Journal of Computers Vol. 27, No. 3, 2016, pp. 135-147
doi:10.3966/199115592016102703013

Using Early-warning and Active Data Migration Technologies for RAID-based Storage Systems

Yin Yang[1] Wenyi Li[1] Subhoyt Sandeep[2]

[1] School of Accounting, Wuhan Textile University, Wuhan 430200, China
cs_yangyin@hust.edu.cn, 108515256@qq.com

[2] School of Computer Science, Virginia Commanwealth University, Virginia VA 23284, USA
subhoyts@vcu.edu

Received 7 February 2015; Revised 27 May 2015; Accepted 28 August 2015

Abstract. The reliability of RAID storage systems can be increased according to the data redundancy method, which uses multiple disks, and parts of these disks are used to store redundant information, when disk failure occurs, system can utilize redundant information to recover RAID systems; however this will lead to more time to recover data and reduce the storage performances. Even worse, during data rebuilding, the possibility of secondary disk failure will increase. This paper proposes a new RAID-based system, which is defined as REA: RAID storage system with Early-warning and Active data migration. REA can predict RAID systems potential faults according to the disk health degree model. Once the fault has been predicted, REA uses active data migration technology to protect the data. REA can effectively avoid the complex checksum computation of the rebuild process and dynamically adjust the speed of data migration according to the system I/O load in order to reduce the RAID systems' performance. The overall results indicate that REA can achieve very high prediction accuracy and improve performance and reliability of storage systems with few side effects. Numerical results using SPC traces and standard benchmarks show that REA can effectively improve RAID reliability and performance.

Keywords: active data migration, disk health degree, early-warning, parameter health degree

1 Introduction

The rapid expansion of modern IT technology and the explosive growth of data banks place increasing demand on the storage systems capacities, performance, availability and reliability. Users and enterprises now desire information safety and reliability highly, and loss of data results in incalculable financial loss for these parties. Traditional RAID uses multiple disks, and many of these disk drives are used to place parity information. When one disk drive fails, the redundant information can be used to rebuild the disk via the RAID encoded mode [1]. However, this process leads to more time to recover data and has a detrimental effect on the performance of storage systems by the increased disk R/W and computational overhead. Further, following disk size increases, the RAID system needs longer data recovery time, so the faulty rate of the second disk increases in the data rebuilding period.

Although RAID technology plays a role in enhancing the reliability of storage systems, traditional RAID systems easily lose data. To address this loss of data limitation, we propose new RAID-based storage systems, which we have named REA: RAID storage systems with Early-warning and Active data migration.

REA creates a disk health degree model to predict potential faults. We first create a parameter health degree model using SMART parameters. Then, we choose some key parameters and confirm the weight of the key parameters to create a disk health degree model. Finally, we define and confirm the disk health degree early-warning threshold and compare it with the disk health degree to predict the disk faults. In case of the disk or system failure has been predicted, the faulty information should be given at once. It utilizes active data migration technology to protect the data [5].

For disk early-warning, REA can copy data from early-warning disk to spare disk; For early-warning of one array in multiple arrays, data can be moved into other backup array of the multiple arrays; For sector detecting, REA can repair part of disk media errors while recover the data from fault zone. REA effectively avoid lengthy process of data rebuilding after disk failure. In addition, the system's faults not really happen in the process of active data migration. So it can effectively avoid the complex checksum computation of the rebuild process and greatly reduce the impact on system performances. REA can dynamically adjust the speed of data migrate according to the system I/O load to reduce the impact on entire system's performance.

The purpose of this paper is to implement the function of early-warning and active data migration on the RAID system, further enhance the reliability of the RAID system. It can provide the most basic availability guarantee for data storage. REA should be implemented under the premise of impacting the system performances as little as possible. When the disk sector failure of the REA system appears, it should do the best to repair the sector. Meanwhile, when early-warning of a disk or an array in multiple arrays appears, it should move the data into a new free disk or a backup array.

The rest of this paper is organized as follows. The next section describes the related works. Section 3 introduces the system module and key technology of REA. Early-warning and active data migration technologies are presented. Section 4 evaluates the REA system and provides a detailed analysis of results. Finally, our conclusions are drawn in Section 5.

2 Related works

There are many reliability improvement approaches for RAID-based storage systems. These approaches main include the improvement of RAID reconstruction algorithms, such as PR [5], PRO [8], WorkOut [9], VDF [10], PUSH [18], S^2-RAID [19] and others [11]. PR analyzed the automatically rebuilding data stored in a failed disk into a spare disk in rebuild algorithms, and proposed a track-based rebuild algorithm that rebuilds lost data in tracks. PRO allows the reconstruction process in a RAID-structured storage system to rebuild the frequently accessed areas prior to rebuilding infrequently accessed areas to exploit access locality. WorkOut proposed a novel scheme to significantly improve RAID reconstruction performance, it outsources all write requests and popular read requests originally targeted at the degraded RAID set to a surrogate RAID set during reconstruction. VDF proposed a storage system consisting of a buffer cache and disk arrays to improve the reliability and performance, which give higher priority to cache the blocks on the faulty disks when the storage system fails, thus reducing the I/Os directed to the faulty disks. PUSH method incorporated PUSH-type transmissions to node reconstruction, where the reconstruction proceeding is divided into multiple assignments accomplished by surviving nodes in a pipelining manner, and also proposed two PUSH-based reconstruction schemes (i.e., PUSH-Rep and PUSH-Sur), which can not only exploit the I/O parallelism of PULL-Sur, but also maintain sequential I/O accesses inherited from PULL-Rep. RAID reconstruction process in case of a failure takes prohibitively long time. S^2-RAID method allows the disk array to reconstruct very quickly in case of a disk failure. The idea is to form skewed sub-arrays in the RAID so that reconstruction can be done in parallel dramatically speeding up data reconstruction process and hence minimizing the chance of data loss. Xie and Wang proposed a multi-level caching-based reconstruction optimization method, which collaboratively utilizes storage cache and disk array controller cache to diminish the number of physical disk accesses caused by reconstruction. In this paper, we use early-warning and active data migration techniques to improve the reliability of RAID system, which can predict the failure about sector, disk and array, and move data from any failure storage device to backup device.

Some researchers have studied different methods for analyzing the reason for disk and array reliability. Goldszmidt uses D-FAD to build statistical models for predict soon-to-fail disks [7]. Hughes, Murray, Kreutz-Delgado and Elkan introduced Wilcoxon rank sum statistical to test prediction model, which increases failure prediction accuracy and lower false prediction rates [2]. Murray, Hughes and Kreutz-Delgado use the multiple-instance learning framework and the naive Bayesian classifier method to analyze prediction failure algorithm [4]. Hamerly and Elkan proposed a disk drive failures prediction method based on Bayesian methods by the measurements of drive internal conditions [3]. Wang, Miao, Ma, Tsui and Pecht used Mahalanobis distance method to detect HDD anomaly. Critical parameters were selected using failure modes, mechanisms, and effects analysis (FMMEA), and the minimum redundancy maximum relevance (mRMR) method. A self-monitoring, analysis, and reporting technology (SMART) data set is used to evaluate the performance of the developed approach [16].

Different from rank sum hypothesis [2], statistical model [7], Bayesian [3], machine learning [4] and other principles [16] of disk failure prediction, we use the SMART method to achieve early-warning. We compare disk health degree with disk health degree early-warning threshold to judge the disk reliability and confirm the rank of disk reliability. Meanwhile, in our paper, we detect sector condition and collect some running information: the state of the disk health, the state of array, temperature and the fan speed in the array, and so on. Once the fault has been predicted, the early-warning information should be given immediately, REA can utilize disk self-healing technology, disk migration technology or array self-healing technology to protect the data in RAID-based storage systems. So, we not only can predict the disk's failure, but also protect the reliability of the whole storage system.

3 REA system module and key technology

3.1 System module

An early-warning module is divided into two modules: fault detection module and fault decision module. Fault decision module is divided into five sub-modules: configuration interface, strategy center, array health gatherer, fault analysis, and early-warning schedule. The REA system builds an appropriate WEB page and manages a configuration interface by the WEB page, and utilizes the WEB browser to realize the function of client configuration and status information acquisition. Array health gatherer collects system health information of different device objects in regular time. Strategy center sets the early-warning scheme through user configured information, and then performs fault analysis of the REA system through health information and the early-warning scheme. For abnormal information, the early-warning schedule module triggers active data migration.

An active data migration module is divided into four sub-modules: disk I/O agent, disk migration and array self-healing, disk self-healing, and disk attribute management. The disk I/O agent includes many child agents. When the upper client sends a disk I/O request, each child agent will be invoked. It provides a registration interface, which is used to install child agents. Disk self-healing, disk migration and array self-healing will command the registration interface to install child agents. Child agents of disk self-healing analyze the implementation status of write and read requests by cutting out disk I/O. If these requests fail during the read-write process, disk attribute management model will trigger disk self-healing to restore failure sectors by recording the failure, and may evaluate the health status to decide whether disk migration and array self-healing is needed to improve data reliability. The child agent of disk migration intercepts disk I/O, when the requested disk is in the procedure of disk migration, it needs to redirect I/O requests to the objective disk. When the array's health can't meet standards, array self-healing is triggered, together with array data migration, according to the background thread migration. Disk attribute management model manages system properties, consisting of the reliability state and property information of other modules.

According to the REA system module, the operation way of REA is shown in Fig.1. REA use fault detection module to monitor the running conditions including disk, array and computer-case. This paper creates disk health degree model and array health model to predict the potential fault of RAID, and confirm disk health degree early-warning threshold, we compare disk health degree with disk health degree early-warning threshold, and thereby the potential predictable disk failures can be predicted. According to the fault decision module, once the fault has been predicted, the early-warning information should be given immediately, it takes active data migration technology to protect the data. REA can copy data from imminent faulty location to safety area. Active data migration technology is divided into the following three levels: sector level, disk level, array level. When detecting information is at sector level, REA can utilize disk self-healing technology, which try to repair part of disk media errors while recovering the data from fault zone; when early-warning information is at disk level, REA can utilize disk migration technology, which can quickly copy the whole fault disk to other spare disks, for the data of sectors whose media error occurs, the method of data checking can be used to recomputed in order to recover the original data(this data copy can avoid the complex checksum computation in the long RAID rebuilding time); When early-warning of one array in multiple arrays, REA can utilize array self-healing technology, which can complete array self-healing in multiple arrays storage system, and take advantage of the remaining hot spare disks to completely copy the early-warning fault array. So in RAID system, we can utilize early-warning and active data migration technologies to enhance the reliability of the RAID storage system.

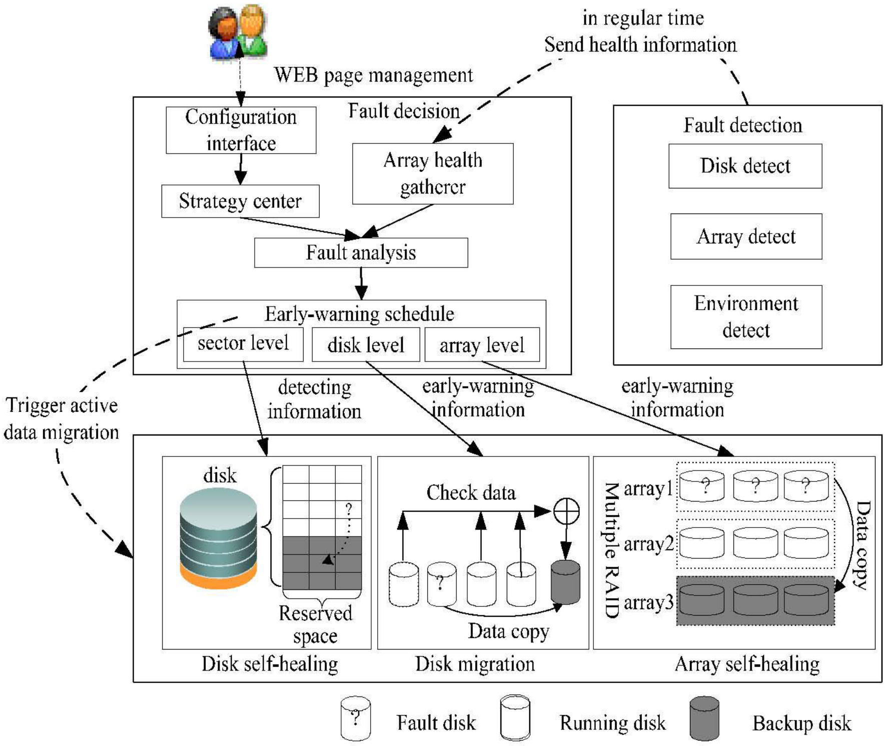


Fig. 1. The operation way of REA

3.2 Key technology

The key point of REA realization is early-warning and active data migration technologies. In early-warning technology, REA can monitor the health conditions including disk, array and environment. When the system is in poor health, the early-warning information should be given immediately, which provides ground for active data migration. In active data migration technology, when disk has part of bad sectors, REA can continue working and recover the data of bad sectors, so the system do not require immediately using new disk and data rebuilding, this disk can continue working through self-healing. Meanwhile, when REA receives the early-warning information, the system can copy fault data to other spare disks or backup arrays and migrate process do not affect normal work flow; the influence of system performances should be reduced to the minimum.

3.2.1 Early-warning technology

The early-warning technology is based on SMART technology. SMART is an acronym for Self-Monitoring, Analysis and Reporting Technology and was designed by IBM. It detects lots of disk status by using various methods and sensors. The current status of the disk drive is all the time checked by the devices. Then, the measured values are handled by several algorithms and the corresponding parameters are changed by the results. SMART monitors and reports failure situations in disk drive devices, whether from destroy or normal wear.

Pinheiro, Weber and Barroso proposed that even with SMART parameters, it was found that 36% of failures had zero counts on all measures [12]. Further up, depended only on those counts, disk SMART can predict about 70% hard drive failures. Many researchers attempted to correlate SMART parameters with failure statistics [2-4]. They use disk failure methods that acquire better prediction rates based on some importance SMART parameters at false rates of about 0.2%. These results suggest that the derived models achieve very good failure prediction accuracy.

The overall prediction and migration process in REA includes two parts: an early-warning process (Fig.2) and an active data migration process (Fig.3). We create two health degrees model: parameter health degree (PHD) and disk health degree (DHD). We use data value, threshold value and attribute value of SMART parameters to compute the value of PHD. Users real-time acquire the Data value (DT) of disk SMART parameter. Threshold value (TH) is the failure restricts value settings by disk drive manufacturer. Attribute value (AV) has been adjusted to the maximum regular value as default. We use smartmontools (an open source disk control and monitor tool) to obtain the values of the DT, TH and AV of disk SMART parameters. The parameter health degree formula is,

$$PHD = 100 \times \left((DT - TH)/(AV - TH) \right) \quad \textbf{(1)}$$

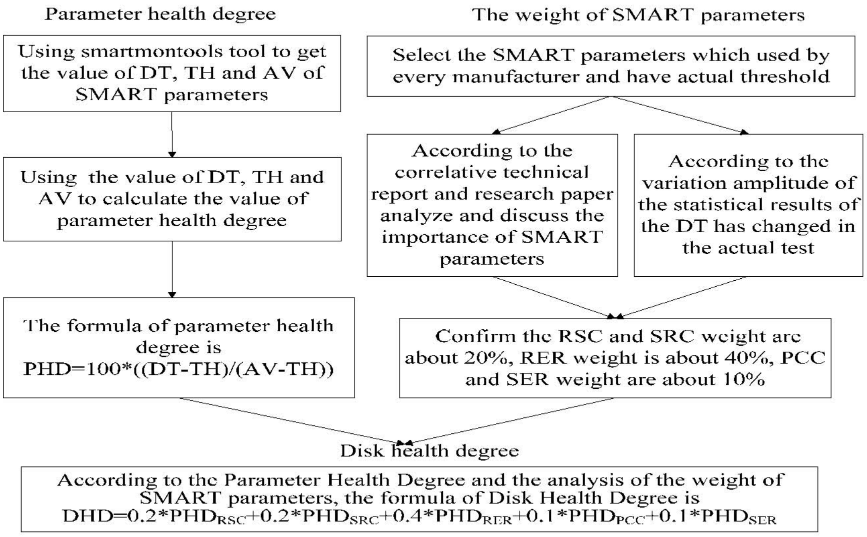

Parameter health degree
The weight of SMART parameters
Using smartmontools tool to get the value of DT, TH and AV of SMART parameters
Select the SMART parameters which used by every manufacturer and have actual threshold
Using the value of DT, TH and AV to calculate the value of parameter health degree
According to the correlative technical report and research paper analyze and discuss the importance of SMART parameters
According to the variation amplitude of the statistical results of the DT has changed in the actual test
The formula of parameter health degree is PHD=100*((DT-TH)/(AV-TH))
Confirm the RSC and SRC weight are about 20%, RER weight is about 40%, PCC and SER weight are about 10%
Disk health degree
According to the Parameter Health Degree and the analysis of the weight of SMART parameters, the formula of Disk Health Degree is DHD=0.2*PHDRSC+0.2*PHDSRC+0.4*PHDRER+0.1*PHDPCC+0.1*PHDSER

Fig. 2. The early-warning process in REA

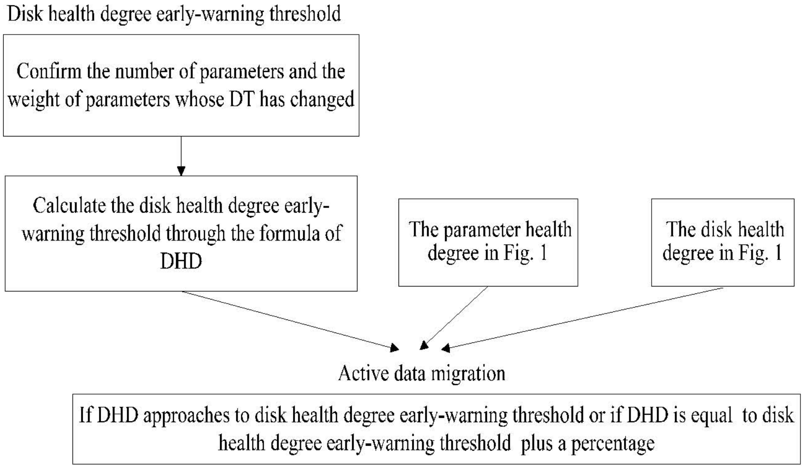

Disk health degree early-warning threshold
Confirm the number of parameters and the weight of parameters whose DT has changed
Calculate the disk health degree early-warning threshold through the formula of DHD
The parameter health degree in Fig. 1
The disk health degree in Fig. 1
Active data migration
If DHD approaches to disk health degree early-warning threshold or if DHD is equal to disk health degree early-warning threshold plus a percentage

Fig. 3. The active data migration process in REA

The different parameters have a differential impact on disk health conditions. Five disk SMART parameters (Reallocated Sectors Count (RSC), Spin-up Retry Count (SRC), Raw Read Error Rate (RER), Power Cycle Count (PCC) and Seek Error Rate (SER)) have greater affect the disk drives faulty and for each DT can compare with threshold value. The parameter weights are determined by two factors: technical reports [13-15] and research papers [2, 12]. These technical reports and research papers analyze the results from long-time tests on a large number of disks, confirming which of the SMART parameters actually do seriously reflect the reliability of disks. Although these studies don't give the specific rankings and weight values of these important SMART parameters, they point out which SMART parameters are most predictive of disk reliability. The second aspect is that through the analysis of the real test results of disk SMART parameters, the amplitude of DT variety can be observed and then it can be confirmed which SMART parameters can seriously affect disk reliability.

From the analysis above, we can estimate SMART parameters' weight values. We assume the average weight of five SMART parameters is 20%. If technical reports and research papers analyzing and discussing the SMART parameter are very important, and the amplitude of DT variety is large, these SMART parameters discussed have a relatively greater weight (RER is about 40%). If technical reports and research papers analyzing and discussing the SMART parameter are relatively important, and the amplitude of DT variety is relatively small, these SMART parameters discussed have a relatively smaller weight (PCC is about 10%). If technical reports and research papers analyzing and discussing the SMART parameter are not very important, and the amplitude of DT variety is large, these SMART parameters discussed have a relatively smaller weight (SER is about 10%). And, if technical reports and research papers analyzing and discussing the SMART parameter are important, and the amplitude of DT variety is relatively small, these SMART parameters discussed have a relatively big weight (RSC and SRC are about 20%).

According to the formula of PHD and the weight of important disk SMART parameters, the formula of disk health degree is,

$$DHD = 0.2 \times PHD_{RSC} + 0.2 \times PHD_{SRC} + 0.4 \times PHD_{RER} + 0.1 \times PHD_{PCC} + 0.1 \times PHD_{SER} \quad \textbf{(2)}$$

The Disk Health Degree Early-warning Threshold (DHDET) is confirmed by two conditions [17]:

(1) Of the five importance SMART parameters, the number of the SMART parameters whose DT has changed.

In case of the DT of a disk SMART parameter has changed, which shows that this SMART parameter has negative influence on DHD; In case of the DT of some disk SMART parameters have simultaneously changed, it shows that these parameters have a greater negative impact on DHD. In Table 1, the calculation of DHD includes the value of PHD and the parameter weight. The PHD of the parameters RSC, SRC, RER, PCC and SER is calculated with the value of DT set to 70, 99, 110, 70 and 70, respectively. The PHD value for each parameter is as follows: 53 for RSC, 67 for SRC, 54 for RER, 63 for RCC and 57 for SER. The value of DHD is subsequently calculated from these values of PHD, and the formula of DHD.

Table 1. The DHD of the number of difference parameters whose DT has changed

The number of DT changed parameters	The name of DT changed parameters	DHD
Zero	None	100
One	RSC	90.6
Two	RSC;SRC	84
Three	RSC;SRC;RER	65.6
Four	RSC;SRC;RER;PCC	61.9
Five	RSC;SRC;RER;PCC;SER	57.6

(2) The weight of impotence SMART parameters whose DT has changed.

In case of the weight of DT changed parameters also has an influence on DHD. The greater the weight of a changed DT, the greater is the negative influence on DHD. The method to calculate the DHD of every parameter in Table 2 is as follows: the PHD of the calculated parameter is 80, while the PHD of all other parameters is 100. For example: as the weight of PCC is 10 %, we see that $DHD = 0.2 \times 100 + 0.2 \times 100 + 0.4 \times 100 + 0.1 \times 80 + 0.1 \times 100 = 98$.

Table 2. The DHD of the different weight parameters whose DT has changed

The weight of *DT* changed parameter	The name of *DT* changed parameter	*DHD*
10%	*PCC*	98
10%	*SER*	98
20%	*RSC*	96
20%	*SRC*	96
40%	*RER*	92

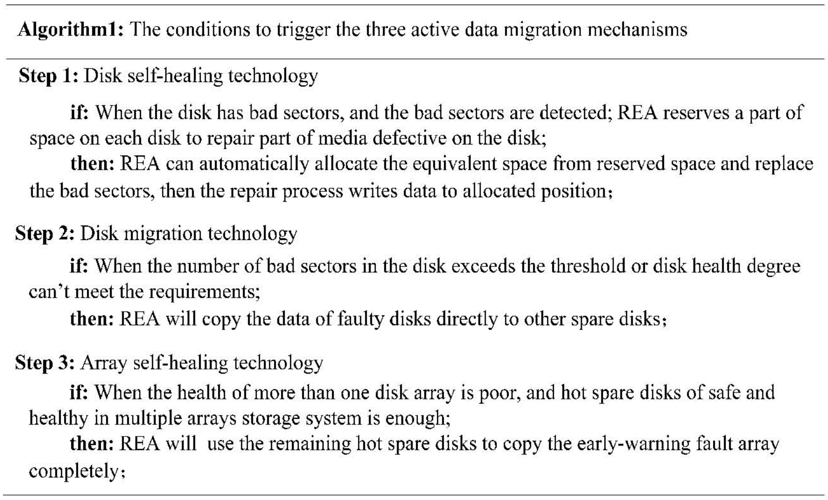
From the values in Table 1 (the number of parameters for which DT has changed) and the DHD method (the weight of impotence SMART parameters whose DT has changed), we may compute DHDET. This paper uses one SMART parameter variation as an example to explain the computation of DHDET. In the case only the DT of RSC has changed, and the value of RSC's PHD reduces from 100 to 0, all other measures are 100. By the DHD formula, when the DT of RSC has changed from 100 (AV) to 36

(TH), other parameters are 100 (AV), the PHD of RSC is accompanied with DT change, and others are 100 (PHD). The value of DHD is calculated from the formula of PHD and DHD. When the weight is 20%, the DHDET of RSC and SRC are $0.2\times 0+(0.2+0.4+0.1+0.1)\times 100=80$; when the weight is 10%, the DHDET of PCC and SER is $0.1\times 0+(0.2+0.2+0.4+0.1)\times 100=90$; when the weight is 40%, the DHDET of RER is $0.4\times 0+(0.2+0.2+0.1+0.1)\times 100=60$. The calculation process of these DHDET values is the same as for RSC.

DHDET hints the imminence of data loss; if the DHD value of a disk equals to the DHDET value, the disk has failed. Hence, to obtain the results of prediction, the DHDET should be offset by a positive delta. If the DHD of a disk is approaching the DHDET, the failure information should be provided at once.

3.2.2 Active data migration technology

Active data migration is divided into three levels: sector level, disk level and array level, and REA can utilize sector self-healing, disk migration or array self-healing respectively. The conditions to trigger activate data migration are shown in Fig.4.



```
Algorithm1: The conditions to trigger the three active data migration mechanisms
Step 1: Disk self-healing technology
    if: When the disk has bad sectors, and the bad sectors are detected; REA reserves a part of space on each disk to repair part of media defective on the disk;
    then: REA can automatically allocate the equivalent space from reserved space and replace the bad sectors, then the repair process writes data to allocated position；
Step 2: Disk migration technology
    if: When the number of bad sectors in the disk exceeds the threshold or disk health degree can't meet the requirements;
    then: REA will copy the data of faulty disks directly to other spare disks；
Step 3: Array self-healing technology
    if: When the health of more than one disk array is poor, and hot spare disks of safe and healthy in multiple arrays storage system is enough;
    then: REA will  use the remaining hot spare disks to copy the early-warning fault array completely；
```

Fig. 4. The conditions to trigger the active data migration mechanisms in REA

When the disk has bad sectors, REA can recover the data from the fault zone of one disk and continue working, so the storage system does not immediately require new disk usage and RAID data rebuild. When the number of bad sectors exceeds the maximal value for one disk or the disk otherwise can't meet the health requirements, REA will quickly move the data to the backup disk or to a new disk. When the health of more than one disk array and the state of the array and environment are poor, data can be moved into other backup arrays. REA can effectively avoid the complex checksum computation of the data rebuild process in RAID and dynamically adjust the speed of data migration according to the REA system I/O workload in order to decrease the influence on storage systems performance.

Disk self-healing technology works to partially repair disk media errors while recovering the data from faulty zones. In order to carry out repairs, REA reserves space on each disk to use as placements for defective parts of the disk medium. When the disk has bad sectors, sectors from the reserved space can be allocated to take the place of the faulty sectors. Subsequent requests for access to bad sectors are then redirected to those reserved sectors that have been assigned as replacements for those bad sectors.

When disk health indicates imminent failure, REA will copy the data of faulty disks directly to other spare disks. While data migration is going on, the system can quickly copy the whole faulty disk to other free disks. For the data of sectors where a media error occurs, data checking can be used to recover the original data.

When the health of more than one disk array is poor, data errors may occur at any time, and data migration must therefore start as soon as possible. If safe and healthy hot spare disks in the multiple arrays' storage system are sufficient, array self-healing can be completed, using those remaining spare hot disks to completely copy the early-warning fault array. In this paper, disk migration and array-level self-healing are considered as the same.

4 Evaluation methodology

4.1 Experimental settings

In our experiment, we use an iSCSI storage server to install the prototype for REA system. Storage clients were connected to the storage server use a Cisco 3750 Gb Ethernet to connect the storage clients. The system configurations of the REA system is listed in Table 3.

Table 3. System configurations of REA system

CPU	Intel iop 80321(500MHZ)
RAM	DDR 512MB
DISK	Seagate Barracuda 7200 160G
FC	Agilent 2G
OS	Linux 2.6.11
NIC	Intel® PRO/1000

4.2 Numerical results and discussions of early-warning

The results of early-warning evaluation are listed in Table 4, including the experiment results of ten disks in a disk array. Through Section 3.2.1, the results indicate the case where RER and SER have changed, and where DHDET is 50. Whereas DHDET indicates when a disk is about to reach a state of data loss, if it should also happen that the DHD of a disk equals the DHDET, then the disk has failed. Hence, to obtain prediction, the DHDET should be offset by a positive delta, take 10% as an example, the reliability early-warning maximal value is 50 + 50*10% = 55. So, in Table 4, about disk reliability, from disk 1 to disk 9 are healthy disk, but disk 10 is unhealthy, and the health of disk 1 is the worst and has least reliability; the health of disk 3 and 6 are better and these have more reliability. If the disk health degree approaches 50, the fault can be predicted and the failure message should be given at once.

Table 4. The test results of disk early-warning and corresponding DHD

	RSC			SRC			RER			PCC			SER			
Attribute	DT	TH	AV	DT	TH	AV	DT	TH	AV	DT	TH	AV	DT	TH	AV	DHD
1	100	36	100	100	97	100	53	6	200	99	20	100	78	30	100	66
2	100	36	100	100	97	100	117	6	200	99	20	100	51	30	100	75
3	100	36	100	100	97	100	117	6	200	99	20	100	85	30	100	80
4	100	36	100	100	97	100	114	6	200	100	20	100	84	30	100	79
5	100	36	100	100	97	100	105	6	200	99	20	100	58	30	100	74
6	100	36	100	100	97	100	118	6	200	99	20	100	83	30	100	80
7	100	36	100	100	97	100	113	6	200	99	20	100	78	30	100	78
8	100	36	100	100	97	100	117	6	200	99	20	100	67	30	100	77
9	100	36	100	100	97	100	118	6	200	99	20	100	78	30	100	78
10	100	36	100	100	97	100	23	6	200	99	20	100	37	30	100	54

4.3 Numerical results and discussions of active data migration

According to the evaluation test in experiments, we give and compare four states of the RAID5 storage systems: Traditional Normal RAID (TNR), Traditional Rebuild RAID (TRR), Disk Self-healing RAID (DSR), and Disk Migration RAID (DMR). TNR consists of striping, with mirroring or parity. There is no added redundancy for handing disk failure, just as with a spanned volume. TRR consists of block-level striping with distributed parity. The parity information is distributed among the disk drives. It requires that all disk drives but one be present to operate. Upon failure of a single disk drive, subsequent reads can be calculated from the distributed parity such that no data is lost. TRR is seriously affected by the general trends regarding array rebuilding time and the change of disk drive failure during rebuild. REA is the RAID system with early-warning and active data migration technologies. DSR is the RAID system with disk self-healing technology. DMR is the RAID system with disk migration technology. We use Iometer by the view of the request size to give the throughput performance. Then, we get the test results of REA according to replay the practical traces during the recovery evaluation.

I/O performance: the throughput performance of the states of the RAID5 storage systems for TNR and DSR are shown in Fig.5 and Fig.6, which measure the random and sequential I/O access for various request size. The same as we have expected, the throughput performance of DSR system is very close to that of traditional normal RAID, which indicates there is little or no influence on RAID performance for the operations of the self-healing RAID system. The main reason that DSR has some minor impact on system performance is that some I/O agents are added in disk self-healing; for one I/O agent, the execution of an I/O request needs to execute more instructions, and each I/O request needs to find out whether bad sectors exist circularly, but the influence is within acceptable limits, so, the throughput of DSR is always close to the TNR. I/O performance: the throughput performance of the two RAID5 storage systems for TNR and DMR is shown in Fig.7, which measure the random and sequential I/O access for various request size.

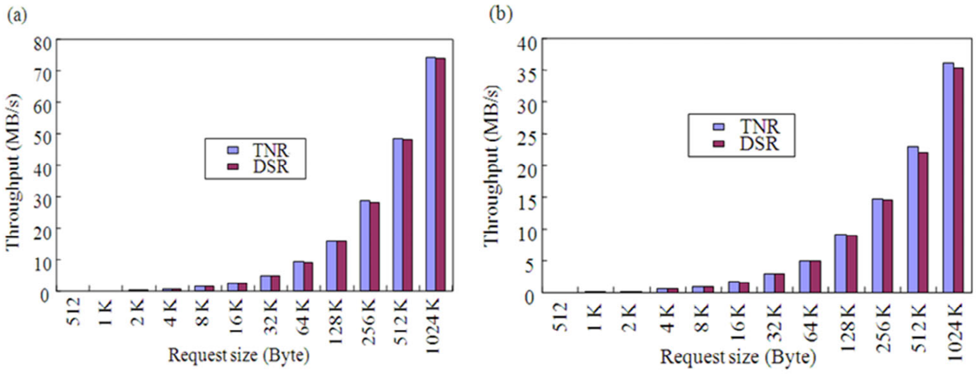


Fig. 5. The throughput of TNR and DSR systems with Iometer. (a) Random read; and (b) Random write

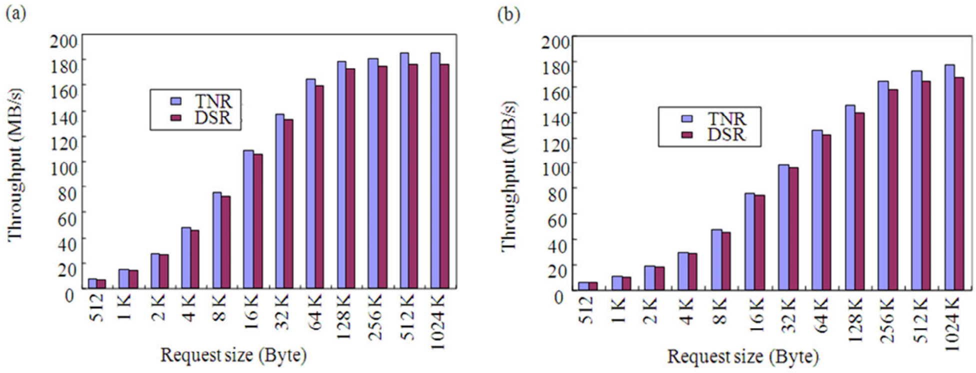


Fig. 6. The throughput of TNR and DSR systems with Iometer. (a) Sequential read; and (b) Sequential write

The disk migration in DMR brings many systems I/O, and system background opens these disk migration threads. Because of the CPU occupancy rate decide the disk migration speed, when the speed is too fast, it can cause a bottleneck in system performance. Dynamically adjusting the speed of data migration decreases the influence on DMR performance. From the line of data migration velocity in Fig.7, when the I/O workload increases in DMR, the speed of data migration would also automatically decrease the influence on DMR performance. The data migration in DMR has some influence on normal performance of I/O workload, compared with the traditional rebuild RAID system (Table 5), the impact on the RAID is much smaller.

We compare the throughput of TNR, DSR, DMR and TRR under the condition that the system reading is sequential and the request size is set to 1024 K. TNR throughput can exceed 180 MB/s, DSR throughput can exceed 160 MB/s, DMR throughput can exceed 140 MB/s, and TRR throughput is only 50 MB/s. The throughput of DSR, DMR and TRR respectively decreased by 11%, 22% and 72% compared with TNR, and the throughput of DSR and DMR respectively increased by 220% and 180% compared with TRR. So RAID rebuilding (TRR) performance is poorer than that of the traditional normal RAID (TNR),

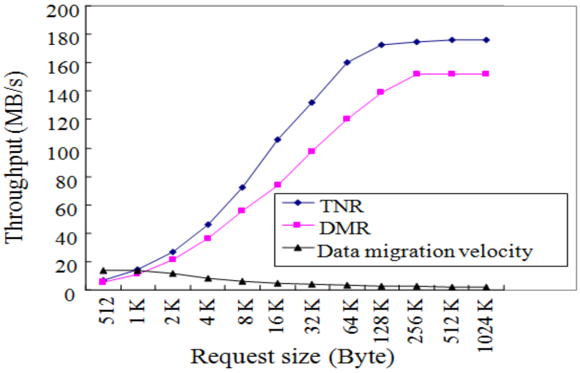


Fig. 7. The throughput of TNR and DMR systems with Iometer

Table 5. The throughput of TRR systems

Request type	Request size (K)	Throughput (MB/s)
Sequential read	1024	50
Sequential write	1024	48
Random read	1024	32
Random write	1024	15

but the performance of disk self-healing (DSR) and disk migration (DMR) RAID system is almost the same as that of the traditional normal RAID system (TNR), and it is obviously higher than the performance of traditional rebuild RAID (TRR).

The performance influence about average response time and recovery time for DSR, DMR and TRR RAID systems are given by the SPC traces [6] and benchmarks. We measured the average response time and recovery time performance according to online recovery operation.

As we can see from Fig.8(a), for the average response time of performance evaluation, DSR only needs seeking out the redirecting address of the fault sector in the disk hash table to keep normal read and write. DSR has little I/O request overhead on the disk, so that the influence on the value of average response time is small, especially using Financial 1 trace, DSR and DMR are close to 0. DMR reads the data from the disk which is in danger of becoming corrupted and writes it to a backup disk while dealing with normal read and write requests from clients. The whole access process requires many I/O requests which seriously influence on the average response time performance. TRR requires access these data on other disk arrays. According to the XOR method, system writes these data in failure disk to a new disk. The XOR operate consumes a great lot of I/O requests time as well as CPU time, which has greater influence on the average response time of the system. So, the average response time of TRR is higher than DSR and DMR. As seen in Fig.8(b), for the recovery time of performance evaluation, DSR finishes system failure recovery while dealing with I/O requests, which leading to the recovery time is close to zero. TRR requires read the data on other disk arrays, and according to the XOR method, system writes these data, so that this operation occupies more time than DMR. So, the average response time and recovery time of disk self-healing (DSR) and disk migration (DMR) RAID systems are obviously better than the performance of traditional rebuild RAID (TRR).

From Fig.9, the average response time and recovery time performance of Benchmark tools the same as the description of Fig.8. So, the average response time and recovery time of disk self-healing (DSR) and disk migration (DMR) RAID systems using Benchmark tools is obviously better than the performance of traditional rebuild RAID (TRR).

We use 8-disk RAID5 to test our method, and run experiment about PRO, PR, WorkOut + PRO, WorkOut + PR, DSR and DMR respectively. According to the three traces, the recovery time and average response time performance have shown as Fig.10 and Fig.11.

From Fig.10, it can be seen that the recovery time performance of DSR and DMR are much better than PRO, PR, WorkOut + PRO and WorkOut + PR. According to Fig.11, we can obtain the similar results that the average response time performance of DSR and DMR are much better than PRO, PR, WorkOut + PRO and WorkOut + PR. The comparison results of benchmarks are the same as traces. According to

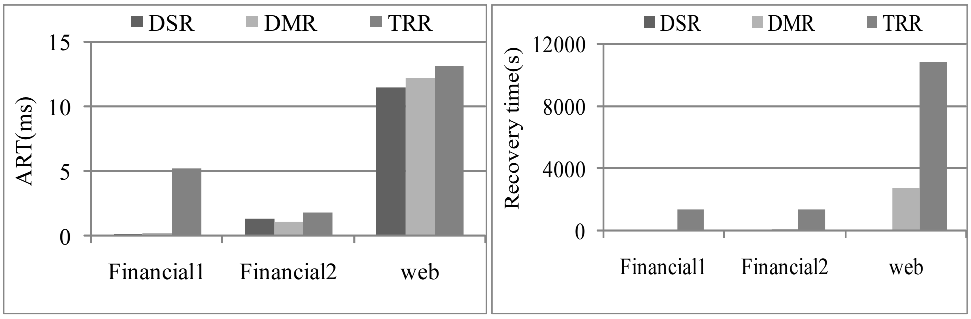


Fig. 8. The performance with three traces.

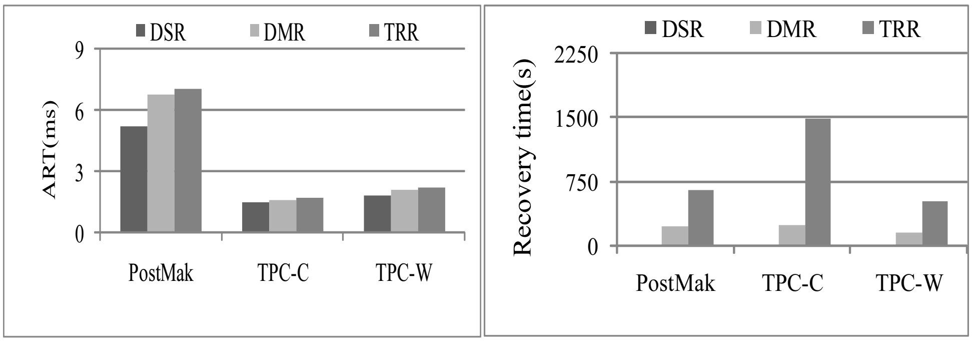


Fig. 9. The performance with three benchmarks.

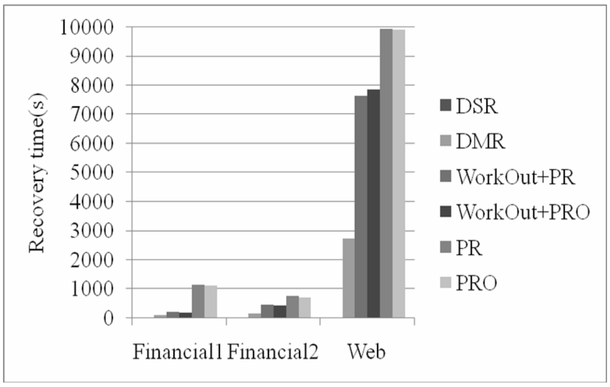


Fig. 10. Comparisons of recovery time performance driven by three traces

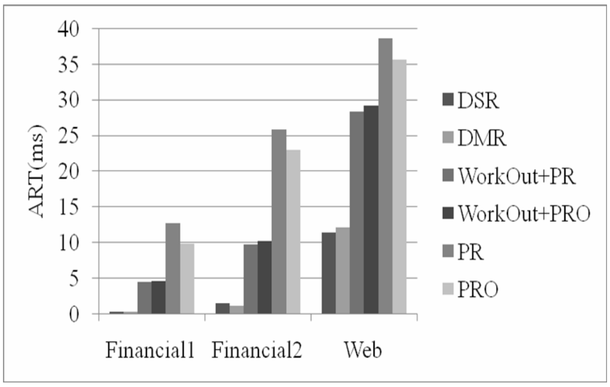


Fig. 11. Comparisons of average response time performance driven by three traces

the results of recovery time and average response time performance, REA effectively avoid lengthy process of data rebuilding after disk failure for DMR. In addition, the system's faults not really happen in the process of active data migration. So it can effectively avoid the complex checksum computation of the rebuild process and greatly reduce the impact on system performances for DMR. REA can dynamically adjust the speed of data migrate according to the system I/O load to reduce the impact on entire system's performance.

5 Conclusion

Traditional RAID system rebuilding is a challenge and threat to the performance and reliability of systems. This paper proposes a RAID storage system, which as REA using early-warning and active data migration technologies. The early-warning technology is based on SMART technology, which can effectively predict system faults in advance. Once a fault has been predicted, REA can utilize active data migration technology to remove data from locations where faults are imminent to more appropriate locations. Evaluation experiment using SPC traces and benchmarks show that REA can effectively increase the reliability and performance of RAID system.

However, this work can be improved and there are many alternative avenues of research to be explored. First, we have only used the SMART technology for prediction of failures. Because the disk is mechanical equipment and the operating environments of storage systems are varied, we plan to incorporate machine learning methods to train disk data and use signals from healthy and failed disks, which can increase the prediction accuracy and lower false alarms. Second, deploying REA is a trade-off between the cost of false positives and the loss of service quality in active data migration if preventive actions are not performed. Nevertheless the reported results and the advantages of the methodology provide evidence that this is a fruitful avenue of research to pursue. Finally, we plan to build a cloud failure prediction model by analyzing the attribute parameters, and the working and running state of the disk medium and cloud nodes, combining some relative techniques and methods. We can use this model to protect the cloud data on failed disks and cloud nodes by active data migration on system level.

Acknowledgement

This work was sponsored by the Natural Science Foundation of Hubei Province under Grant No. 2015CFB235, the National Basic Research Program of China (973 Program) under Grant No.2011CB302303 and the National Natural Science Foundation of China under Grant No.60933002.

References

[1] Patterson, D. A., Gibson, G., & Katz, R. H. (1988). A case for redundant arrays of inexpensive disks (RAID). *ACM SIGMOD Record*, *17*, 109-116.

[2] Hughes, G. F., Murray, J. F., Kreutz-Delgado, K., & Elkan, C. (2002). Improved disk-drive failure warnings. *IEEE Transactions on Reliability, 51*, 350-357.

[3] Hamerly, G., & Elkan, C. (2001). Bayesian approaches to failure prediction for disk drives. In C. E. Brodley & A. P. Danyluk (Eds.), *Proceedings of the eighteenth international conference on machine learning* (pp. 202-209). San Francisco, CA: Morgan Kaufmann.

[4] Murray, J. F., Hughes, G. F., & Kreutz-Delgado, K. (2005). Machine learning methods for predicting failures in hard drives: A multiple-instance application. *The Journal of Machine Learning research, 6*, 783-816.

[5] Lee, J. Y. B., & Lui, J. C. S. (2002). Automatic recovery from disk failure in continuous-media servers. *IEEE Transactions on Parallel and Distributed Systems, 13*, 499-515.

[6] UMass Trace Repository. (2007). *OLTP application I/O and search engine I/O*. Retrieved from http://traces.cs.umass.

edu/index.php/storage.

[7] Goldszmidt, M. (2012). Finding soon-to-fail disks in a haystack. In *Proceedings of the 4th USENIX Conference on Hot Topics in Storage and File Systems* (pp. 8-8). Berkeley, CA: USENIX Association.

[8] Tian, L., Feng, D., Jiang, H., Zhou, K., Zeng, L., Chen, J., Wang, Z., & Song, Z. (2007, February). PRO: A popularity-based multi-threaded reconstruction optimization for RAID-structured storage systems. Paper presented at the Proceedings of the 5th USENIX Conference on File and Storage Technologies, San Jose, CA.

[9] Wu, S., Jiang, H., Feng, D., Tian, L., & Mao, B. (2009, February). *WorkOut: I/O workload outsourcing for boosting RAID reconstruction performance*. Paper presented at the Proceedings of the 7th USENIX Conference on File and Storage Technologies, Berkeley, CA.

[10] Wan, S., Cao, Q., Huang, J., Li, S., Li, X., Zhan, S., Yu, L., Xie, C., & He, X. (2011, June). *Victim disk first: An asymmetric cache to boost the performance of disk arrays under faulty conditions*. Paper presented at the Proceedings of the 2011 USENIX Annual Technical Conference, Portland, OR.

[11] Xie, T., & Wang, H. (2008). MICRO: A multilevel caching-based reconstruction optimization for mobile storage systems. *IEEE Transactions on Computers, 57*, 1386-1398.

[12] Pinheiro, E., Weber, W.-D., & Barroso, L. A. (2007, February). Failure trends in a large disk drive population. Paper presented at the Proceedings of the 5th USENIX Conference on File and Storage Technologies, San Jose, CA.

[13] Seagate. (1997). *Seagate statement on enhanced smart attributes* (Seagate Technology Paper TP-67D). Scotts Valley, CA: Seagate Technology.

[14] Wikipedia. (n.d). Self-monitoring, analysis, and reporting technology. Retrieved from http://en.wikipedia.org/wiki/S.M.A.R.T

[15] Hungary, H. D. S. (2007). *Can we believe S. M. A. R. T.?* (Technical Committee T13) Washington, DC: Information Technology Industry Council.

[16] Wang, Y., Miao, Q., Ma, E. W. M., Tsui, K.-L. & Pecht, M. G. (2013). Online anomaly detection for hard disk drives based on Mahalanobis distance. *IEEE Transactions on Reliability, 62*, 136-145.

[17] Yang, Y., Tan, Z. H., Wan, J. G., Xie, C. S., Yu, J., & He, J. (2013). A reliability optimization method for RAID-structured storage systems based on active data migration. *Journal of Systems and Software, 86*, 468-484.

[18] Huang, J. Z., Liang, X. H., Qin, X., Cao, Q., & Xie, C. S. (2015). PUSH: A pipelined reconstruction I/O for erasure-coded storage clusters. *IEEE Transactions on Parallel and Distributed, 26*, 516-526.

[19] Wan, J. G., Wang, J. B., Xie, C. S., Yang, Q. (2014). S^2-RAID: Parallel RAID architecture for fast data recovery. *IEEE Transactions on Parallel and Distributed, 25*, 1638-1647.