Single node failure repair method for distributed storage system based on genetic algorithm and MSR code

Xiaoyue Qiu 1, *, Tiantian Wang 1, Enze Lin 2, Zongpeng Lu 3
1 School of Information Engineering, CHANG’AN UNIVERSITY, Xian, 710000, China
2 College of Meterials Science and Engineering, Hunan University, Changsha 410000, China
3 School of Electronic and Control Engineering, Chang’an University, Xi’an 710064, China

*Corresponding author e-mail: 2019902701@chd.edu.cn

Abstract. In view of the minimum storage regeneration (MSR) single fault node repair plan, design a kind of new single repair fault node topology. Consider link bandwidth available and the node processing power of heterogeneity, the repair delay and repair traffic node of combinatorial optimization. Topology can be converted to Steiner tree model with constraints, and design the corresponding hybrid genetic algorithm to solve the approximate global optimal solution. Finally, a compromise between the two optimization objectives is achieved. Simulation results show that under the same MSR code size, the repair delay of this topology is only 60%-80% of that of the traditional tree topology and 30%-40% of that of the traditional star topology. While the repair traffic for this topology is 9% to 65% higher than for the traditional star topology, the repair traffic is only 40% to 55% higher than for the traditional tree topology.

Keywords: Distributed storage system, single node repair, MSR code, genetic algorithm.

1. Introduction
In the current era of cloud computing, global traffic is growing rapidly, and the amount of data in the Internet, the Internet of Things, mobile terminals, security monitoring, finance and other fields shows a "blowout" growth trend. Stored in the cloud server data in the growth rate and even beyond Moore's law, cloud storage system become one of the key component of cloud computing [1]. The rapid growth of data poses more severe challenges to the performance and scalability of storage systems. The traditional storage system uses a centralized method to store data, which makes the security and reliability of data cannot be guaranteed, and can not meet the needs of big data application [2]. Distributed storage system with its huge storage potential, the advantages of high reliability and easy extensibility and become the key of the large data storage system was applied to reduce the storage load, dredge network congestion, and other areas of the business, and its high throughput and high availability become the main in the cloud storage system. In order to deal with mass data storage requirements, size of the storage system is often very large, generally contain ranging from several thousand to tens of
thousands of storage nodes. As early as 2014, the number of nodes in a single cluster of Chinese Internet company Baidu exceeded 10,000 [2]. In the past two years, the number of nodes managed and scheduled by VStation, Tencent Cloud’s distributed scheduling system, can reach 100000. However, a large number of cluster nodes often produce such as power supply damage, system maintenance and network interruption caused frequent node failures. According to the statistical data of some large distributed storage systems, about 2% of nodes fail every day on average [3]. Thus it can be seen that the global cloud server failure leads to the loss of data from time to time. Obviously, reliable to store data and effective repair failure data of fault tolerant technology for distributed storage system is of great importance is self-evident. Therefore, it is very important to timely and effectively repair the failed nodes to ensure that the data can be read and downloaded normally.

When node failure occurs in the storage system with erasure code redundancy strategy, the repair topology constructed by the traditional single failure node repair scheme can be divided into two types: star type and tree type. The star repair topology directly connects the new node with many data providing nodes to download the data needed to reconstruct the failed node. Star topology, however, may not be able to effectively avoid the path of the low bandwidth available in network topology, so its repair delay longer [3]. Fix another tree topology structure of a new node as root repair tree. Tree topology while make full use of the available bandwidth higher link in the network topology, but has lower repair delay. However, due to the increase in data transmission path in the process of repair, can lead to additional repair flow in the storage network topology, lead to the entire network load has increased dramatically. This defect can lead to other network problems that ultimately affect the overall performance of the system. Traditional multi-fault node repair schemes are mainly divided into centralized and decentralized. In the centralized repair scheme, all repair traffic in the system is concentrated in the storage management node, and the storage management node is also responsible for the repair operation of all failed nodes. Therefore, all the processing capacity of a link node itself and its connection will become the bottleneck of the whole repair operation, lead to a node failure cases of repair efficiency is low. However, most of the current researches on distributed repair schemes focus on the optimization of the erasure code encoding mechanism to make it more in line with the needs of multi-node repair, so as to reduce the repair traffic. Moreover, a repair tree with the same number of fault nodes and disjoint edges is constructed in the system to complete the repair of each fault node, so as to achieve the maximum utilization of network topology bandwidth [4]. However, compared with the ordinary erasure codes, the modified coding mechanism is more complex, which will bring more additional storage and coding and decoding costs to the system. The repair topology of multiple edge disjoint repair trees may be different due to the construction order of repair trees, which leads to the first constructed repair tree monopolizing the link with excellent available bandwidth, When there are too many failed nodes, the repair rate of the whole topology may not be optimal.

Aiming at the defects of traditional repair process in distributed storage system, a single fault node repair topology based on minimum storage regenerative code environment was proposed and solved by hybrid genetic algorithm. In order to improve the repair rate of single fault node and reduce the traffic cost in the repair process as much as possible.

2. Related work
At present, study of erasure codes redundancy strategy literature mainly concentrated in the repair process and to improve the coding mechanism of topology optimization. In the field of the optimal topological building repair, according to the number of nodes in the topology can repair, it can be divided into two items: a single repair fault node topology optimization and repair fault node topology optimization.

The first mock exam node topology optimization is proposed. Huang and others [5] propose a pipeline repair scheme. The data can be transferred to each node in one to one mode and integrated. The network traffic overhead is also reduced while the efficiency of the repair process is reduced. However, the processing delay of each node in the cluster is not considered. If the cluster scale is large, it is difficult to guarantee the security and stability of the topology. By analyzing the heterogeneity of available
bandwidth in cluster, Jia et al. [6] designed a routing algorithm to optimize the process of providing repair data from surviving nodes to new nodes. G et al. [7] proposed the selection criteria for the topology of data supply nodes and trees in the restoration process for redundant regenerated code scenes, and analyzed their advantages. Wan et al. [8] comprehensively considered the impact of the actual network topology on the repair process, and proposed STNR algorithm to integrate the repair traffic in the intermediate nodes. STNR constructed the optimal repair topology by combining the minimum cost path heuristic algorithm and ant colony algorithm, which can improve the repair rate and reduce the repair traffic in the cluster, however, its coding is still the most basic RS code, and there is no further optimization in the field of improved coding method. Zhang et al. [9] took into account the actual network topology and adopted FPGA instead of switch to integrate and forward the repair traffic, and considered the heterogeneity of the processing capacity of the nodes when selecting the repair nodes to optimize the repair delay and traffic. However, due to the FPGA is expensive, it is difficult to application in a real environment.

Since professor Holland of the University of Michigan formally introduced the monograph of genetic algorithm (GA) in 1975, GA has achieved excellent results in various fields. At present, there are a lot of classical algorithms based on greedy mechanism for unconstrained minimum spanning tree problem, which can converge to the optimal solution with excellent time cost [10], such as kruskalst [11], primmst [12]. However, the problems to be solved in practical engineering and scientific research usually can not be described by a simple minimum spanning tree, but a variety of different constraints will be imposed, such as the degree constraint of limiting the number of associated edges of nodes in the tree [13]; The number of leaf nodes is limited by the leaf constraint [14].

In the aspect of coding mechanism improvement: Wang et al. [15] for regenerating traditional minimum bandwidth (MBR) code defects, according to the ideas of the different data and repeated cycle VFR code is constructed. Guo [16] based on sparse random matrix of exercise such as the high probability of full rank characteristics of sparse random erase error correcting code is proposed. The local reconfiguration coding (GRC) designed by Microsoft team on the basis of RS code [17] can reduce the repair flow of failed node repair process at the cost of reducing storage efficiency. Dimakis et al. [18] improved the basic MDS code to get the regeneration code. The regeneration code repair process reduces the repair traffic in the whole network by downloading the preprocessed data from more surviving nodes. Jiekak S. et al. [19] proved that the repair traffic of regenerative codes in the process of repairing failed nodes is only 16% of RS codes with the same disaster tolerance level.

In this paper, the star topology proposed in this paper has lower repair speed, but the repair flow in the topology is small. Tree topology has good repair rate, but in the topology repair flow is larger. By combining the advantages of the two topologies and adding the influence of heterogeneous node processing capacity on the repair process, a new three-layer repair topology is designed to realize the trade-off of repair rate and repair traffic in the overall repair process.

3. Method

3.1. Problem description of single failure node repair process

Generally, there are three types of repair modes: accurate repair, function repair and accurate repair of system node data by comparing the data stored by the original node and the data obtained by repairing the failed node. The data recovered by the new node may not be the same as the data stored by the failed node. As long as the repair mode can guarantee the same reliability is called functional repair. The random network coding is used to repair the nodes as function repair. It is necessary to carry out the operation in a large finite field, and the calculation complexity is high. However, accurate repair requires that the data of the failed node can be recovered accurately. Then the repair and reconstruction rules of the data will no longer be updated after the node repair, so the cost is relatively small and more meaningful. Only the accurate repair of system node data requires accurate repair of system nodes, which can ensure that there will always be system parts in the storage system after node repair, that is, raw data without coding. When repairing the failed nodes, each node only sends one coding block, which
can achieve the optimal value of the minimum storage regeneration code repair bandwidth. However, the gradual construction of code block algorithm needs to read n-1 nodes when repairing the failed nodes, so it will bring large disk reading overhead. Compared with erasure codes, the minimum storage regeneration (MSR) codes achieve the optimal storage cost and greatly reduce the bandwidth cost of a single failed node.

3.2. Optimization objective repair model

In the minimum memory regeneration (MSR) code, when using the gradual construction coding block algorithm to repair a single failed node, the remaining n-1 surviving nodes participate in the repair, and each node only sends one coding block, in which the size of each coding block is M/K (n – k), so the single node repair bandwidth of the gradual construction coding block algorithm is (n – 1) * M/k (n – k). The disk reading cost is n – 1. It can be seen that the repair bandwidth of the algorithm depends on the disk reading cost. The smaller the disk reading cost is, the smaller the repair bandwidth is.

![Three layer structure of repair topology](image)

*Figure 1. Three layer structure of repair topology*

Top with the new node as a root node, to guarantee the repair amount of data required for a single fault node; Intermediate node layer contains a total of n - d - 1 intermediate node, for sending the provider node repair data integration to the respective subtree, in order to reduce the topology of the repair flow; Provider layer contains D a provider node, the node is directly connected to the intermediate node or the root node, the data needed to provide repair fault node.

This topology has the following three advantages. First of all, using the repair tree topology, avoid storage network topology of the link is not good, so as to reduce the repair delay; Second, to join in the topology is similar to the middle of the star topology repair node aggregation mechanism, as far as possible to reduce additional repair operations; Finally, the depth of the repair tree topology is only three layers, the largest extent, ensure the repair topology of robustness.

In order to find design repair the global optimal solution of the topology, the right ZhangCai using undirected connected graph G = (V, e, w) with MSR code redundancy mechanism of the network topology of distributed storage system is simplified. V={V1, V2... vn} said set of N nodes in the system; E={E1, E2... , EM}, where Es (1≤ s≤ m) = (VI, VJ) represents the link between node VI and VJ in the topology; W = {(omega 6, VJ) | 1 I n, 1 or less or less J n} or less or less instructions available bandwidth on the link. When VI=VJ, it represents the processing ability of VI node processed.

\[
T_{repair} = \min \left( c_{time} \max \left( \frac{\beta}{\omega_{(v_i,v_j)}} + \alpha \cdot \frac{D_i}{\text{process}_i} + \alpha \cdot \frac{D_j}{\text{process}_j} \left| (v_i, v_j) \in E \right| \right) + \frac{c_{traffic}}{\sum_{mid=1}^{n-d-1} \text{degree}_{\text{mid}} - 2} \right)
\]

s.t. \( \text{degree}_{\text{newcomer}} \geq d - k + 1 \)
3.3. Solution method
How to find the repair topology in the storage system network. The tree repair topology is transformed from the traditional minimum spanning tree problem to the Steiner tree problem with constraints after adding intermediate nodes to achieve the aggregation of repair business and the degree constraint of the root node of the topology. The steps of the genetic algorithm are as follows:

Step 1: initialization parameter initialization population size parameters fill size, maximum generation number g mutation and crossover probability of PC, PM;

Step 2: (population initialization) Repeat the filling size algorithm once to get the initial population P (0), and the filling size is, t=0;

Step 3: (Crossover Operation) Select two individuals randomly from the current population P (T) and execute Algorithm 2 with probability PC to generate hybrid offspring, otherwise the parent individual will directly add hybrid offspring. Repeat crossover operation until hybrids is equal to the number of population size, and after all the hybrid composed of all the individual individual set records for O1;

Step 4: Mutation operation) Select an individual randomly from the current population P (T) and execute algorithm 3 using probability PM to generate mutant offspring, otherwise the parent individual will be directly added to the mutant offspring. Mutation was repeated until the number of offspring was equal to the population size, which was recorded as O2;

Step 5: (local search) Algorithm 4 continuously performed union oxygen 1 union O2 for each individual in set P (T), and performed local search to obtain three subpopulations of equal size, L1, L2, and L3.

Step 6: (select) computing set P (T) all the individual fitness 1 \bigcup \bigcup oxygen in oxygen levels \bigcup L1 \bigcup L2 \bigcup L3 according to the formula (11), implement the algorithm 5, select the with the size of the individual as the next generation of population P (T + 1), T = T + 1;

Step 7: (Termination) When the current evolutionary generation number T reaches the maximum evolutionary generation number G, the individual with the lowest fitness function value in the current population is the approximate solution of the global optimal solution, which is output as the optimal tree repair topology. Otherwise, skip step 3 and continue.

4. Experiment

4.1. Experimental details
In the experiment of this paper, pycharm2019.1.3 and python3.7 are used to write the simulation program of genetic algorithm, and the comparative experiment is performed in the hardware environment of Intel (R) core (TM) i5-7450h CPU @ 2.60GHz, 16G RM and 1T SSD. This experiment simulates the repair process of a single fault node in distributed storage.

The experimental scenario is as follows: the space of the data M is 1GB, and the number of storage nodes in the distributed storage system are 10, 20, 40 and 80 respectively. The internal network topology adopts undirected complete graph and MSR code redundancy strategy. The coding parameter triples correspond to (10, 6, 8), (20, 12, 16), (40, 24, 32), (80, 48, 64) respectively according to the number and scale of nodes in the system.

| Bandwidth B (KB/s) | Percent (%) |
|-------------------|-------------|
| B<30              | 2.8         |
| 30≤ B<200         | 11.2        |
| 200≤ B<500        | 26.4        |
| 500≤ B<1000       | 55.9        |
| 1000≤ B           | 3.7         |
4.2. Experimental scheme and result analysis

Experiment 1 (comparison of repair delay): in the above experimental scenario, the repair delay of failure node in system is compared between star repair topology using regenerative code, the tree repair topology using primmst minimum spanning tree algorithm using MSR code and the global optimal repair topology solved by genetic algorithm in this chapter.

Figure 2. Repair delay of three repair topologies in different scenarios

The results of experiment 1 as shown in figure 2, this chapter demonstrates that the proposed genetic algorithm to construct global optimal repair topology in four different sizes of clusters with the optimal repair delay. Due to the traditional topology star topology can choose the link was limited by its scope, even with greedy mechanism to search the local optimum solution, when choosing repair topological link still can't avoid some of the available bandwidth with weak link, cause major repair delay. The repair delay of the three-layer repair topology constructed by this algorithm is 9%-31% less than that of the tree repair topology. This is because the processing capacity of storage nodes is heterogeneous in the experimental scenario, which makes the PrimmST algorithm based on greedy mechanism unable to find the global optimal solution.

Experiment 2 (comparison of algorithm parameters): this experiment compares and analyzes the relationship between the repair delay of single failure node and the parameters $C_{\text{Time}}$ and $C_{\text{traffic}}$ in the three-tier repair topology designed in this paper under different storage sizes.

Table 2. Effects of four kinds of $(C_{\text{Time}}, C_{\text{traffic}})$ on MSR code repair delay in different scenarios

| $(C_{\text{time}}, C_{\text{traffic}})$ | (0.25,0.75) | (0.5,0.5) | (0.75,0.25) | (1,0) |
|--------------------------------------|------------|-----------|-------------|-----|
| (10, 6, 8)                          | 6.42       | 6.42      | 6.42        | 6.42|
| (20, 12, 16)                        | 1.35       | 1.35      | 1.35        | 1.35|
| (40, 24, 32)                        | 0.83       | 0.83      | 0.82        | 0.81|
| (80, 48, 64)                        | 0.30       | 0.29      | 0.28        | 0.27|

The results of Experiment 2 are shown in Table 2. Although the repair delay does not change with the change of $C_{\text{time}}$ (traffic) in the storage scale of (10, 6, 8) and (20, 12, 16), the repair delay decreases with the increase of $C_{\text{time}}$ in the storage scale of (40, 24, 32) and (80, 48, 64). This is because in the fitness function of the genetic algorithm designed in this paper, the larger the parameter $C_{\text{time}}$ is, the more attention the algorithm pays to reducing the repair delay, but it will increase the repair traffic in repair topology.

5. Conclusion

For the distributed storage system is widely used in repair the fault node data, not only can consider cluster topology link bandwidth heterogeneity are available, and also analyzed the processing power of all nodes heterogeneity and repair the flow state of the network effects on cluster itself. Redundant
mechanism in view of the MSR yards single node failure in distributed storage system, considering the heterogeneous node processing power and bandwidth available for node repair rate, the influence of and the influence of repair topology to repair flow, established a three layer model for repair. An optimal maintenance topology construction method based on genetic algorithm was designed to optimize the maintenance rate and maintenance flow at the same time. Through simulation experiment, comparing the structure of the method to repair topology with repair traditional repair star topology and tree topology performance difference, discussed the influence of the parameters on the results. The simulation results show that the algorithm is constructed to repair topology structure can improve the efficiency of nodes to repair, repair flow optimization.

References
[1] Shuang Zhang, Xingdong Feng. Distributed identification of heterogeneous treatment effects [J]. Computational Statistics, 2021 (prepublish).
[2] Ravi Kiran Raman, Lav Raj Varshney. Dynamic distributed storage for scaling blockchain [P].: US10965448, 2021-03-30.
[3] Yang Liu. DC/DC converter for distributed storage and solar systems [P].: US10951040, 2021-03-16.
[4] Dhanashankar Venkatesan, Jagadish Kumar Mukku, Ripulkumar Hemantbhai Patel. Creating Snapshots of a Storage Volume in a Distributed Storage System [P].: US2021073079, 2021-03-11.
[5] Eric EILEBRECHT, Karthick KRISHNA MOORTHY, Maneesh SAH. QUERYING DATA IN A DISTRIBUTED STORAGE SYSTEM [P].: US2020349166, 2020-11-05.
[6] Eric EILEBRECHT, Karthick KRISHNA MOORTHY, Maneesh SAH. QUERYING DATA IN A DISTRIBUTED STORAGE SYSTEM [P].: US2020349166, 2020-11-05.
[7] David Lane Smith. DISTRIBUTED STORAGE SYSTEM FOR LONG TERM DATA STORAGE [P].: US2020341689, 2020-10-29.
[8] Mohamed Sohail, Junping Zhao. INTELLIGENT DATA STORAGE AND MANAGEMENT FOR CLOUD COMPUTING [P].: US2020344299, 2020-10-29.
[9] Mohamed Sohail, Junping Zhao. INTELLIGENT DATA STORAGE AND MANAGEMENT FOR CLOUD COMPUTING [P].: US2020344299, 2020-10-29.
[10] David Lane Smith. DISTRIBUTED STORAGE SYSTEM FOR LONG TERM DATA STORAGE [P].: US2020341689, 2020-10-29.
[11] Bo Zhao, Junzhi Ren, Jian Chen, Da Lin, Ruwen Qin. Tri-level robust planning-operation co-optimization of distributed energy storage in distribution networks with high PV penetration [J]. Applied Energy, 2020, 279.
[12] Shravan Kumar Vallala, Dhanashankar Venkatesan. Window-based priority tagging of IOPs in a distributed storage system [P].: US10782887, 2020-09-22.
[13] Asfand Yar Ali, Abdul Basit, Tanvir Ahmad, Affaq Qamar, Javed Iqbal. Optimizing coordinated control of distributed energy storage system in microgrid to improve battery life [J]. Computers and Electrical Engineering, 2020, 86.
[14] Energy; Researchers at University of Cassino and Southern Lazio Have Reported New Data on Energy (Planning of Distributed Energy Storage Systems In Mu Grids Accounting for Voltage Dips) [J]. Energy Weekly News, 2020.
[15] Guidehouse Insights Report: Distributed Energy Storage Capacity Deployments for Top 10 US Markets Expected to Increase [J]. Manufacturing Close - Up, 2020.
[16] Zhiguo Du, Du Zhiguo, Wu Zhihui, Wen Bin, Xiao Kehui, Su Ruijia. Traceability of animal products based on a blockchain consensus mechanism [J]. IOP Conference Series: Earth and Environmental Science, 2020, 559 (1).
[17] Yuyu Jiang. The Application of Blockchain Technology in Management Accounting [J]. International Journal of Social Science and Education Research, 2020, 3(8).
[18] Google LLC; Patent Issued for Distributing Data On Distributed Storage Systems (USPTO
10,678,647) [J]. Internet Weekly News, 2020.
[19] Google LLC; Patent Issued for Distributing Data On Distributed Storage Systems (USPTO 10,678,647) [J]. Internet Weekly News, 2020.
[20] HashBlock Inc.; Patent Application Titled "Systems And Methods For Decentralized Distributed Storage Using Blockchain" Published Online (USPTO 20200136809) [J]. Information Technology Newsweekly, 2020.
[21] EMC IP Holding Company LLC; "Secure Distributed Storage Of Encryption Keys" in Patent Application Approval Process (USPTO 20200136822) [J]. Information Technology Newsweekly, 2020.
[22] Ullah Shafaat,Khan Laiq,Badar Rabiah,Ullah Ameen,Karam Fazal Wahab,Khan Zain Ahmad,Rehman Atiq Ur. Consensus based SoC trajectory tracking control design for economic-dispatched distributed battery energy storage system. [J]. PloS one, 2020, 15(5).
[23] Gang Wang,Min-Yao Niu,Fang-Wei Fu. Constructions of (r,t)-LRC Based on Totally Isotropic Subspaces in Symplectic Space Over Finite Fields[J]. International Journal of Foundations of Computer Science, 2020, 31 (03).
[24] Veritas Technologies; Patent Issued for Systems And Methods For Distributing Information Across Distributed Storage Devices (USPTO 10, 592, 337) [J]. Computers, Networks &amp; Communications, 2020.
[25] BlueTalon Inc.; Patent Issued for Distributed Storage Processing Statement Interception And Modification (USPTO 10, 594, 737) [J]. Computers, Networks &amp; Communications, 2020.