_j)$ for $j$ less than or equal to $k$. Thus, equation (4) provides the result. *square*

In fact we can construct an example showing that the $GRV - k$ algorithm may produce a solution which is a factor of 2 from optimal for a system in which $d^* = 2$ showing that this bound is tight.

## 5 The Simulation Framework for Multi-File Caching

### 5.1 The Model of the Disk Cache

To evaluate various alternative algorithms for scheduling jobs in a Multi-File caching environment, we setup an appropriate machinery for file caching and cache replacement policies. More specifically, the machinery compares GRV and FCFS job admissions in combination with the least recently used (LRU) replacement policy. Although cache replacement policies have been studied extensively in the literature these have only dealt with transfers between computing system's memory hierarchy, database buffer management and in web-caching where the model for cache replacement assumes instantaneous replacements. That is that the request to cache an object is always serviced immediately and once the object is cached, the service on the object is carried out instantaneous. As a result the literature gives us very simplistic simulation models for the comparative studies of cache replacement policies. Such models are inappropriate in the practical scenarios for MFC. For instance, once a job is selected for service, all its files must be read into the cache and this involves very long delays

We develop and implement an appropriate simulation model that takes into account the inherent delays in locating the file, transferring the file into the cache and holding the file in the cache while it is processed. The sizes of the files we deal with impose these long delays. We capture these in the general setup of our simulation machinery. The machinery considers the files to exist in various states and undergo state transitions, from state to state, conditionally when they are subjected to certain events.

### 5.2 The States of a File in a Disk Cache

Each file object associated with the cache is assumed to be in some state. The file may not have been referenced at all in which case it is assumed to be in state $S_0$. When a reference is made to the file (an event which we characterize subsequently), the file makes a transition to state $S_1$. A file in state $S_1$ implies that it as been referenced with at least one pending task waiting to use it but has neither been cached nor in the process of being cached.

A file is in state $S_2$ if it has been previously referenced but not cached and there are no pending tasks for the file. A file is in state $S_3$ if it has been cached but not pinned and in state $S_4$ if space reservation has been made for the file, and is in the process of being cached. A file is in state $S_5$ if it is cached and pinned. Each of these states is characterized by a set of conditions given by the file status, number of waiting tasks, the last time the file was cached, the job identifier that initiated the caching of the file, and the setting of a cache index.

At some intermittent stages, all files in state $S_2$ that have not be used for a specified time period are flushed from memory. At this stage all accumulated information, such as the number of reference counts accumulated since its first reference, is lost. The file is said to be set back into state $S_0$. For our simulation runs all files in state $S_2$ that have not been referenced in the last five days are cleared.

Any subsequent reference to the file would initiate a new accumulation of historical information on the references made to the file. The various states of a file is summarized as follows:

$S_0$**:** Not in memory and not-referenced.
$S_1$**:** Referenced, not cached but has pending tasks.
$S_2$**:** Referenced, not cached and has no pending tasks.
$S_3$**:** Cached but not pinned.
$S_4$**:** Space reserved but not Cached. Caching in progress.
$S_5$**:** Cached and pinned.

### 5.3 The Event Activities of the Disk Cache

A file that is referenced, cached, processed and eventually evicted from the cache is considered to undergo some state changes. The possible events that force the files to undergo state changes are now described. The events affecting the state changes of the files are caused by the actions of the tasks that are invoked by the jobs and related system actions. Figure 3 illustrates some of the details of the simulation framework used in processing jobs. Jobs that arrive at a host are maintained in the search structure $T_1$. The search structure $T_1$ is either a simple queue or a balanced search tree depending on the algorithm for processing the jobs. For FCFS job scheduling, $T_1$ is represented as queue and for the GRV algorithm, $T_1$ is represented as binary search tree. A significant addition to the nodes of $T_1$ is that each job retains a list of file requests to be processed. The files requested by the jobs are maintained in a search structure $T_2$, and each

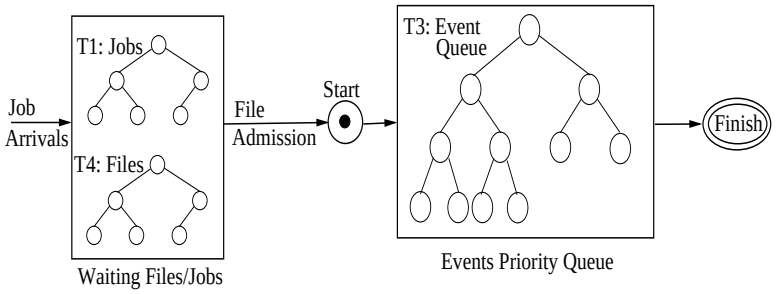


**Fig. 3.** A Simulation Framework for Evaluating the Performance of MFC Algorithms

node of $T_2$ corresponds to a unique file and also maintains a list of all the jobs that have requested it.

A job scheduling policy is used to select the job whose files are to be cached next. For FCFS, the job in front of the queue $T_1$ is selected next. In the GRV algorithm, we evaluate the selection criterion for all waiting jobs and select the recommended one based on the potential available cache space. When a job is selected, all its files are then scheduled to be brought into the disk cache. For each such file of a job a task event is initiated by creating a task token that is inserted into the event queue $T_3$. Each task token is uniquely identified by the pair of values of the job and file identifiers. A task token is subjected to five distinct events at different times. These events are: *Admit-File ($E_0$) Start-Caching ($E_1$)*, *End-Caching ($E_2$)*, *Start-Processing ($E_3$)* and *End-Processing ($E_4$)*. Two other events are the *Cache-Eviction ($E_5$)* and the *Clear-Aged-file ($E_6$)*. The special event, *Clear-Aged-file ($E_5$),* when it occurs, causes the all the information (e.g., history of references to a file) for files that have been dormant for a stipulated period to be deleted. The entire activities within this framework are executed as a discrete event simulation. The activities of the simulation may be summarized by the finite state machine, with conditional transitions. This is depicted as a state transition diagram of Figure 4. The simulation is event driven and the three different file processing orders are modeled accordingly with the order of insertions and deletions of files in the data structure $T_4$.

### 5.4 File Processing in MFC with Delays

The event selected next is an admission if the arrival time of the next job is earlier than the time of the earliest event in $T_3$. If a job arrival is earliest, it is inserted into the admission structure $T_1$. The corresponding files being requested are then inserted into the structure $T_2$. Note that the job also maintains its list of files as well. On the other hand if the top event of the event queue $T_3$ is earlier, it is removed and processed accordingly.

The processing of the event may reinsert an event token into the event queue unless the event is the completion of a task. Each time a job completes, we determine the potential available free cache space and schedule the next job to

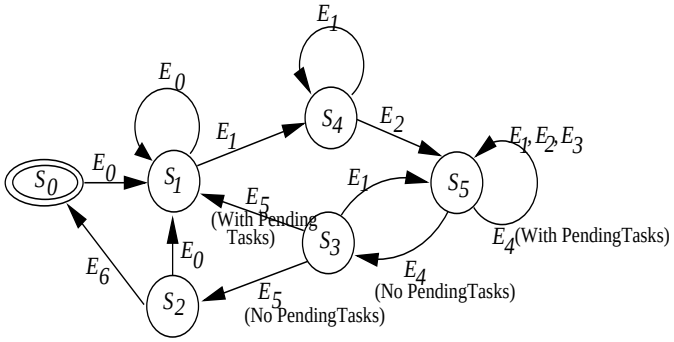


**Fig. 4.** A Finite State Machine Diagram with Conditional Transitions

be processed. In GRV, the potential available cache space is used in the algorithm and not the actual free space. The *potential available free cache space* is the sum of unoccupied space and total size of all unpinned files.

We evaluate the jobs admission policy and the cache replacement policy separately but not independently. In evaluating GRV, a file in cache is assigned a zero file size and in evaluating a replacement policy, all files that appear in the batch of files to be cached are first pinned there by restricting them from being candidates for eviction.

## 6 Experimental Setup

We conducted experiments using the simulation framework described above to compare the GRV algorithm with a naive FCFS job scheduling of the Multi-file Caching problem when the cache replacement algorithm is LRU. Two performance metrics were used: the average response time of a job and the average queue length jobs with workloads of varying jobs arrival rates. Our implementation is a straight forward translation of the *Finite State Machine (FSM)*, with conditional transitions, to a C++ code. When all the files of a job are cached the tasks associated with the jobs, process the files at a steady rate of 10 MBytes per second. This implies that the processing time of a job is the time to process the file with the largest size.

### 6.1 Workload Characteristics and Simulation Runs

We subjected the simulation model to workloads in our experiments where the job inter-arrival times are exponentially distributed with mean inter-arrival times are 40, 60 and 120 seconds. Each job makes a request for $n$ files where n is a uniform number between 1 and 25. The file sizes are also uniformly distributed between 500 Kbyte to 4 Gbytes.

The simulation runs were carried out on a Redhat Linux machine with 512 MBytes of memory. We evaluated the performance metrics of the average response time per job and the average queue length when the cache replacement

policy is LRU. For each configuration and for each workload, a number of runs were done with cache sizes varying from 70 to 130 gigabytes. For each run and for each cache size, we applied a variance reduction method by averaging the statistics that we compute independently for 5 segments of the workload.

### 6.2 Discussion of Results

Figures 5a, 5b and 5c show the graphs of the response times for the synthetic workloads for the respective mean inter-arrival times of 40, 60 and 120 seconds. These graphs indicate that the GRV clearly gives a better response time than a simply FCFS job scheduling. GRV performs even better for higher arrival rates. The higher the arrival rate of jobs GRV used to base its selection on and consequently give even better performance. However as disk cache sizes increase, the performance of the two algorithms converge.

The graphs of the average queue length shown in the Figures 6a, 6b and 6c demonstrate similar trends as the graphs of the average response times. This was expected since the average queue length is strongly correlated with the response time for a fixed arrival rate. FCFS admission policy cannot be discarded entirely. As Figures 5c and 6c illustrate, for sufficiently low rate of arrivals and siginificantly large disk cache size, FCFS job scheduling can perform better than GRV. Using different cache replacement policies, e.g., greedy dual size, the same relative results are likely to be achieved. This is left for future work.

## 7 Conclusions and Future Work

We have identified a new type of caching problem that appears whenever dependencies exist among files that must be cached. This problem arises in various scientific and commercial applications that use vertically partitioned attributes in different files. Traditional caching techniques that make caching decisions one file at a time do not apply here as requests can only proceed if all files requested are cached. Since the problem of optimally loading the cache to maximize the value of satisfied requests is $NP$ hard, we settled on approximation algorithms that were shown analytically to produce solutions bounded from the optimal one. The MFC problem is also of theoretical interest in its own right because of its connection to the well known dense k-subgraph and the fact that any approximation to MFC can be used to approximate the latter problem with the same bounds from optimality.

Our simulation studies show that our new algorithms compare very favorably with file caching based on first come first serve ($FCFS$) scheduling of requests. The results indicate that system throughput using schedules based on Algorithm GRV are consistently higher than the FCFS based schedules and the queue length is shorter. Future work in this area involves conducting further detailed simulations with both synthetic workloads and real workloads derived from file caching activities of data intensive applications. We also intend to pursue these studies under file replacement policies that replace combinations of files rather than the more traditional algorithms that replace one file at a time.

## Acknowledgment

This work is supported by the Director, Office of Laboratory Policy and Infrastructure Management of the U. S. Department of Energy under Contract No. DE-AC03-76SF00098. This research used resources of the National Energy Research Scientific Computing (NERSC), which is supported by the Office of Science of the U.S. Department of Energy.

## References

1. Pei Cao and Sandy Irani. Cost-aware WWW proxy caching algorithms. In *USENIX Symposium on Internet Technologies and Systems*, 1997.
2. Uriel Feige, David Peleg, and Guy Kortsarz. The dense k-subgraph problem. *Algorithmica*, 29(3):410–421, 2001.
3. U. Hahn, W. Dilling, and D. Kaletta. Adaptive replacement algorithm for disk caches in hsm systems. In *16 Int'l. Symp on Mass Storage Syst.*, pages 128 – 140, San Diego, California, Mar. 15-18 1999.
4. Samir Khuller, Anna Moss, and Joseph (Seffi) Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.
5. S. O. Krumke, M. V. Marathe, D. Poensgen, S. S. Ravi, and Hans-Chris. Wirth. Budgeted maximum graph coverage. In *int'l. Workshop on Graph Theoretical Concepts in Comp. Sc., WG 2002*, pages 321 – 332, Cesky Krumlov, Czech Republic, 2002.
6. Ekow J. Otoo, Doron Rotem, and Arie Shoshani. Impact of admission and cache replacement policies on response times of jobs on data grids. In *Int'l. Workshop on Challenges of Large Applications in Distrib. Environments*, Seatle, Washington, Jun., 21 2003. IEEE Computer Society, Los Alamitos, California.
7. A. Shoshani, Bernado L., H. Nordberg, D. Rotem, and A. Sim. Multidimensional indexing and query coordination for tertiary storage management. In *Proc. of SSDBM'99*, pages 214 – 225, 1999.
8. M. Tan, M.D. Theys, H.J. Siegel, N.B. Beck, and M. Jurczyk. A mathematical model, heuristic, and simulation study for a basic data staging problem in a heterogeneous networking environment. In *Proc. of the 7th Hetero. Comput. Workshop*, pages 115–129, Orlando, Florida, Mar. 1998.
9. J. Wang. A survey of web caching schemes for the internet. In *ACM SIGCOMM'99*, Cambridge, Massachusetts, Aug. 1999.
10. Kesheng Wu, Wendy S. Koegler, Jacqueline Chen, and Arie Shoshani. Using bitmap index for interactive exploration of large datasets. In *SSDBM'2003*, pages 65–74, Cambridge, Mass., 2003.
11. N. Young. On-line file caching. In *SODA: ACM-SIAM Symposium on Discrete Algorithms (A Conference on Theoretical and Experimental Analysis of Discrete Algorithms)*, 1998.

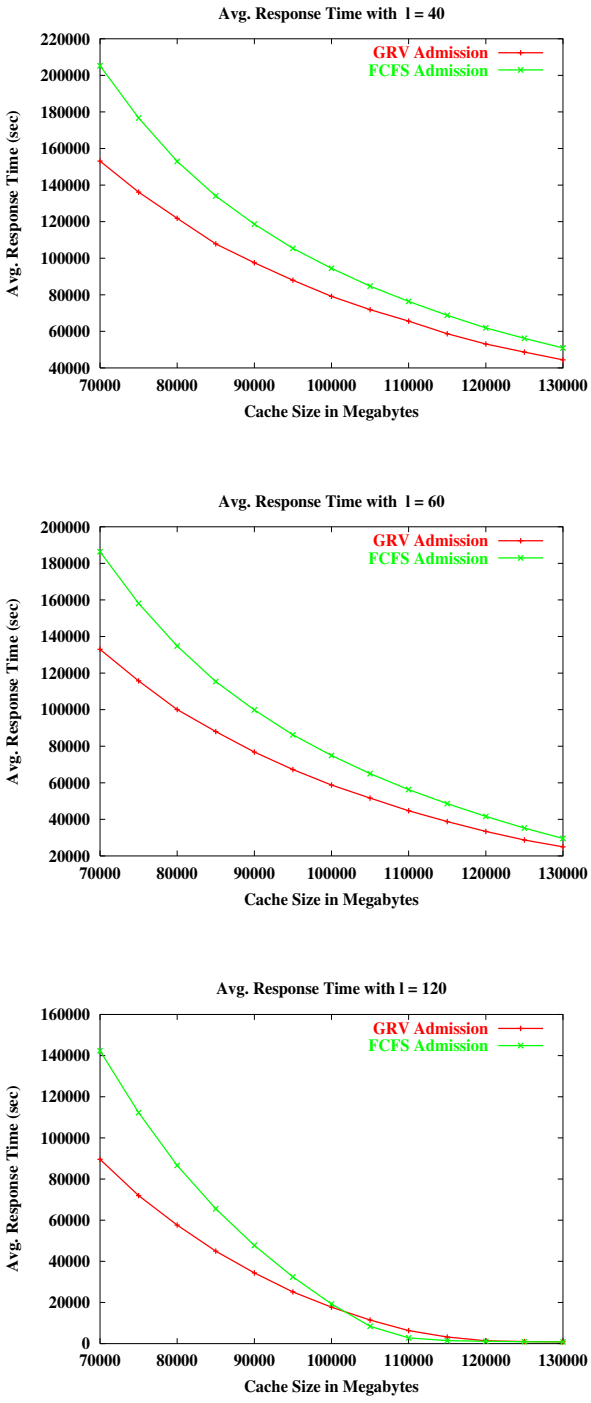


**Fig. 5.** Job Response Times of GRV vs FCFS Multi-File Caching Algorithm

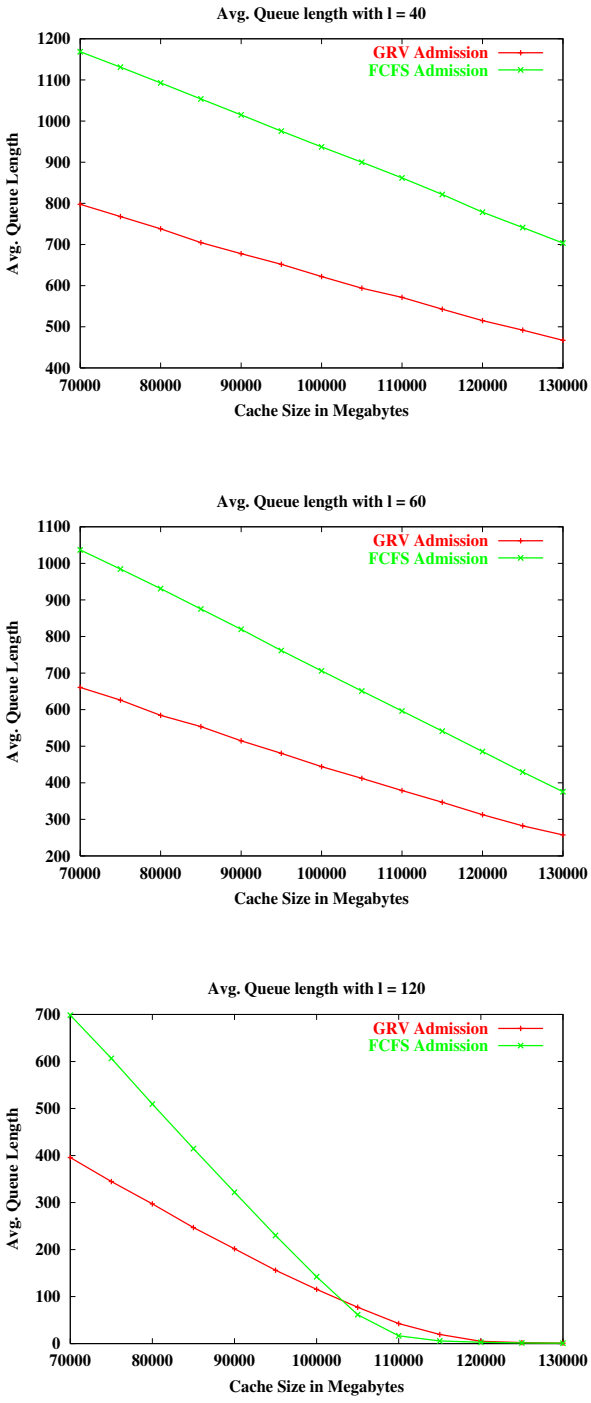


**Fig. 6.** Average Queue Length of GRV vs FCFS Multi-File Caching