COMPARING MULTIPLE FILE COPIES WITH A PRIMARY COPY USING MINIMAL COMMUNICATION

Khaled A. S. Abdel-Ghaffar*
Department of Electrical
and Computer Engineering
University of California
Davis, CA 95616

Amr El Abbadi†
Department of Computer Science
University of California
Santa Barbara, CA 93106

Abstract

For data consistency in distributed database systems, it is necessary to compare remotely located copies of a file. The cost of such comparison is in the amount of communication required to identify the erroneous pages in the different copies. We assume that the sites at which the copies reside can communicate only with a primary site that contains an exact copy of the file. The minimum amount of communication necessary to identify any given number of erroneous pages is determined and a technique to attain this minimum is presented.

Index Terms: Communication efficiency, fault location, replicated files, Reed-Solomon codes.

*This author was supported in part by an IBM Faculty Development Award.
†This author was supported in part by the National Science Foundation under grant IRI-9117094.

I. INTRODUCTION

Multiple copies of a file are often distributed to remote sites for accessibility purposes. One popular method for file replication is the primary-secondary approach [12], where a designated copy is called the primary copy, and the rest of the copies are referred to as the secondary copies. All updates are directed to the primary copy, who is responsible for updating the secondary copies. In this correspondence, we are interested in efficiently comparing secondary copies with the primary copy in order to detect any corruptions in the data stored at the secondary copies. Our approach uses ideas upon which Reed-Solomon codes are based.

Consider a file which is divided into N units, $P_1, \ldots, P_N$, called *pages*. Such a file will be denoted by $(P_1, \ldots, P_N)$. An exact copy of this file is located at a site, referred to as the *primary site*, while M copies are located at different sites, referred to as the *secondary sites*. Let $(P_{1,m}, \ldots, P_{N,m})$, where $m = 1, \ldots, M$, be the copy of the file that resides in the m^{th} secondary site. Thus, $P_{n,m}$ is the n^{th} page residing in the m^{th} secondary site. Ideally, all copies are identical and $P_{n,m} = P_n$ for $m = 1, \ldots, M$ and $n = 1, \ldots, N$. However, due to site failures, some pages may be corrupted. We assume that among the total number of NM pages that resides in all M secondary sites, at most f of them are corrupted. The goal is to be able, through (error-free) communication between sites, to compare the copies of the file residing in the M sites in order to determine the corrupted pages. We assume that secondary sites can only communicate with the primary site.

Since the pages may be quite large, it is common to compute for each page P_i a concise representation, called the *page signature* of P_i, and denoted by p_i [7]. Each page signature has length b bits, which is typically much less than the length of the page. Thus, two different pages may have the same page signature. Metzner [7] has developed a simple technique to assign signatures to pages using feedback shift registers. In this technique, it is assumed that the number of pages in the file, N, is less than the total number of page signatures 2^b, i.e., $b > \log_2 N$. This condition means that there are fewer bits in the page address than in the signature of a page. We will assume that the condition $b > \log_2 N$ holds throughout this paper.

Let $p_{n,m}$ denote the signature of page $P_{n,m}$. If $p_{n,m} \neq p_n$, then page $P_{n,m}$ is corrupted, i.e., $P_{n,m} \neq P_n$. However, the converse is not always true in the sense that $P_{n,m}$ may be corrupted although $p_{n,m} = p_n$. The probability that this happens can be made arbitrarily small by making b large. In practice, no attempt is usually made to identify all the corrupted pages since the sizes of the pages are very large and, therefore, comparing pages may be prohibitively expensive. For practical purposes, we are often content if all pages $P_{n,m}$ with corrupted signatures, i.e., with $p_{n,m} \neq p_n$, are identified. If b is large enough, we can be sufficiently confident that all corrupted pages have corrupted signatures and, hence, all of them will be identified. Let

$$\Phi_m = \{n : 1 \leq n \leq N \text{ and } p_{n,m} \neq p_n\} \tag{1}$$

be the set of indices of pages with corrupted signatures in the m^{th} secondary site and $f_m = |\Phi_m|$ be their number. Since we have at most f corrupted pages, $f_1 + \cdots + f_M \leq f$.

In order to identify the corrupted page signatures, the secondary sites exchange information with the primary site. The amount of communication is measured by the number of transmitted bits. For normalization, the cost of communication is defined to be the number of sequences, each of length b bits, that need to be transmitted. Each such sequence is called a *signature*. For example, a secondary site may send the bit-wise exclusive-or of a number of page signatures. This contributes one signature to the cost of communication. It should be noted that the cost of communication may depend on the location of the corrupted pages. Here, we are concerned with the number of signatures that need to be transmitted, in the worst case, in order to identify all the pages with corrupted signatures.

All algorithms proposed in the literature for comparing multiple copies of a file are designed for the case of a single secondary site, i.e., for $M = 1$. Metzner [7],[9] has developed several protocols that require the transmission of $f[(\log_2 N) + 1]$ signatures, where N is assumed to be a power of 2. In case of $f = 1$, Fuchs, Wu, and Abraham [5] proposed a different protocol that requires the exchange of $(\log_2 N) + 1$ signatures. Barbara, Garcia-Molina, and Feijoo [2] developed an algorithm to identify up to two pages with corrupted signatures, i.e., $f = 2$, after the exchange of $(\log_2 N)((\log_2 N)+1)/2+1$ signatures. Metzner and Kapturowski [10] introduced an algorithm that requires an exchange of $f+1$ signatures to identify up to f pages with corrupted signatures. A class of randomized strategies to identify any given number of pages was presented by Barbara and Lipton [3]. Schwarz, Bowdidge, and Burkhard [11] have developed a mechanism that is also able to detect and identify missing and extraneous pages. It should be noted that all these algorithms may fail, in some cases, to identify the pages with corrupted signatures unless $f = 1$. Recently, an algorithm to solve the problem is proposed that requires the transmission of $\min\{N, 2f\}$ signatures [1]. This algorithm always succeeds in identifying the pages with corrupted signatures. It is shown in [1] that this solution is optimal, in the sense that any solution that is guaranteed to identify the pages requires the transmission of at least $\min\{N, 2f\}$ signatures.

In this correspondence, we will consider the general case where the number of secondary sites, M, is arbitrary. Of course, by applying a protocol for the case $M = 1$ to each secondary site, we obtain a protocol for any value of M. The number of signatures that need to be sent in this case is M times the number of signatures that need to be sent in the former protocol. For example, the protocol presented in [1] yields a technique that requires the transmission of $M \min\{N, 2f\}$ signatures in order to identify up to f pages with corrupted signatures in M copies of a file composed of N pages. Since N is typically larger than $2f$, the number of signatures that need to be sent is $2Mf$. We will introduce in Section III a technique that requires the transmission of at most $(M-1)\min\{N, f\} + \min\{N, 2f\}$ signatures. For

$N \geq 2f$, this expression reduces to $(M+1)f$. Thus, a saving in communication costs by a factor of $2M/(M+1)$ can be achieved. In Section II we will prove that the new solution in optimal, in the sense that any solution that is guaranteed to identify up to f pages with corrupted signatures in M copies of a file composed of N pages, requires the transmission of $(M-1)\min\{N,f\}+\min\{N,2f\}$ signatures.

II. A LOWER BOUND

The following lower bound uses the same counting arguement upon which the Singleton bound is based [6].

Lemma 1 *The number of signatures that need to be transmitted in order to identify up to any f pages with corrupted signatures in M copies of a file composed of N pages is at least*

$$(M-1)\min\{N,f\}+\min\{N,2f\}.$$

Proof. Suppose, to get a contradiction, that it suffices to transmit at most

$$t < (M-1)\min\{N,f\}+\min\{N,2f\} \tag{2}$$

signatures in total. We will show that it is impossible to determine the pages with corrupted signatures even if it is known that all of them, if there are any, reside in one secondary site. First, we will show that at least $\min\{N,f\}$ signatures need to be sent or received by each secondary site in order to determine if the site contains any pages with corrupted signatures. Suppose that the m^{th} secondary site, where $m=1,\ldots,M$, sends or receives t_m signatures. There are exactly $2^{b\min\{N,f\}}$ tuples of page signatures that agree on the last $N-\min\{N,f\}$ pages but differ in at least one page signature among the first $\min\{N,f\}$ pages. If $t_m < \min\{N,f\}$, then there are two tuples of page signatures $\mathbf{p} = (p_1,\ldots,p_{\min\{N,f\}},p_{\min\{N,f\}+1},\ldots,p_N)$ and $\mathbf{p}' = (p'_1,\ldots,p'_{\min\{N,f\}},p_{\min\{N,f\}+1},\ldots,p_N)$ that agree on the last $N-\min\{N,f\}$ pages but disagree in at least one page signature among the first $\min\{N,f\}$ pages, such that the signatures of these tuples are identical. Thus, if all other sites, including the primary site, have page signatures $\mathbf{p}$, it is impossible to determine whether the m^{th} site has page signatures $\mathbf{p}$ or $\mathbf{p}'$. Therefore, each secondary site should send or receive $\min\{N,f\}$ signatures before determining which secondary site contains pages with corrupted signatures. Thus, in the worst case, $(M-1)\min\{N,f\}$ signatures must be sent or received by the secondary sites with correct copies before determining whether the remaining secondary site has any pages with corrupted signatures. Without loss of generality, assume that the remaining site is the M^{th} site. This site may have up to $\min\{N,f\}$ pages with corrupted page signatures. Thus, if t_M is the number of signatures sent or received by the M^{th} site, then according to (2),

$$t_M \leq t-(M-1)\min\{N,f\} < \min\{N,2f\}. \tag{3}$$

However, we have just argued that at least $\min\{N,f\}$ signatures need to be sent or received by each site, and in particular the M^{th} site, in order to determine whether it has any pages with corrupted signatures. Thus, $t_M \geq \min\{N,f\}$. This contradicts (3) in case $N \leq f$, and thus proves the lemma in this case. So, we assume in the following that $N > f$. There are exactly $2^{b\min\{N,2f\}}$ tuples of page signatures that agree on the last $N-\min\{N,2f\}$ pages but differ in at least one page signature among the first $\min\{N,2f\}$ pages. Furthermore, the number of distinct signatures that the M^{th} site can send or receive is 2^{bt_M}. From (3), there are two tuples of page signatures $\mathbf{p}'' = (p''_1,\ldots,p''_{\min\{N,2f\}},p_{\min\{N,2f\}+1},\ldots,p_N)$ and $\mathbf{p}''' = (p'''_1,\ldots,p'''_{\min\{N,2f\}},p_{\min\{N,2f\}+1},\ldots,p_N)$ that agree on the last $N-\min\{N,2f\}$ pages but disagree in at least one page signature among the first $\min\{N,2f\}$ pages, such that the signatures of these tuples are identical. Suppose that all other sites, including the primary site, has page signatures $(p''_1,\ldots,p''_f,p'''_{f+1},\ldots,p'''_{\min\{N,2f\}},p_{\min\{N,2f\}+1},\ldots,p_N)$. There are at most f differing page signatures between this tuple and each of $\mathbf{p}''$ and $\mathbf{p}'''$. Hence, the corrupted page signatures should be determined if the page signatures of the M^{th} site are either $\mathbf{p}''$ or $\mathbf{p}'''$. However, in both cases, the t_M signatures sent or received by the M^{th} site are identical, which implies that there is no way to determine the corrupted page signatures. This contradiction concludes the proof. □

III. ALGORITHMS

We present in this section two algorithms for the primary site to determine all pages with corrupted signatures using the minimum amount of communication as specified by Lemma 1. Algorithm I is used when $N \leq f$, while Algorithm II is used when $N > f$.

Algorithm I $(N \leq f)$

Step 1 Initialize $m = 1$.

Step 2 Request the N page signatures $p_{1,m}, \ldots, p_{N,m}$ from the m^{th} secondary site.

Step 3 Determine $\Phi_m = \{n : 1 \leq n \leq N \text{ and } p_{n,m} \neq p_n\}$.

Step 4 If $m < M$, increase m by 1 and return to step 2, otherwise, stop.

Algorithm I is simply an exhaustive technique to determine the pages with corrupted signatures. The following result obviously holds.

Lemma 2 *For $N \leq f$, Algorithm I determines all pages with corrupted signatures using MN transmitted signatures.*

In the following, we assume that $N > f$. Before presenting Algorithm II, some notation need to be introduced. Let α be a primitive element in $GF(2^b)$. Since $b > \log_2 N$, it follows that the elements α^j, where $j = 1, \ldots, N$, are distinct. For each positive integer j, we define the signatures $s_j = \sum_{n=1}^{N} p_n \alpha^{jn}$ and $s_{j,m} = \sum_{n=1}^{N} p_{n,m} \alpha^{jn}$, where $m = 1, \ldots, M$. These signatures are the syndromes of a Reed-Solomon code corresponding to the vectors $(p_1, \ldots, p_N)$ and $(p_{1,m}, \ldots, p_{N,m})$, respectively. The following remarks are based on the fact that the $J \times J$ Vandermonde matrix $[\alpha^{jn_j}]$, where $1 \leq j \leq J \leq N$ and $1 \leq n_1 < \cdots < n_J \leq N$, is nonsingular [4],[6].

Remark 1 *For any given* $E_1, \ldots, E_N \in GF(2^b)$, *the system of equations* $\sum_{n=1}^{N} e_n \alpha^{jn} = E_j$, *where* $j = 1, \ldots, N$, *has a unique solution for* $(e_1, \ldots, e_N)$.

Remark 2 *Suppose that* $(e_1, \ldots, e_N)$ *has weight at most* f. *If* $\sum_{n=1}^{N} e_n \alpha^{jn} = 0$, *where* $j = 1, \ldots, f$, *then* $e_n = 0$ *for* $n = 1, \ldots, N$.

Remark 3 *For any given* $E_1, \ldots, E_J \in GF(2^b)$, *where* $1 \leq J \leq N$, *the system of equations* $\sum_{n=1}^{N} e_n \alpha^{jn} = E_j$ *for* $j = 1, \ldots, J$, *has at most one solution for* $(e_1, \ldots, e_N)$ *of weight less than or equal to* $\lfloor J/2 \rfloor$. *This solution can be obtained using a Reed-Solomon decoder [4].*

Algorithm II $(N > f)$

Step 1 Initialize $m = 1$.

Step 2 Request the f signatures $s_{1,m}, \ldots, s_{f,m}$ from the m^{th} secondary site.

Step 3 Try to compute a vector $(e_{1,m}, \ldots, e_{N,m})$ of weight at most $\lfloor f/2 \rfloor$ as a solution to the system of equations

$$\sum_{n=1}^{N} e_{n,m} \alpha^{jn} = s_{j,m} - s_j, \quad \text{where} \quad j = 1, \ldots, f. \tag{4}$$

If such a solution exists, let w_m denote the weight of the vector $(e_{1,m}, \ldots, e_{N,m})$ and go to step 6. If no solution exists, continue to the next step.

Step 4 Request the signatures $s_{f+1,m}, \ldots, s_{\min\{N,2f\},m}$ from the m^{th} secondary site.

Step 5 Find a solution $(e_{1,m}, \ldots, e_{N,m})$ of weight at most f to the system of equations

$$\sum_{n=1}^{N} e_{n,m} \alpha^{jn} = s_{j,m} - s_j, \quad \text{where} \quad j = 1, \ldots, \min\{N, 2f\}. \tag{5}$$

Step 6 If $m < M$, increase m by 1 and return to step 2. Otherwise, continue to the next step.

Step 7 If a solution to (4) of weight at most $\lfloor f/2 \rfloor$ exists for all m, go to step 8. Otherwise, for $m = 1, \ldots, M$, let

$$\Phi_m = \{n : 1 \leq n \leq N \text{ and } e_{n,m} \neq 0\},$$

and stop.

Step 8 Initialize $m = 1$.

Step 9 If $w_m \leq \lfloor (w_1 + \cdots + w_M)/2 \rfloor$, let

$$\Phi_m = \{n : 1 \leq n \leq N \text{ and } e_{n,m} \neq 0\},$$

and go to step 12. Otherwise, continue to the next step.

Step 10 Request the signatures $s_{f+1,m}, \ldots, s_{\min\{N,2f\},m}$ from the m^{th} secondary site.

Step 11 Find a solution $(e'_{1,m}, \ldots, e'_{N,m})$ of weight at most f to the system of equations

$$\sum_{n=1}^{N} e'_{n,m} \alpha^{jn} = s_{j,m} - s_j, \quad \text{where} \quad j = 1, \ldots, \min\{N, 2f\}. \tag{6}$$

Let

$$\Phi_m = \{n : 1 \leq n \leq N \text{ and } e'_{n,m} \neq 0\}.$$

Step 12 If $m < M$, increase m by 1 and return to step 9, otherwise, stop.

We will show that this algorithm identifies Φ_m correctly in steps 7, 9, and 11 for all $m = 1, \ldots, M$. Furthermore, it requires the transmission of at most $Mf + \min\{N - f, f\}$ signatures. First, let m be fixed. From the definitions of s_j and $s_{j,m}$, we have

$$\sum_{n=1}^{N} (p_{n,m} - p_n) \alpha^{jn} = s_{j,m} - s_j. \tag{7}$$

We consider the following cases.

Case 1 $f_m \leq \lfloor f/2 \rfloor$: In this case, the weight of the vector $(p_{1,m} - p_1, \ldots, p_{N,m} - p_N)$ is at most $\lfloor f/2 \rfloor$. From (7) and Remark 3, it follows that there exists a unique solution to (4) given by

$$e_{n,m} = p_{n,m} - p_n, \quad \text{for} \quad n = 1, \ldots, N.$$

In particular, $w_m = f_m$ and steps 4 and 5 are skipped. If the algorithm stops at step 7, then the set Φ_m as derived in this step is exactly Φ_m as defined in (1). Similarly, if the algorithm reaches step 9 with $w_m \leq \lfloor (w_1 + \cdots + w_M)/2 \rfloor$, then the set Φ_m as derived in this step is exactly equal to Φ_m as defined in (1). On the other hand, if step 9 is reached and $w_m > \lfloor (w_1 + \cdots + w_M)/2 \rfloor$, then the algorithm proceeds to steps 10 and 11. From (7) and the fact that there are at most f pages with corrupted signatures, Remark 1 (in case $N \leq 2f$) and Remark 3 (in case $N > 2f$) imply that there exists a unique solution to (6) of weight at most f given by

$$e'_{n,m} = p_{n,m} - p_n, \quad \text{for} \quad n = 1, \ldots, N.$$

Hence, the set Φ_m as derived in step 11 is exactly equal to Φ_m as defined in (1).

Case 2 $f_m > \lfloor f/2 \rfloor$: Two possibilities need to be considered.

Case 2.1 There is no solution to (4) of weight at most $\lfloor f/2 \rfloor$: In this case the algorithm, after executing the m^{th} iteration of step 3, proceeds to steps 4 and 5. From (7) and the fact that there are at most f pages with corrupted signatures, Remark 1 (in case $N \leq 2f$) and Remark 3 (in case $N > 2f$) imply that there exists a unique solution to (5) of weight at most f given by

$$e_{n,m} = p_{n,m} - p_n, \quad \text{for} \quad n = 1, \ldots, N.$$

The algorithm stops after executing step 7 and the set Φ_m as derived in this step is exactly equal to Φ_m as defined in (1).

Case 2.2 There is a solution to (4) of weight at most $\lfloor f/2 \rfloor$: Let $(e_{1,m}, \ldots, e_{N,m})$ denote the solution and w_m be its weight. We will argue that $w_m \geq f - f_m + 1$. Since $w_m \leq \lfloor f/2 \rfloor$ and $f_m > \lfloor f/2 \rfloor$, then $e_{n,m} \neq p_{n,m} - p_n$ for some n. From (4) and (7), we have

$$\sum_{n=1}^{N} [e_{n,m} - (p_{n,m} - p_n)]\alpha^{jn} = 0, \quad \text{where} \quad j = 1, \ldots, f.$$

If $e_{n,m} \neq p_{n,m} - p_n$ for at most f values of n, then by Remark 2, it follows that $e_{n,m} = p_{n,m} - p_n$ for all values of n. This contradicts the fact that $e_{n,m} \neq p_{n,m} - p_n$ for some n. Therefore, $e_{n,m} \neq p_{n,m} - p_n$ for at least $f+1$ values of n. On the other hand, since $e_{n,m} \neq 0$ for exactly w_m values of n and $p_{n,m} - p_n \neq 0$ for exactly f_m values of n, it follows that $e_{n,m} = p_{n,m} - p_n = 0$ for at least $N - w_m - f_m$ values of n. Thus, $e_{n,m} \neq p_{n,m} - p_n$ for at most $w_m + f_m$ values of n. From this argument, we conclude that $w_m + f_m \geq f + 1$, i.e., $w_m \geq f - f_m + 1$.

Since $f_1 + \cdots + f_M \leq f$ and $f_m > \lfloor f/2 \rfloor$, it follows that $f_{m'} \leq \lfloor f/2 \rfloor$ for all $m' \neq m$. From Case 1, there exists a solution to (4) for all executions of step 3 and the algorithm proceeds to step 8. Since $f_{m'} \leq \lfloor f/2 \rfloor$ for all $m' \neq m$, it follows from Case 1 that $w_{m'} = f_{m'}$. Hence,

$$w_m \geq f - f_m + 1 \geq \sum_{m' \neq m} f_{m'} + 1 = \sum_{m' \neq m} w_{m'} + 1.$$

From this, we conclude that $w_m > \lfloor (w_1 + \ldots + w_M)/2 \rfloor$. Hence, the algorithm proceeds from step 9 to steps 10 and 11. From (7) and the fact that there are at most f pages with corrupted signatures, Remark 1 (in case $N \leq 2f$) and Remark 3 (in case $N > 2f$) imply that there exists a unique solution to (6) of weight at most f given by

$$e'_{n,m} = p_{n,m} - p_n, \quad \text{for} \quad n = 1, \ldots, N.$$

Hence, the set Φ_m as derived in step 11 is exactly equal to Φ_m as defined in (1).

We conclude that the algorithm identifies Φ_m correctly for all $m = 1, \ldots, M$. Next, we will show that the algorithm requires the transmission of at most $Mf + \min\{N - f, f\}$

signatures. Notice that f signatures are transmitted at each iteration of step 3. We consider the following cases.

Case I $f_m \leq \lfloor f/2 \rfloor$ for all $m = 1, \ldots, M$: Then, Case 1 applies for all m. Hence, step 4 is never executed. If $w_m \leq \lfloor (w_1 + \cdots + w_M)/2 \rfloor$, then step 10 is not executed after the m^{th} iteration of step 9. On the other hand, there is at most a single value of m such that $w_m > \lfloor (w_1 + \cdots + w_M)/2 \rfloor$. If such an m exists, the algorithm proceeds after the m^{th} iteration of step 9 to step 10 and $\min\{N-f, f\}$ signatures are transmitted. Thus, at most $Mf + \min\{N-f, f\}$ signatures are transmitted in this case.

Case II $f_m \geq \lfloor f/2 \rfloor$ for some m: Since $f_1 + \cdots + f_M \leq f$, this value of m is unique. We consider two possibilities.

Case II.I Case 2.1 applies: Therefore, no solution to (4) exists at the m^{th} iteration of step 3 and $\min\{N-f, f\}$ signatures are transmitted at step 4. Since $f_{m'} \leq \lfloor f/2 \rfloor$ for all $m' \neq m$, then Case 1 applies for all such m'. Thus, step 4 is skipped after all executions of step 3 except for the m^{th} one. Furthermore, the algorithm stops at step 7. We conclude that $Mf + \min\{N-f, f\}$ signatures are transmitted.

Case II.II Case 2.2 applies: Since $f_{m'} \leq \lfloor f/2 \rfloor$ for all $m' \neq m$, then from the discussion of Cases 1 and 2.2, it follows that step 4 is never executed. As stated in Case 2.2, we have $w_m > \lfloor (w_1 + \cdots + w_M)/2 \rfloor$. Therefore, the algorithm proceeds after the m^{th} iteration of step 9 to step 10 and $\min\{N-f, f\}$ signatures are transmitted. Since $w_m > \lfloor (w_1 + \cdots + w_M)/2 \rfloor$ implies that $w_{m'} \leq \lfloor (w_1 + \cdots + w_M)/2 \rfloor$ for all $m' \neq m$, it follows that step 10 is executed only once. We conclude that $Mf + \min\{N-f, f\}$ signatures are transmitted.

We have thus established the following result.

Lemma 3 *For $N > f$, Algorithm II determines all pages with corrupted signatures using at most $Mf + \min\{N - f, f\}$ transmitted signatures.*

From Lemmas 1, 2, and 3, the following result can be deduced.

Theorem 1 *The number of signatures that need to be transmitted in order to identify up to any f pages with corrupted signatures in M copies of a file composed of N pages is*

$$(M-1)\min\{N, f\} + \min\{N, 2f\}.$$

References

[1] K. A. S. Abdel-Ghaffar and A. El Abbadi, "An optimal strategy for comparing file copies," submitted to *IEEE Trans. Parallel and Distributed Syst.*, Oct. 1991.

[2] D. Barbará, H. Garcia-Molina, and B. Feijoo, "Exploiting symmetries for low-cost comparison of file copies," in *Proc. Int. Conf. Distributed Comput. Syst.*, San Jose, CA, pp. 471–479, June 1988.

[3] D. Barbará and R. J. Lipton, "A class of randomized strategies for low-cost comparison of file copies," *IEEE Trans. Parallel and Distributed Syst.*, vol. 2, pp. 160–170, Apr. 1991.

[4] R. E. Blahut, *Theory and Practice of Error Control Codes.* Reading, MA: Addison-Wesley, 1984.

[5] W. Fuchs, K. L. Wu, and J. A. Abraham, "Low-cost comparison and diagnosis of large remotely located files," in *Proc. Symp. Reliability Distributed Software and Database Syst.*, Los Angeles, CA, pp. 67–73, Jan. 1986.

[6] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes.* Amsterdam: North-Holland, 1977.

[7] J. J. Metzner, "A parity structure for large remotely located replicated data files," *IEEE Trans. Comput.*, vol. C-32, pp. 727–730, Aug. 1983.

[8] —, "Reliable and efficient broadcast of files to a group of locally interconnected stations," in *Proc. GLOBECOM '86*, Houston, TX, pp. 1762–1767, Dec. 1986.

[9] —, "Efficient replicated remote file comparison," *IEEE Trans. Comput.*, vol. C-40, pp. 651–660, May 1991.

[10] J. J. Metzner and E. J. Kapturowski, "A general decoding technique applicable to replicated file disagreement location and concatenated code decoding," *IEEE Trans. Inform. Theory*, vol. IT-36, pp. 911–917, July 1990.

[11] T. Schwarz, R. W. Bowdidge, and W. A. Burkhard, "Low cost comparisons of file copies," in *Proc. Int. Conf. Distributed Comput. Syst.*, Paris, France, pp. 196–202, May 1990.

[12] M. Stonebraker, "Concurrency control and consistency of multiple copies of data in distributed INGRES," *IEEE Trans. Software Engineering*, vol. SE-5, pp. 188–194, May 1979.