ICIC Express Letters
Volume **11**, Number **8**, August **2017**

ICIC International ©2017 ISSN 1881-803X
pp. 1359–1366

A SERIAL HYBRID STRATEGY FOR FILES IN DISTRIBUTED STORAGE SYSTEMS

Weibo Zhou[1,2], Yong Zhong[1] and Yang Wang[1,2]

[1]Chengdu Institute of Computer Applications
Chinese Academy of Sciences
No. 9, Sec. 4, Renmin South Road, Chengdu 610041, P. R. China

[2]University of Chinese Academy of Sciences
No. 19(A), Yuquan Road, Shijingshan District, Beijing 100049, P. R. China
findz@qq.com

Received January 2017; accepted April 2017

Abstract. *The rapid growth of the vast amount of data constantly increases the node scale in distributed storage systems, which leads to more and more frequent node failures. The reliability study on the distributed files has become the focus in the storage field of big data. Through the analysis on the studies about the replication strategy, erasure code strategy, and hybrid strategy, we propose a serial hybrid strategy and its reliability is modeled. The experimental comparison and analysis show that the proposed strategy has a better performance in the storage space utilization, reliability, data repair bandwidth, and time of file reading and writing.*

Keywords: Distributed files, Reliability modeling, Big data, Serial hybrid strategy

1. **Introduction.** Nowadays, data are in the trend of explosive growth and the data type also transforms from the structured data to the diversified data. All kinds of semi-structured and unstructured data have sprung up and the traditional method of data storage cannot solve such large-scale data. Distributed storage has become the main way to solve the mass data storage. It divides the data to be processed into several data blocks and stores them into several storage nodes which are connected through the network. Such a method represented typically by the Hadoop distributed file system [1] has the advantage of high throughput, high availability, and low cost. However, the rapid growth of the vast amount of data constantly increases the scale of nodes, and to save cost most of the storage nodes in the cluster use the equipment with low reliability, which inevitably leads to more and more frequent node failures. According to the statistics in the Hadoop cluster deployed by Facebook, this cluster is up to 3000 nodes and has node failure number up to over 20 nodes every day [2]. To improve the storage reliability, through the analysis about the replication strategy, erasure code strategy, and hybrid strategy, this research illustrates their respective advantages and disadvantages, and further proposes a serial hybrid strategy for distributed files.

The organization of this paper is as follows: related works are presented in Section 2; the proposed serial hybrid strategy is described in Section 3; test and analysis are shown in Section 4; Section 5 concludes this work.

2. **Related Work.** The current study on data reliability in the distributed storage systems primarily focuses on the replication strategy [3], the erasure code strategy [4], and a hybrid strategy which combines both the replication strategy and the erasure code strategy.

2.1. **Replication strategy.** The basic idea of the replication strategy is to copy the original data to obtain multiple replications of the data and dispersedly store each replication into different storage nodes. This is equivalent to carrying out multiple full backups of the original data and the whole storage system only needs to keep effective either the original data or at least one of the multiple replications.

The replication strategy is simple, practical, and can effectively improve the data reliability. Meanwhile the parallel access of multiple replications can improve the data access efficiency of the whole storage system and it can also achieve the effect of load balancing to some extent. However, the replication strategy has a high storage cost and its storage space linearly grows as the number of replications increases. Further, the pressure of the system network bandwidth is higher when data are being written.

The current study on the replication strategy mainly focuses on the replication coefficient [5], replication placement [6,7], replication consistency [8], replication repair [9], and so on.

2.2. **Erasure code strategy.** The erasure code strategy which assures the data of high reliability by increasing less redundancy originates from the communication field and is later introduced into the distributed storage system. Its basic idea is to divide the original file into m data blocks and encode them to obtain n ($n > m$) coded blocks. When some nodes are failed in the system, data can be restored by any k (usually $k = m$) coded blocks.

Compared with the replication strategy, the erasure code strategy greatly reduces the storage overhead and obtains high storage efficiency on the premise of guaranteeing the data reliability. However, this strategy generates large time overhead on encoding and decoding, which greatly reduces the data access efficiency. When restoring a block of data, it needs to transmit k blocks of data in the network, which causes heavy pressure on the network bandwidth.

The current study on the erasure code strategy mainly centers on the coding method [10], renewable code technology [11], and local repairable code technology [12].

2.3. **Hybrid strategy.** Rodrigues et al. [13] take quantitative comparison on the above two strategies and present their advantages and disadvantages, which provides a theoretical support for the parallel hybrid strategy. The DiskReduce [14] designed by the Parallel Data Lab at Carnegie Mellon University takes the asynchronous coding method, handles the hot data by the replication strategy to improve the system I/O throughput and the computing parallelism, and deals with the cold data by the erasure code strategy and deletes redundant replications to save storage space. However, this hybrid strategy is essentially a bundle of the replication strategy and the erasure code strategy and it does not consider the large data conversion overhead in these two strategies under the condition of more frequent conversion from the hot data to cold data.

In addition, the hybrid strategy is just to simply combine the replication strategy and the erasure code strategy, and uses the advantage of the erasure code strategy to rectify the disadvantage of the large storage overhead of the replication strategy. However, the hybrid strategy does not rectify the disadvantage that exits in the erasure code strategy and is only to reduce the probability of using the erasure code strategy. Once the erasure code strategy is needed to recover the file, it still needs to download the data that is nearly the size of the entire file in order to recover the whole file. In this case, the time delay of data access and the pressure of network bandwidth will get higher as the increase of the file. Further because the file size cannot be determined, during the coding procedure of the erasure code strategy, different coding redundancies should be used in order to obtain a certain degree of reliability, which increases the difficulty in using the erasure code strategy.

3. Serial Hybrid Strategy.

3.1. **Strategy description.** Normally in a distributed storage system if a file fails, it does not mean that all the Blocks in that file have failed but a Block or some Blocks have failed. We only need to recover these failed Blocks so that the entire file will be restored. Therefore, this paper presents a serial hybrid strategy, and this strategy changes the erasure objects and coding objects and carries out the erasure adjustment on the individual Blocks rather than originally the whole file. Since each Block size is fixed, a unified coding redundancy can be set at the time of coding. Because each Block size is smaller compared with the whole file, when a failed Block is being restored, only an effective code with the size of a Block is needed. This not only ensures the data reliability, but also greatly reduces the time delay of data access and the pressure of network bandwidth. Meanwhile according to user's interest, this strategy can also improve the reliability and access efficiency of a certain Block or a certain part of the Block, which makes different reliability and access efficiency regarding multi Blocks of the same file. Such difference can avoid the increase of file replications due to users' interest only in part of the data of a large file and thus the storage resources are saved and the system resource utilization is improved.

3.2. **Reliability modeling.** The proposed strategy is based on the replication strategy, and regarding each Block (called BigBlock afterwards) in the replication strategy, it is divided into several SmallBlocks. SmallBlocks are encoded and BigBlocks are under erasure. Under the proposed strategy, it is assured that the multiple replications from each BigBlock are independent and irrelevant to certain number of SmallBlocks. The reliability model of the proposed strategy is shown in Figure 1.

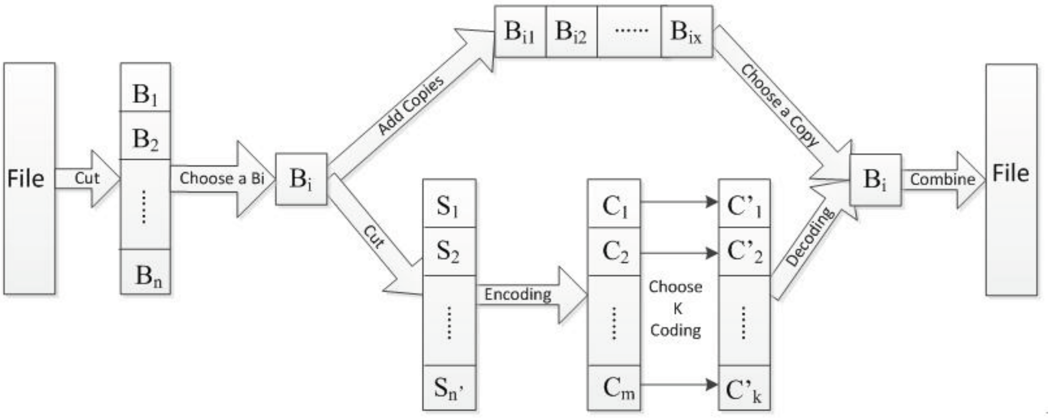

FIGURE 1. Reliability model of serial hybrid strategy

And its reliability mathematical model can be expressed as

$$Fr = \prod_{i=1}^{n} Br_i \tag{1}$$

where Br_i denotes the reliability of the ith Block and

$$Br_i = 1 - (1 - Br_{Rep})(1 - Br_{EC}) \tag{2}$$

where Br_{Rep} denotes the Block reliability of replication strategy and Br_{EC} denotes the Block reliability of erasure code strategy.

In the serial hybrid strategy model, the entire file is divided into n Blocks. Regarding any data block B_i, its copies can be increased by means of Add Copies; meanwhile the data block can be further divided into several smaller blocks SmallBlock to construct the erasure code system for B_i. When the data needs to be restored, B_i can be obtained by retrieving any copy or by decoding the erasure code. Finally all the blocks of data are incorporated to get the entire file. As can be seen from Figure 1, the copies in the serial hybrid strategy are based on the data block and the copy numbers can be set according to the heat degree of the data block or the user's interest. Different data blocks of a file may have different copy numbers so that storing large file will have better flexibility. In the erasure code strategy, each Block is encoded and thus a failure Block can be recovered by only an effective code of a Block size, which can not only guarantee the reliability and greatly reduce the recovery cost when faced with a large file.

In addition, it can be noticed that the reliability mathematical model of the proposed strategy is consistent with the replication strategy. However, the difference is that all Br_is are fixed and consistent in the replication strategy while every Br_i in the proposed strategy can guarantee the needed access efficiency and reliability according to Br_{Rep} and Br_{EC} (Br_{Rep} and Br_{EC} are used to respectively guarantee the access efficiency and reliability).

4. Test and Analysis.

4.1. Test setting. In order to verify the advantage of the proposed strategy, a Hadoop cluster of 20 nodes is built by a virtual machine to carry out the test. The cluster consists of 2 NameNodes and 18 DataNodes and each node has 512MB memory and 8GB hard disk capacity. The operating system is CentOS 6.5 and the version of Hadoop is 2.6.0.

To further simplify the calculation, the BigBlock size is set as 256MB, the replication strategy (Rep) [6] uses three replications, the erasure code strategy (EC) uses an RS$(2n, n)$ [15] coding scheme, the parallel hybrid strategy (PH) uses a fixed replication and an RS$(2n, n)$ parallel coding scheme, and the proposed strategy (SH) uses a fixed strategy and an RS$(16, 8)$ serial coding scheme.

4.2. Test 1: storage capacity and reliability. The storage capacity test and the reliability test of each strategy are presented in Figure 2 and Figure 3 respectively. In this test, under the condition of similar storage consumption of the four kinds of storage strategies, the decreasing amplitude about the reliability of the replication strategy (Rep) is higher as the file increases. However, the erasure code strategy (EC) and the parallel hybrid strategy (PH) have reliability enhancement as the file increases. The decreasing amplitude about the reliability is relatively low in the proposed strategy (SH) as the file

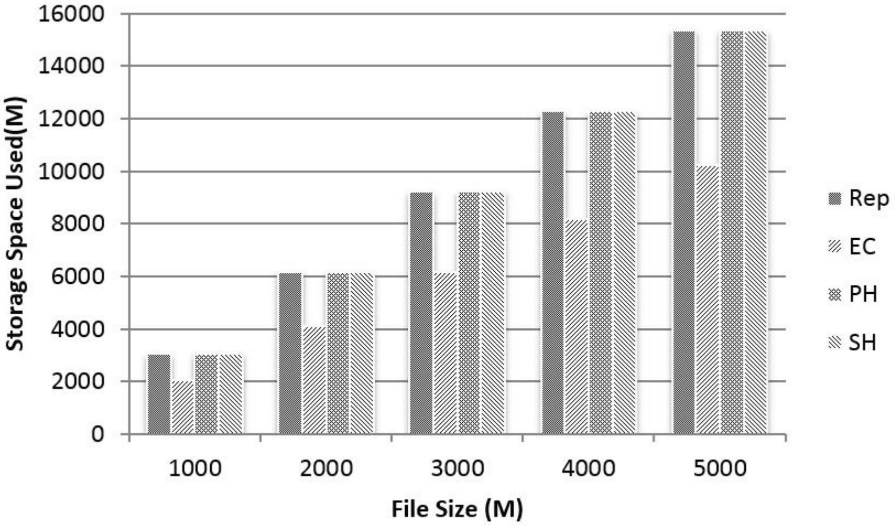

FIGURE 2. Storage space usage of each strategy

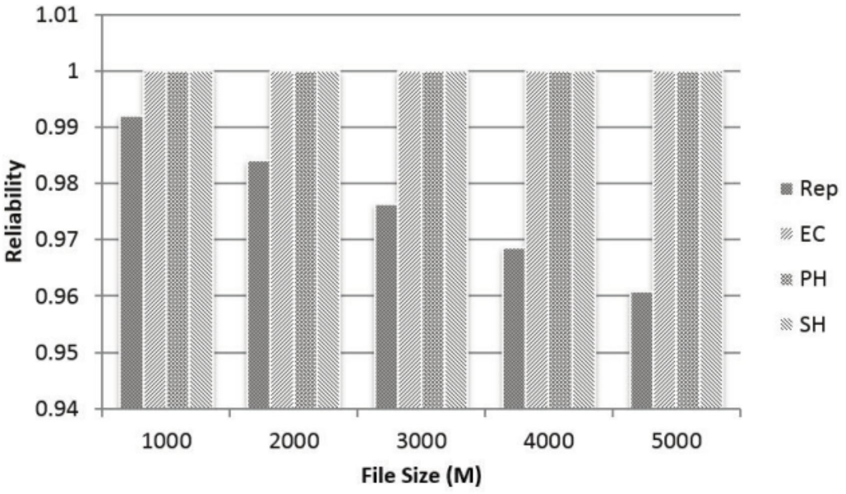

Figure 3. Reliability of each strategy

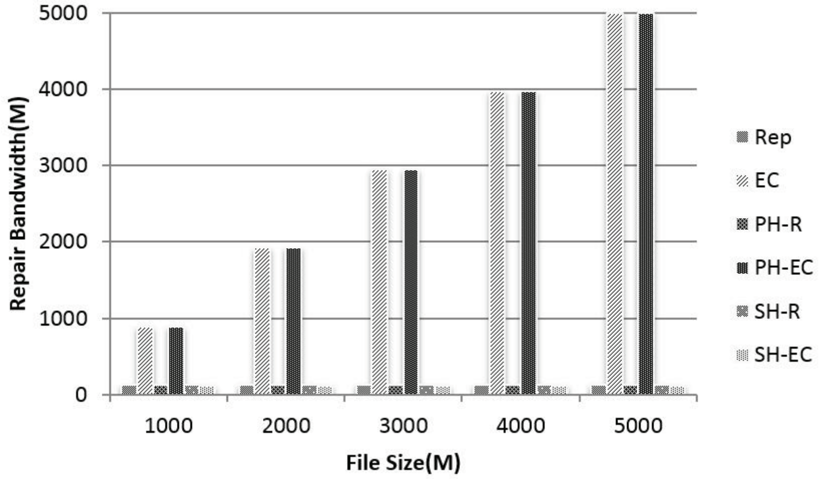

Figure 4. Data repair bandwidth of each strategy

increases, and the reliability can be improved through increasing the encoding redundancy, which can meet the requirement of high reliability. Thus the replication strategy achieves the high reliability requirement with high cost and the erasure code strategy, the parallel hybrid strategy, and the proposed strategy can obtain the high reliability requirement with low cost.

4.3. **Test 2: data repair bandwidth.** The data repair bandwidth of each strategy is shown in Figure 4. In this test, the repair bandwidth of the replication strategy (Rep), the replication strategy of parallel hybrid strategy (PH-R), the replication strategy of the proposed strategy (SH-R), and the erasure code strategy of the proposed strategy (SH-EC) is similar to the size of BigBlock, but the repair bandwidth of erasure code strategy and the erasure code strategy of parallel hybrid strategy (PH-EC) is similar to the size of File. Thus when repairing the data, the strategy that uses the file erasure will bring huge bandwidth pressure on the cluster and this pressure becomes higher as the file increases.

4.4. **Test 3: file writing and reading.** Because the erasure code strategy (EC) and parallel hybrid strategy (PH) carry out the erasure on the whole file, their efficiency of file writing and reading is significantly lower than the replication strategy. Thus the

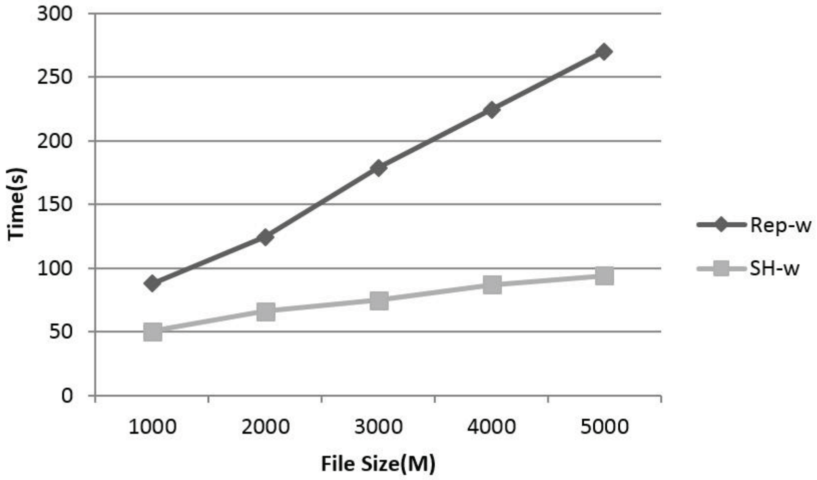

Figure 5. File writing time comparison

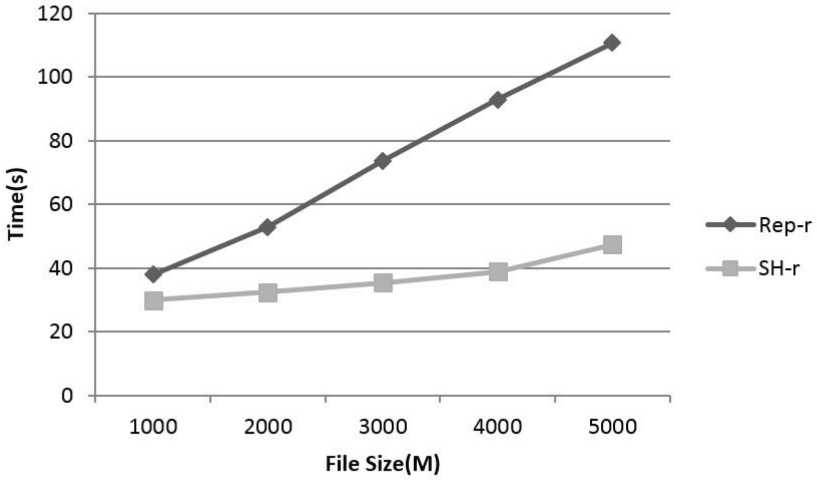

Figure 6. File reading time comparison

replication strategy (Rep) and the proposed strategy (SH) are compared in this section and results are presented in Figure 5 and Figure 6.

In this test, through the file reading and the file writing comparisons, the proposed strategy (SH) has more advantages than the replication strategy. In the process of file writing, once the proposed strategy (SH) finishes writing a replication of a Block, it immediately informs the management node about the completion of writing the data of the replication strategy, and the Client can continue to apply for a new Block; meanwhile, the coding and distributing task in the erasure code (SH-EC) of the proposed strategy can concurrently be finished in the cluster, and thus the total time consumption is much lower than the replication strategy. In the process of file reading, when the proposed strategy (SH) recovers the data, it can concurrently transmit SmallBlock from multi storage nodes to the destination storage node, and its cost of decoding time is relatively lower than that of the network transmission time, and thus the reading file efficiency is superior to that of the replication strategy (Rep). We can conclude that the efficiency of file reading and writing in the proposed strategy is much superior to that in the replication strategy (Rep).

From the above tests, the superiority of the proposed serial hybrid strategy over erasure code strategy and hybrid strategy is as follows. The copy strategy cannot solve the problem of low cost under high reliability requirement. The erasure code strategy cannot deal with the bandwidth of data recovery. Although the hybrid strategy reduces the cost of reliability by the erasure code and uses copy strategy with top priority to restore data so as to lower the bandwidth of data recovering, the strategy does not essentially solve the problem of the recovery bandwidth of erasure code. And once the copy fails, when the data in the erasure code is restored, its data recovery bandwidth is the same as the erasure code strategy. When using data blocks as unit to duplicate, decode, and encode copies, the proposed serial hybrid strategy can fundamentally solve the problem of the recovery bandwidth of erasure code by taking the advantages and avoiding the disadvantages of both the copy strategy and the erasure code strategy. In addition, regarding the frequently accessed and user's interested data blocks, more copy numbers can be set in the proposed method to enhance the data accessing efficiency.

5. **Conclusions.** The reliability study on the distributed files has become the focus in the storage field of big data. Through the analysis on the replication strategy, erasure code strategy, and hybrid strategy, this paper illustrates their respective advantages and disadvantages, proposes a serial hybrid strategy, and carries out the mathematical modeling for its reliability. The two parts of the proposed strategy are respectively used to ensure its high data access efficiency and to achieve its high reliability under the condition of low redundancy. In addition, the proposed strategy can also increase the replications of the hot data in the cold file and guarantee the hot data's higher data access efficiency. Through the experimental comparison and analysis, the proposed strategy shows a good performance in the storage space utilization, reliability, data repair bandwidth, and the time delay of file reading and writing. The innovation on the encoding theory and the improvement of encoding and decoding granularity and speed will be our future research.

Acknowledgment. This work is partially supported by the Sci.&Tec. Support Plan of Sichuan Province under the grant No. 2014GZ0013 and the Sci.&Tec. Innovation Talent Project of Sichuan Province under the grant No. 2014-063. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] J. R. Chen and J. J. Le, Reviewing the big data solution based on Hadoop ecosystem, *Computer Engineering & Science*, vol.35, no.10, pp.25-35, 2013.

[2] M. Sathiamoorthy et al., XORing elephants: Novel erasure codes for big data, *VLDB Endowment*, vol.6, no.5, pp.325-336, 2013.

[3] T. T. Liu et al., Multiple-replicas management in the cloud environment, *Journal of Computer Research and Development*, vol.48, no.3, pp.254-260, 2011.

[4] X. H. Luo and J. W. Shu, Summary of research for erasure code in storage system, *Journal of Computer Research and Development*, vol.49, no.1, pp.1-11, 2012.

[5] L. Shi et al., Feedback mechanism based prediction method of dynamic replicas number, *Journal of System Simulation*, no.1, pp.193-199, 2011.

[6] W. W. Lin, An improved data placement strategy for Hadoop, *Journal of South China University of Technology (Natural Science Edition)*, vol.40, no.1, pp.152-158, 2012.

[7] P. Luo and X. Gong, Research and improvement of data placement strategy for HDFS, *Computer Engineering and Design*, vol.35, no.4, pp.1127-1131, 2014.

[8] Y. M. Tian et al., Replica consistency detection in distributed file system, *Journal of Computer Research and Development*, vol.49, pp.276-280, 2012.

[9] F. Ren et al., Analysis of data recovery strategy in large scale distributed storage system, *China Internet*, vol.2, pp.7-12, 2013.

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

[11] A. G. Dimakis et al., Network coding for distributed storage systems, *IEEE Trans. Information Theory*, vol.56, no.9, pp.2000-2008, 2008.

[12] D. S. Papailiopoulos and A. G. Dimakis, Locally repairable codes, *IEEE International Symposium on Information Theory*, pp.2771-2775, 2012.

[13] R. Rodrigues and B. Liskov, High availability in DHTs: Erasure coding vs. replication, *International Conference on Peer-To-Peer Systems*, pp.226-239, 2005.

[14] B. Fan et al., DiskReduce: RAID for data-intensive scalable computing, *Workshop on Petascale Data Storage*, pp.6-10, 2009.

[15] D. Y. Tao, X. H. He and Z. H. Wu, The implementation of RS encoding and decoding, *Journal of Sichuan University (Natural Science Edition)*, vol.34, no.6, pp.868-872, 2000.