Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2012, Article ID 868943, 9 pages
doi:10.1155/2012/868943

Research Article

Two-Layer Storage Scheme and Repair Method of Failure Data in Wireless Sensor Networks

Yulong Shen,[1, 2] Xiaowei Dang,[1] Min Shu,[1] Ning Xi,[1] and Jianfeng Ma[1]

[1] *School of Computer Science and Technology, Xidian University, Shaanxi, Xi'an 710071, China*
[2] *Department of Computer Science, Wayne State University, Detroit, MI 48202, USA*

Correspondence should be addressed to Yulong Shen, ylshen@mail.xidian.edu.cn

Received 20 July 2012; Accepted 8 September 2012

Academic Editor: Fu Xiao

Distributed data storage is a key technology in the data collection in wireless sensor networks. The storage scheme based on network coding is applied to data collection in wireless sensor networks because of its high reliability and low overhead. However, it is an open problem to reduce data repair communication overhead caused by the failure of storage nodes. This paper focuses on this issue and presents a two-layer distributed data storage scheme. The lower-layer nodes store the encoded data blocks and the upper-layer nodes store the re-encoded blocks that are responsible for failure data recovery. Based on the two-layer data storage scheme, a data repair method is proposed to decrease the repair communication overhead with only sacrificing lower storage overhead. Compared with MSR, interference alignment-based scheme and group interference alignment scheme, the proposed method has lower repair communication overhead. We prove that the proposed method can reduce the repair communication overhead to $o(1/\sqrt{k})$ times and it is suitable to resource-constrained distributed wireless sensor networks.

1. Introduction

Wireless sensor networks, Internet of things and M2 M make the computer networks extended to things. The data collection in wireless sensor networks perform environmental monitoring, information transmission, data storage, and independent provided service. The collected data are distributed stored at the data storage nodes. In the case of node failure, high reliability and low-overhead distributed data storage is an open problem which has received widespread attention in recent years [1, 2]. With the consideration of resource-constrained property of the storage nodes in the data collection in wireless sensor networks, some effective distributed storage methods are proposed in [3–8]. The distributed storage scheme based on network coding is one of them and it has been researched extensively. It encodes the original data into a number of encoded data blocks and then stores them at different storage nodes. To reconstruct the original data, users only need to get a reasonable number of encoded data blocks (not less than the original data amount). Compared with the data backup method, the network coding technique has advantages of low storage overhead and robustness.

But it also brings the repair problem: When a storage node fails, the data on nonfailure nodes are used to repair the failure data to keep the same level reliability. To repair the failure encoded data blocks, the traditional method is that the newly added storage node collect sufficient encoded data (not less than the original data amount), then decode the encoded data to get the original data, and reencode the original data blocks to recover the failure data. As a result, traditional method causes great repair communication overhead which makes it not optimal for the wireless sensor networks because of the strict limitation of energy. Therefore, communication overhead becomes the primary factor of designing data repair algorithm. In order to reduce the repair communication overhead, some researchers focus their attention on designing data repair algorithms. Among these repair algorithms, the interference alignment repair algorithm presented in [9] and regenerating codes repair algorithm introduced in [7, 8] are outstanding.

For regenerating codes repair algorithm, there are two interesting points on its optimal tradeoff curve: minimum-bandwidth regenerating (MBR) codes and minimum-storage regenerating (MSR) codes [7, 8]. For MSR codes, its core constructive techniques are interference alignment and the network coding. With the interference alignment technique, MSR codes can reduce the repair communication overhead. The more interference elements are aligned, the more communication overhead is reduced. In order to decrease the interference dimension to a maximum extent, a common eigenvector should be computed and used. But when the interference dimension is more than 3, the complexity of computing the common eigenvector will increase greatly and that will become a significant challenge to the wireless network nodes with limited calculation ability. Dimakis et al. [7] have proved that the repair communication overhead of exact-MSR codes repair algorithm can be achieved with interference alignment only when the code rate is not higher than 1/2. For MBR codes, it can decrease the repair communication overhead to the minimum by sacrificing significant storage overhead. Its storage overhead is about 2 times of MSR. So for the given redundancy, the MBR codes are no longer optimal in terms of reliability.

The interference alignment repair algorithm is based on a hybrid storage model. In this model, the data consist of two parts. One is systematic part which is composed of the original data blocks. The other is nonsystematic part which is composed of the linear combination of the systematic part. Using the interference alignment algorithm to repair the failure data, the newly added node should first collect sufficient data to reduce the interference factors to 3 and then use the interference alignment technique to repair the failure data. The method proposed in [9] can reduce the repair communication overhead, but it is still high.

This paper focuses on optimizing the repair communication overhead of distributed storage. Having analyzed the tradeoff between storage overhead and repair communication overhead and turned the flat storage structure into hierarchical storage structure, this paper proposes a distributed storage method, which is based on two-layer storage structure, and a data repair algorithm. The proposed algorithm decreases the repair communication overhead by sacrificing lower storage overhead. The two-layer storage structure has two kinds of encoded data. They are the encoded network coding data and reencoded data. The lower-layer nodes are responsible for the original data reconstruction. The upper-layer nodes are responsible for failure data recovery. The data repair algorithm based on two-layer storage structure can ensure the restored data and the original encoded data have the recoverability property. Moreover, the two-layer storage structure keeps the storage system in a dynamic steady state all the time. That is, the data reliability of the entire system is stable. The analysis shows that the proposed method can greatly reduce the repair communication overhead by sacrificing lower storage overhead. This paper also proves that the proposed method can reduce the repair communication overhead to $o(1/\sqrt{k})$ times of the traditional method at least and satisfies the basic requirements of sensor networks.

This paper is organized as follows. Section 2 is related works. Section 3 proposes two-layer data storage scheme and data repair method. Section 4 evaluates the repair communication overhead of the proposed method. The conclusion is in Section 5 .

2. Related Work

There is a tradeoff between the communication overhead of repairing failure data and the storage overhead in a distributed data storage system. To ensure the availability of the stored data, some methods to balance the storage overhead and the communication overhead of repairing the failure data are proposed in [1, 7, 9–14]. There are three kinds of recovery algorithms: regenerating codes recovery algorithm [1, 7, 11–14], interference alignment recovery algorithm [1, 9], and tree-structured recovery algorithm based on network topology [10].

For the regenerating codes recovery algorithm, MSR codes and MBR codes [7] can be its representation because of most researchers having a huge interest in them. With the consideration of the minimum storage overhead, MSR codes have minimal storage overhead on a single storage node. MSR codes use the interference alignment technique to reduce the repair communication overhead and the repair process can be seen in [7]. Compared with the traditional backup method, MSR codes repair algorithm can significantly reduce the repair communication overhead by interference alignment technique. But MSR codes have some deficiencies. (1) Using interference alignment technique to reduce the interference dimension to 1 should use the common eigenvectors of all the interfering elements. But the complexity of computing the common eigenvectors will greatly increase with the increase of interference elements. It will be a great challenge to the wireless storage nodes with limited calculation ability. (2) Repair communication overhead of exact-MSR codes repair algorithm can be achieved only when the coding rate is at most 1/2 [7], otherwise its desired results cannot be guaranteed.

MBR codes consider the repair problem with the view of minimum repair communication overhead. Repair communication overhead of MBR codes is equal to its single-node storage overhead. And its repair communication overhead is minimal among all the known data repair algorithms. The meticulous process of building MBR codes is in [7]. Nevertheless, the shortcoming of MBR codes is its great storage overhead. That is because each data block is stored twice. So for the given redundancy, the reliability of MBR codes is no longer optimal.

With minimal storage overhead for a single storage node, the data repair method based on interference alignment reduces repair communication overhead by merging interference elements. The interference alignment technique can reduce the interference elements to 1 by collecting sufficient data. Then the failure data can be repaired by solving linear equations. The details are shown in [9]. The shortcomings of the interference alignment repair algorithm are as follows: (1) Interference alignment repair algorithm is mainly acted on the systematic data. When non-systematic data are failure, they are turned into systematic. (2) Compared with the traditional method, the interference alignment repair algorithm does not decrease the repair communication overhead significantly.

Data repair algorithm based on network topology is named tree-structured data regeneration. This method views the data repair as recoding of the encoded data blocks. This method is based on the random network coding theory. And it is chiefly used to reduce the repair time. Compared with the traditional method, its repair communication overhead is not reduced.

For these shortcomings of the available data repair algorithms, this paper analyzes the tradeoff between the storage overhead and repair communication overhead, transforms the flat node storage method into hierarchical network coding storage method which is inspired by the tree-structured data regeneration, and then proposes the distributed network coding storage method and repair algorithm.

3. Two-Layer Storage Structure and Data Repair Method

In this section, two-layer storage structure and data repair method are proposed. In the two-layer storage structure, the encoded data blocks are organized into two layers.

3.1. The Construct of Two-Layer Data Storage Structure. There are two types of encoded data blocks in the two-layer storage structure and they constitute the lower-layer and the upper-layer of the two-layer data storage structure, respectively. The network coding data blocks of the original data consist of the lower-layer data blocks; while the upper-layer encoded data blocks are the linear combination of the lower-layer data blocks. The construction process is as follows.

The original data of size M bits are divided equally into k blocks (of size M/k bits each), represented by a k-dimension vector $E = [e_1, e_2, \dots, e_{k-1}, e_k]^T$. These k data blocks are expended into n encoded blocks by linear network coding, represented by a n-dimension vector $B = [b_1, b_2, \dots, b_{n-1}, b_n]^T$, that is, $B = AE$:

$$B = AE = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,k} \\ a_{2,1} & a_{2,2} & \dots & a_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,k} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_k \end{bmatrix}, \quad (1)$$

where A denotes the $n * k$ encoding matrix. The n encoded data blocks are allocated to the lower-layer storage nodes. The corresponding nodes that stored these encoded data blocks are the lower-layer nodes of the two-layer storage structure. From the property of the network coding, we know that the data at lower layer can reconstruct the original data.

The traditional data repair method illustrates that the data repair is essentially the process of solving linear equations. The communication overhead of the data recovery depends on the required data blocks, which are also the number of linear equations. Thus, reducing the communication overhead of data recovery is to decrease the number of equations.

To reduce the number of equations, the upper-layer encoded data scheme is proposed. The upper-layer nodes store the reencoded data from the lower-layer data by (m, n) code, where m is the number of the upper-layer encoded data blocks. Similar to the original data blocks, the upper-layer encoded data can also be denoted by $C = [c_1, c_2, \dots, c_{m-1}, c_m]^T$, that is, $C = FB'$:

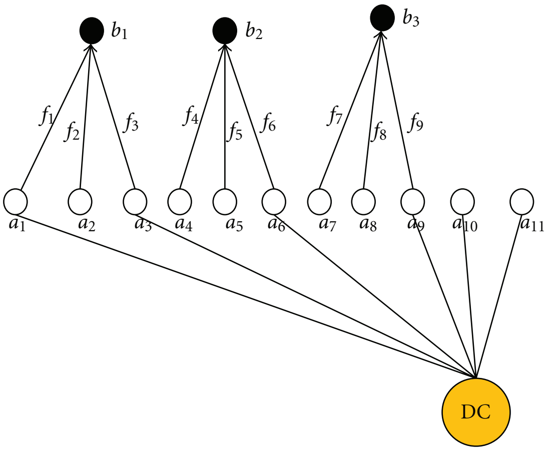

FIGURE 1: Two-layer encoded data structure with (3, 11, 6) code.

$$C = FB' = \begin{bmatrix} f_{1,1} & f_{1,2} & \dots & f_{1,mp} \\ f_{2,1} & f_{2,2} & \dots & f_{2,mp} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m,1} & f_{m,2} & \dots & f_{m,mp} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{mp} \end{bmatrix}, \quad (2)$$

where F is the $m * mp$ encoding matrix of the upper layer encoded data. p is the number of the lower layer data blocks for encoding a data block of the upper layer and $p < k$. Each row of F has p non-zero elements and every column of F has only one nonzero element. Moreover, $m \le \lceil n/p \rceil$ and $n - mp$ are the number of the lower layer data blocks that are not involved in re-encoding the upper layer data blocks. The upper layer encoded data blocks can be expressed by $c_i = f_i B'$. c_i denotes the upper layer encoded data. f_i is a mp-dimension row vector of F_{m*mp}. B', which is the subvector of B, is a mp-dimension row vector and represents the lower-layer encoded data participated in re-encoding the upper-layer data. Afterwards, the upper layer data blocks are stored at different nodes which are different from the lower-layer nodes, and these nodes are the upper-layer nodes of the two-layer storage structure. For convenience, the two-layer encoded data structure is represented by triple (m, n, k) code. Figure 1 is an example and shows a two-layer encoded data structure with (3, 11, 6) code.

3.2. The Methods of Repairing the Failure Data. The exact repair means that the recovered data are exactly the same as the failures. For all of the upper-layer nodes and the lower-layer nodes involved in re-encoding process, when any of them is failure, the failure data can be exactly repaired. For the rest of the lower-layer nodes, when one of them fails, there are two ways, exact repair and functional repair, to recover the failure data. The functional repair is that the newly generated block can contain the different failure data as long as the system maintains the network coding (n, k)

property. Two types of repairs are presented as follows: exact repair and hybrid repair.

3.2.1. Exact Repair. In exact repair, when the upper-layer nodes failed, the new joined node should only collect the lower-layer data blocks which are responsible for the previous re-encoded block to exact repair the failure data. For the lower-layer node participated in re-encoding, such as node a_1 in Figure 1, if it failed to repair the failure data exactly, the data stored at a_2, a_3, and b_1 are only required.

For the rest of the lower-layer nodes, if one of them failed, the upper-layer data blocks are used to realize the exact repair. The subsets of the upper-layer encoded blocks can exactly repair the failure data blocks stored at the lower layer nodes that are not involved in re-encoding process and the number of elements in each subset is p' (p' is the number of the upper-layer data blocks that are used for data repair. To maintain (n, k) network coding property, $pp' \geq k$). Moreover, any two of the subsets are required to have no intersection to ensure the failure data can be accurately repaired. As a result, m should satisfy $m/p' \geq n - mp$. The left of the inequality represents the number of the subsets, while the other side of the inequality represents the number of the lower layer nodes that are not involved in re-encoding process. Then with the help of the well-designed repair matrix, the failure data blocks can be repaired exactly by collecting the corresponding data blocks of the upper layer and combining them linearly afterwards.

3.2.2. Hybrid Repair. Hybrid repair is a hybrid model of the exact repair and functional repair. The hybrid model is: If any of the upper-layer nodes or the lower-layer nodes involved in reencoding process failed, the data stored at them are exactly repaired; however, if the lower-layer nodes which are not involved in the reencoding process failed, the data stored at them are functionally repaired. The functional repair is actually the linear combination of p' ($p' \geq 2$) blocks of the upper-layer data. For example, as show in Figure 1, if a_{10} failed, the data stored at a_{10} can be functionally repaired by the data stored at b_1 and b_2. With the help of the accurately calculated repair coefficients, the repaired data can preserve the network coding (n, k) property. For convenience, the new joined node storing the recovered data is still named a_{10}. To keep the same level reliability of the data, the functional repair must make sure that the repaired lower layer storage system maintains (n, k) network coding property.

The functional repair of the data stored at the lower layer nodes which are not involved in re-encoding process is a linear combination of data blocks of the upper layer. Therefore, the functional repair can be represented by $r_i = \xi_i C$. r_i is the recovered data and ξ_i is a m-dimension row vector for repair. Each row of $S = [\xi_1 \xi_2 \cdots \xi_{j-1} \xi_j]^T$ has p' nonzero elements, where$j$ values at least $n - mp$ and each column of S has only one nonzero element. To ensure the repaired lower layer storage system maintains (n, k) network coding property, we should work out all the encoding vectors that have (n, k) network coding property with that of the data stored at the lower layer in advance and their number is set to n_1 which is not less than $n - mp$. Then strictly calculate the encoding coefficients of the upper layer data and make the final repair vectors equal to the vectors we calculated beforehand. At this time j is equal to n_1. Therefore, the key problem of functional repair is to calculate these coefficients. Their number is $mp + (n - mp)p'$. Among them mp coefficients are used for encoding the upper-layer data and the others are used to repair the failure data. The repair vector of the data repaired can be expressed in the form of the summation of the product of these coefficients. These coefficients can be calculated with the help of repair vectors of the ultimate repaired data. Data repair can also be represented by $R = S[F(AE)]$ and Rank $(R) \leq j$. To maintain (n, k) network coding property, the rank of R must be greater than $n - mp$. The solution of these coefficients exists when the rank of R equals to that of its augmented matrix by choosing the value of the non-zero elements reasonable. Because of $n - mp < mp + (n - mp)p'$, the solution will not be unique.

4. Evaluation

In this section, we analyze the proposed data repair method and evaluate the communication overhead for data repair, namely, repair communication overhead. We also compare it with the existing data repair method. The communication overhead for data repair is represented by the amount of communication data in the data repair process.

4.1. Repair Communication Overhead Evaluation of the Exact Repair

4.1.1. Repair Communication Overhead. Whether the lower-layer data blocks that are involved in re-encoding process or not brings the difference of the repair schemes and even makes the repair communication overhead not the same. From the repair methods mentioned above, we know that the repair communication overhead of the upper-layer data blocks and the lower-layer data blocks involved in re-encoding process is p, and the repair communication overhead of the rest of the lower layer data blocks is p'. p is not necessarily equal to p'. When we compute the communication overhead of data repair, it is not logical to use an accurate value to represent the repair communication overhead of the entire storage system. Therefore, the expectation value of the repair communication overhead provides a good idea to represent the repair communication overhead of the entire storage system. Let m be x, let p be y, and let p' be z. We assume that each storage node has the same failure probability in the entire storage system. When a storage node failed, the expectation of its repair communication overhead is

$$E(x, y, z) = \frac{(x + xy)y}{n + x} + \frac{(n - xy)z}{n + x}, \tag{3}$$

where x, y, and z are subject to

$$k \leq yz \leq \left\lceil \sqrt{k} \right\rceil^2, \tag{4}$$

$$2 \leq y \leq \left\lceil \frac{k}{2} \right\rceil, \tag{5}$$

$$2 \le z \le \left\lceil \frac{k}{2} \right\rceil, \tag{6}$$

$$x \ge z(n - xy), \tag{7}$$

$$n - xy \ge 0, \tag{8}$$

$$\begin{aligned} E(x, y, z) &= \frac{(x + xy)y}{n + x} + \frac{(n - xy)z}{n + x} \\ &= \frac{xy^2 - xyz + nz + xy}{n + x}. \end{aligned} \tag{9}$$

Set $t = n - xy$, then $n/x = y + t/x$. E will be turned into

$$\begin{aligned} E(x, y, z) &= \frac{xy(y - z + 1) + nz}{n + x} \\ &= \frac{y(y - z + 1) + (y + t/x)z}{1 + y + t/x} \\ &= \frac{y^2 + y + (t/x)z}{1 + y + t/x} \\ &= \frac{y(y + 1 + tz/xy)}{1 + y + t/x}. \end{aligned} \tag{10}$$

From the equation, we know that the value of E is related to x, y, and z. x is similarly related to y and z. As a result, the value of E is determined by y and z. According to the relationship between y and z, we will have the following 3 cases.

Firstly, if $y = z$, then $E = y$. From (4), we know that $y_{\min} = \lceil \sqrt{k} \rceil$. At this time, we can calculate that $E = \lceil \sqrt{k} \rceil$.

Secondly, if $y < z$, then $E > y$. For the given value of z, the range of y is determined by the formula (4), and at the same time the range of x is determined by the formula (8). Within the range of x and y, it can be thought that x, y, and z are independent of each other. For formula (3), the partial derivative of x is:

$$\frac{\partial E}{\partial x} = \frac{n(y^2 - yz + y - z)}{n + x} = \frac{n(y - z)(y + 1)}{(n + x)^2}. \tag{11}$$

When x is maximal, E will be minimal. From (8), we can see that $x_{\max} = n/y_{\min}$ and $y_{\min} = 2$. At this time, $z = \lceil k/2 \rceil$, then $E = 2$.

Thirdly, if $y > z$, then $E < y$. Similar to the $y < z$ situation, when x is minimal, E is minimal. From (7), we can see that $x \ge zn/(1 + yz) = n/(y + 1/z)$. From (4), (5), and (6), we know that if $y = \lceil k/2 \rceil$ and $z = 2$, x will be minimal and at this time $x_{\min} = 2n/(1 + k)$, $E = (1 + k)^2/(6 + 2k)$.

Compare the value of E at these 3 situations and we can see that if $k = 3, 4$, the minimal repair communication overhead is $(1 + k)^2/(6 + 2k)$; if $k \ge 5$, the minimal repair communication overhead is 2.

4.1.2. Evaluation of the Repair Communication Overhead

Theorem 1. *If $k \ge 3$ and the relationship between n and k is $k + 1 \le n \le 2k - 1$, the repair communication overhead of exact repair based on the two-layer storage structure is lower than that of MSR.*

Proof. If $k \ge 5$, the repair communication overhead of exact repair based on the two-layer storage structure is 2, while the MSR is $d/(d - k + 1)$ [7], where d is the number of nodes that are involved in data repair and $k \le d \le n - 1$. Let $f(d) = d/(d - k + 1)$ and its derivative of d is a monotone decreasing function. So $f(d)$ will be minimal when $d = n - 1$ and $f(d)_{\min} = (n - 1)/(n - k) = 1 + (k - 1)/(n - k)$. From the condition that $k + 1 \le n \le 2k - 1$, we can know $f(d)_{\min} = 1 + (k - 1)/(n - k) \ge 2$. Therefore, it turns to be correct that the repair communication overhead of the exact repair based on two-layer storage structure is lower than that of MSR when $k \ge 5$. Moreover, when $k = 3, 4$, the repair communication overhead of the exact repair based on two-layer storage structure is lower than 2. That is to say when n and k satisfy $k + 1 \le n \le 2k - 1$, the conclusion is also correct. Hence, theorem 1 proves to be correct. □

Theorem 2. *If $k \ge 3$, the repair communication overhead of exact repair based on two-layer storage structure is lower than that of repair method based on interference alignment which is proposed in [9].*

Proof. The repair communication overhead of the basic interference alignment repair algorithm proved by [9] is $(qk - q + 1)/q$ (q is the number of data pieces stored at a single storage node). Let $f(q) = (qk - q + 1)/q = k - 1 + 1/q$. If $k \ge 3$, then $f(q) > 2$. The repair communication overhead of exact repair based on two-layer storage structure is 2 if $k \ge 5$. For $k \ge 5$, the conclusion is correct. When k are 3, 4, the repair communication overhead of exact repair based on two-layer storage structure is 4/3 and 25/14, respectively. Both of them are smaller than 2, so the conclusion is also correct when k are 3, 4.

For the repair algorithm based on group interference alignment, the repair communication overhead is $p + (k - p)/q$ and it is higher than p (p is the number of storage nodes that a data group contains, $2 \le p < k$). If $k \ge 3$, the repair overhead of group interference alignment is $p + (k - p)/q$ and it is higher than 2. Therefore, the Theorem 2 is obviously correct for the repair method based on group interference alignment. □

4.1.3. Numeral Result. In order to compare the repair overhead of these data repair method above and verify the correctness of the conclusions, the numeral result is shown in Figure 2. The histogram in Figure 2 shows, respectively, the repair overhead of MSR repair, exact repair based on two-layer storage structure, repair based on basic interference alignment and group interference alignment, where (n, k) values are as follows: (5, 3), (6, 4), (9, 5), (12, 6), (15, 7), and (17, 8). Comparing the conclusions of this paper with the numeral results displayed in Figure 2, it can be seen apparently that they are consistent.

4.2. Repair Overhead Evaluation of the Hybrid Repair

4.2.1. Repair Communication Overhead. Similar to the exact repair, the communication overhead of hybrid repair can also be represented by its expectation value. Let m be x, let p be y,

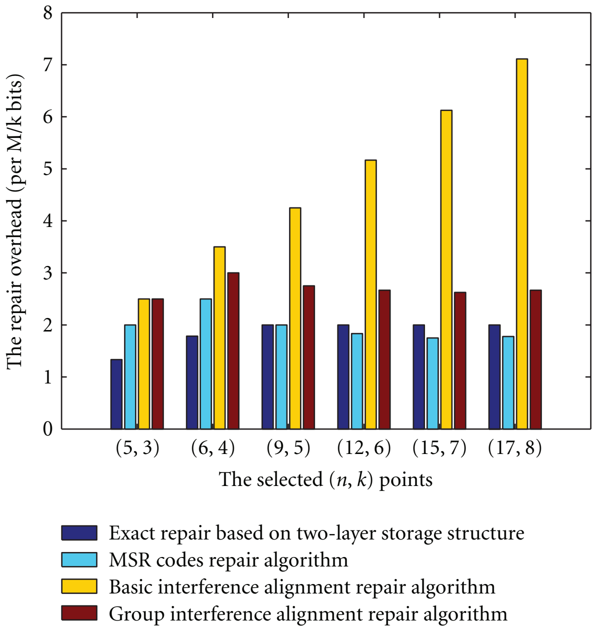

Figure 2: The repair communication overhead evaluation of exact repair.

and let p' be z, and assume that in the entire storage system, the failure probability of each storage node is exactly the same. When a storage node failed, the expectation of repair overhead is

$$E(x, y, z) = \frac{(x + xy)y}{n + x} + \frac{(n - xy)z}{n + x}, \tag{12}$$

where x, y, and z are subject to

$$k \le yz \le \left\lceil \sqrt{k} \right\rceil^2, \tag{13}$$

$$xy < n, \tag{14}$$

$$z < x, \tag{15}$$

$$2 \le y \le \left\lceil \frac{k}{2} \right\rceil, \tag{16}$$

$$2 \le z \le \left\lceil \frac{k}{2} \right\rceil, \tag{17}$$

$$C(x, z) \ge n - xy, \tag{18}$$

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n, \tag{19}$$

$$n - xy \le n_1, \tag{20}$$

$$n > n_1. \tag{21}$$

We can gain

$$\begin{aligned} E(x, y, z) &= \frac{(x + xy)y}{n + x} + \frac{(n - xy)z}{n + x} \\ &= \frac{xy^2 - xyz + nz + xy}{n + x}. \end{aligned} \tag{22}$$

Let $t = n - xy$. From (19), we can draw $n! \to \sqrt{2\pi n}(n/e)^n$ [15]. So, $C(x, z)$ is

$$\begin{aligned} C(x, z) &\approx \frac{x^{x+1/2}e^{1/12n}}{\sqrt{2\pi}z^{z+1/2}(x - z)^{x-z+1/2}} \\ &\approx \frac{x^{x+1/2}(1 + 1/12n)}{\sqrt{2\pi}z^{z+1/2}(x - z)^{x-z+1/2}}. \end{aligned} \tag{23}$$

When $t \to C(x, z)$, the storage overhead will be minimal, and the value of t is $\mathrm{Min}(n_1, C(x, z))$.

From (12), we can see that the value of E is related to x, y, and z. The relationship between x, y, and z can be seen from (18). Formula (13) gives the condition that y, z should be satisfied. When any of x, y, and z is determined, the range of the others will be known. Within the range of them, it can be thought that x, y, and z are independent of each other. For the given value of$y$, the range of x and z can respectively be determined and within the range of x and z, the three variables are independent of each other. For formula (12), the partial derivative of x is

$$\frac{\partial E}{\partial x} = \frac{n(y^2 - yz + y - z)}{(n + x)^2} = \frac{n(y - z)(y + 1)}{(n + x)^2}. \tag{24}$$

And the partial derivative of z is

$$\frac{\partial E}{\partial z} = \frac{n - xy}{n + x}. \tag{25}$$

Combine (23) and (18), we can have

$$f(x, y, z) = \frac{x^{x+1/2}(1 + 1/12n)}{\sqrt{2\pi}z^{z+1/2}(x - z)^{x-z+1/2}} - (n - xy). \tag{26}$$

Formula (18) can be rewritten as $f(x, y, z) \ge 0$. To compute convenient, by the Taylor formula, (26) can be simplified to

$$f(x, y, z) = \frac{A(1 + (1/12n))}{B\sqrt{2\pi}} - (n - xy), \tag{27}$$

$$\begin{aligned} A &= 1 + 105.9(x - 3) + 126.5(x - 3)^2, \\ B &= \left[1 + 11(z - 2) + 11.7(z - 2)^2\right] \\ &\quad \times \left[1 + 1.5(x - z - 1) + 1.375(x - z - 1)^2\right]. \end{aligned} \tag{28}$$

Now, we discuss the value of E. According to the relationship between y and z, we will have the following 3 situations.

First, if $y = z$, then $E = y$. From (13), we will know that $y_{\min} = \lceil \sqrt{k} \rceil$. At this time, we can calculate that $E = \lceil \sqrt{k} \rceil$.

Second, if $y > z$, formulas (24) and (25) show us the value of E increases with the increasing of x, z and the values of x, z decrease with the increase of y which are known from (14) and (20). So for the given y, to make E minimal, x and z should be minimal within their range. When $y = a(\lceil \sqrt{k} \rceil + 1 \le a \le \lceil k/2 \rceil)$, $z = \lceil k/a \rceil$ and the minimum of x can be drawn from (27).

Theorem 3. *Under the condition that the relationship between n and k is $n/k \leq \sqrt{e}/(\sqrt{e}-1)$, for the given y and z, when $f(x,y,z) \geq 0$, it makes E minimal.*

Proof. x, y are known, so (27) is a function on x and

$$\frac{df}{dx} = y + \left[\ln\frac{x}{x-z} - \frac{0.5z}{x(x-z)}\right]x^{x+1/2}(x-z)^{-(x-z+1/2)}. \tag{29}$$

For $y = a$, $z = \lceil k/a \rceil$, let $g(x) = x/(x-z) \; - e^{0.5z/x(x-z)}$, $x/(x-z)$ decreases with the increase of x. Formula (14) shows that $x \leq [n/a]$, so

$$\frac{x}{x-z} \geq \frac{[n/a]}{[n/a] - \lceil k/a \rceil} > \frac{\lceil n/a \rceil}{\lceil n - k/a \rceil} = \frac{n}{n-k} \geq \sqrt{e},$$

$$\sqrt{e} < \frac{x}{x-z} \leq \left\lceil \frac{k}{a} \right\rceil + 1, \tag{30}$$

$$\frac{0.5z}{x(x-z)} < 0.5,$$

then

$$e^{0.5z/x(x-z)} < \sqrt{e}. \tag{31}$$

Therefore, $g(x) > 0$ and $f(x,y,z)$ is an increasing function. To make $(x,y,z) \geq 0$, the value of x must be not less than that of making $f(x,y,z) = 0$. Then, the minimal value of x can be goten on the condition that $f(x,y,z) = 0$.

The minimum value of x can be drawn from Theorem 3 when the relationship between n and k is $n/k \leq \sqrt{e}/(\sqrt{e}-1)$. For the given y and z, $f(x,y,z)$ is a function of a, so x can be set as $x = h(a)$. Substitute the value of x into (12) and then compute the derivative of a. Within the range of$a$, if the derivative is greater than 0, the value of a will be $\lceil \sqrt{k} \rceil + 1$, and at this time $z = [\sqrt{k}]$, $x = h(\lceil \sqrt{k} \rceil + 1)$, then the minimal E can be computed. When the derivative is less than 0, the value of a will be $\lceil k/2 \rceil$, and at this time $z = 2$, $x = h(\lceil k/2 \rceil)$. Therefore, the minimum of E can be gained. If it cannot make sure whether the derivative is greater than 0 or not, we can firstly compute the value of a which makes the derivative equal to 0. Substitute the value of a into (12), then we can get the minimum E.

If $y < z$, formulas (24) and (25) show us that the value of E increases with the increase of z and decreases with the increase of x. At the same time, the values of x and z are both decreased with the increase of y. That can be seen from (14) and (20). As a result, for the given y, to make E minimal, x should be maximal within its range, while z should be minimal within its range. When $y = a$ $(2 \leq a \leq [\sqrt{k}])$, $z = \lceil k/a \rceil$. And when the relationship between n and k is $n/k \leq \sqrt{e}/(\sqrt{e}-1)$, f will be an increasing function and the maximum of x will be $[n/a]$.

Then, formula (12) turns into

$$E(a) = \frac{[n/a]a^2 + [n/a]a + n\lceil k/a \rceil - [n/a]\lceil k/a \rceil a}{n + [n/a]}. \tag{32}$$

We can compute the derivative of a by the equality (32). If the derivative is greater than 0, set $a = 2$, and substitute the value of a into (32) afterwards, then the minimum of E can be gained. If the derivative is less than 0, set $a = [\sqrt{k}]$. Then substitute the value of a into (32). The minimum of E can be gained. If it cannot make sure whether the derivative is greater than 0 or not, we can compute the value of a firstly, which makes the derivative equal to 0. Then put the value of a into (32), we can get the minimum E.

The repair communication overhead of hybrid repair based on two-layer storage structure is $\lceil \sqrt{k} \rceil$ when $y = z$. If the repair communication overhead is greater than $\lceil \sqrt{k} \rceil$ in both cases, $y > z$ and $y < z$, the minimal repair communication overhead of hybrid repair is $\lceil \sqrt{k} \rceil$. If the repair communication overhead is smaller than $\lceil \sqrt{k} \rceil$ in any of the two cases: $y > z$ and $y < z$, the minimal repair communication overhead of hybrid repair is at most $\lceil \sqrt{k} \rceil$. In one word, the repair communication overhead of hybrid repair is at most $\lceil \sqrt{k} \rceil$. □

Theorem 4. *The hybrid repair based on two-layer storage structure can reduce the repair communication overhead to $o(1/\sqrt{k})$ times of the traditional data recovery algorithm.*

Proof. From the analysis mentioned above, it can be concluded that the repair communication overhead of hybrid repair based on two-layer storage structure is at most $\lceil \sqrt{k} \rceil$. Compared with the traditional method whose repair communication is k, the hybrid repair reduces the repair communication overhead to $\lceil \sqrt{k} \rceil / k \approx 1/\sqrt{k}$ times of the traditional method. □

4.2.2. Evaluation of the Repair Communication Overhead

Theorem 5. *If the relationship between n and k is $k+1 \leq n \leq k + \sqrt{k}$, it can make sure that the repair communication overhead of hybrid repair based on two-layer storage structure is lower than that of MSR.*

Proof. The repair communication overhead of hybrid repair is at most $\lceil \sqrt{k} \rceil$. While the repair over head of MSR is $d/(d-k+1)$, where d is the number of storage nodes that involved in data repair and $k \leq d \leq n-1$. Let $f(d) = d/(d-k+1)$, and from its derivative, we can know that $f(d)$ is a decreasing function, and the minimum of $f(d)$ can be gotten at $d = n-1$ and at this time $f(d)_{\min} = (n-1)/(n-k) = 1 + (k-1)/(n-k)$. With the condition $k+1 \leq n \leq k+\sqrt{k}$, we can gain $f(d)_{\min} = 1 + (k-1)/[\sqrt{k}] \geq \lceil \sqrt{k} \rceil$, so if $k+1 \leq n \leq k + [\sqrt{k}]$, the repair communication overhead of hybrid repair based on two-layer storage structure is lower than that of MSR. □

Theorem 6. *For the repair method based on group interference, if $p \geq \lceil \sqrt{k} \rceil$ (p is the number of storage nodes that a data group contains), its repair communication overhead is higher than that of hybrid repair based on two-layer storage structure.*

Proof. For the repair method based on group interference, its repair communication overhead is $p + (k-p)/q$ which is higher than p (q is the number of data pieces stored at a single storage node). That has been proved by [9]. If $p \geq \lceil \sqrt{k} \rceil$, the repair communication overhead of it is higher

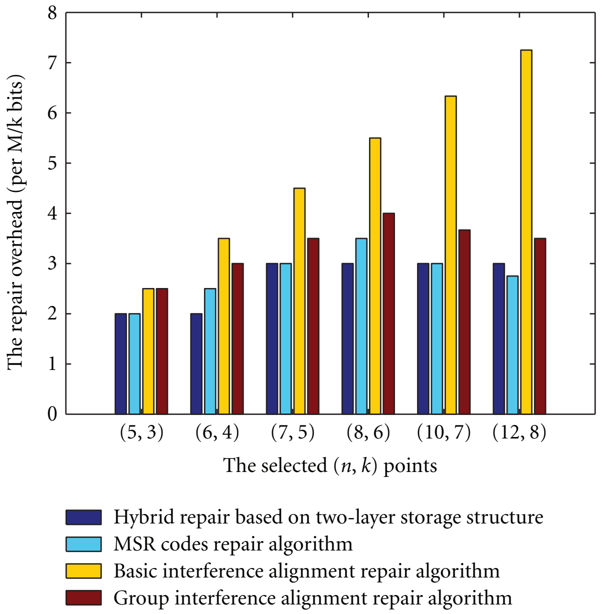

Figure 3: The repair communication overhead evaluation of the hybrid repair.

than $\lceil \sqrt{k} \rceil$. However, the repair communication overhead of hybrid repair based on two-layer storage structure is at most $\lceil \sqrt{k} \rceil$ which is known from the analysis above. Therefore, the theorem is proved. □

For the repair method based on basic interference alignment, since its repair communication overhead is $(qk - q + 1)/q = k-1+1/q > k-1$, the repair communication overhead of hybrid repair based on two-layer storage structure is lower.

4.2.3. Numeral Result. To verify the correctness of the conclusions, the numeral result is shown in Figure 3. The histogram in Figure 3 shows, respectively, the repair communication overhead of MSR repair, exact repair based on two-layer storage structure, repair based on basic interference alignment and group interference alignment, where (n, k) values are as follows: (5, 3), (6, 4), (7, 5), (8, 6), (10, 7), and (12, 8).

Comparing the conclusions of this paper with the numeral results displayed in Figure 3, it can be seen apparently that they are consistent.

5. Conclusion

This paper analyzes the tradeoff between storage overhead and repair communication overhead in the distributed data storage. We turn the flat storage structure into hierarchical storage structure and present a two-layer distributed data storage scheme to improve the repair communication overhead. Based on the two-layer data storage scheme, a data recovery method is proposed to decrease the repair communication overhead with sacrificing lower storage overhead. The proposed method has lower repair communication overhead than that of MSR, basic interference alignment and group interference alignment schemes. We prove this method reduces the repair communication overhead to $o(1/\sqrt{k})$ times. The proposed scheme is suitable for resource-constrained and node frequent failure distributed wireless sensor networks.

Acknowledgments

The authors gratefully acknowledge inspiring discussions with Weisong Shi and Xiaohong Jiang. This work was supported by the National Natural Science Foundation of China (61100153, 61172068) and the Aviation Science Foundation of China (2009198101, 2010ZC31001, 2010ZC31002, 20101981015, and 2011ZC31006).

References

[1] C. Suh and K. Ramchandran, "Exact-repair MDS codes for distributed storage using interference alignment," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT '10)*, pp. 161–165, Dublin, Ireland, June 2010.

[2] N. B. Shah, K. V. Rashmi, P. V. Kumar et al., "Interference alignment in regenerating codes for distributed storage: necessity and code constructions," *IEEE Transactions on Information Theory*, vol. 58, no. 4, pp. 2134–2158, 2012.

[3] H. Weatherspoon and J. D. Kubiatowicz, "Erasure coding versus replication: a quantitiative comparison," in *Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS '02)*, pp. 328–338, Cambridge, Mass, USA, March 2002.

[4] J. Kubiatowicz, D. Bindel, Y. Chen et al., "OceanStore: an architecture for global-scale persistent storage," in *Proceedings of the 9th Internatinal Conference Architectural Support for Programming Languages and Operating Systems (ASPLOS '00)*, pp. 190–201, Boston, Mass, USA, November 2000.

[5] I. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, 1960.

[6] S. Rhea, G. Wells, P. Eaton et al., "Maintenance-free global data storage," *IEEE Internet Computing*, vol. 5, no. 5, pp. 40–49, 2001.

[7] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A survey on network codes for distributed storage," *Proceedings of the IEEE*, vol. 99, no. 3, pp. 476–489, 2011.

[8] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4539–4551, 2010.

[9] Y. Wu and A. G. Dimakis, "Reducing repair traffic for erasure coding-based storage via interference alignment," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT '09)*, pp. 2276–2280, Seoul, Republic of Korea, July 2009.

[10] J. Li, S. Yang, X. Wang, X. Xue, and B. Li, "Tree-structured data regeneration with network coding in distributed storage systems," in *Proceedings of the 17th International Workshop on Quality of Service (IWQoS '09)*, pp. 1–5, Shanghai, China, July 2009.

[11] D. Cullina, A. G. Dimakis, and T. Ho, "Searching for minimum storage regenerating codes," in *Proceedings of the Allerton Conference on Control, Computing, and Communication*, 2009.

[12] K. V. Rashmi, N. B. Shah, P. Vijay Kumar, and K. Ramchandran, “Explicit construction of optimal exact regenerating codes for distributed storage,” in *Proceedings of the 47th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1243–1249, Urbana-Champaign, Ill, USA, October 2009.

[13] N. B. Shah, K. V. Rashmi, P. V. Kumar, and K. Ramchandran, “Explicit codes minimizing repair bandwidth for distributed storage,” in *Proceedings of the IEEE Information Theory Workshop (ITW ’10)*, pp. 1–5, Bangalore, India, January 2010.

[14] K. V. Rashmi, N. B. Shah, and P. V. Kumar, “Optimal exact-regenerating codes for distributed storage at the MSR and MBR points via a product-matrix construction,” *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5227–5239, 2011.

[15] P. R. Beesack, “Improvements of Stirling’s formula by elementary methods,” University of Beograd Publications, Elektrotehničkog fakulteta, Serija: Matematika i Fizika, 1969.