RAID Storage Systems with Early-warning and Data Migration

Yin Yang[12]
[1]School of Computer Sci. and Tech. Huazhong University of Sci. and Tech
[2]Wuhan National Laboratory for Optoelectronics
Wuhan, China
yy16036551@smail.hust.edu.cn

Zhihu Tan[12*]
[1]School of Computer Sci. and Tech. Huazhong University of Sci. and Tech
[2]Wuhan National Laboratory for Optoelectronics
Wuhan, China
stan@mail.hust.edu.cn

Wei Liang
The 722 Institute of China Shipbuilding Industry Corporation
Wuhan, China
liangwei19830725@gmail.com

ChangSheng Xie[12]
[1]School of Computer Sci. and Tech. Huazhong University of Sci. and Tech
[2]Wuhan National Laboratory for Optoelectronics
Wuhan, China
cs_xie@mail.hust.edu.cn

***Abstract*—This paper proposes a new RAID-based system, we defined as REM: RAID system with Early-warning and data Migration. REM can predict RAID potential fault according to disk health degree and array health. Once the fault has been predicted, it takes data migration to protect the data, which allocates reserved space in every disk and arranges to redirect the data in the non-repairable sectors to the reserved space; when the number of the bad sectors in the disk exceeds threshold or the disk is in poor health, the system will copy data of the unreliable disk to a new one; when an array of multiple arrays early-warning, data can be moved into other backup array of the multiple arrays. Test results show that REM can enhance the reliability of system and reduce the entire system's performance.**

Keywords- early-warning; data migration; reliability;

I. INTRODUCTION

The rapid development of modern information technology and the explosive growth of information data put forward high requirement on the storage system capacity, performance, high availability and reliability. It has become the focus of user and enterprise to ensure information data safe and reliable. Once the data is lost, the loss will be incalculable. In order to enhance the reliability of storage system, RAID technology [1] emerges as the times require. Traditional RAID is combined by multiple disks, and parts of the disk are used to store redundant information. When the member disk of the RAID goes wrong, we can use the redundant information to rebuild the storage system by RAID encoded mode [2]. However, this process will result in longer data recovery time and negative impact on system I/O performances because of additional disk I/O and calculate. What's more, as disk capacity increases, it needs more time to recover the system, thus the error rate of the second disk will increase in the time window of data rebuild [3].

Although RAID technology plays a role to enhance the reliability of storage system, traditional RAID is easy to lose data. Consequently, this paper proposes a new RAID-based system, we defined as REM: RAID system with Early-warning and data Migration. REM can monitor the running conditions including disk, array and computer-case, and creates disk health degree model and array health model to predict the potential fault of RAID. Different from rank sum hypothesis methods [4], non-parametric statistical methods [5], bayesian approaches [6], machine learning methods [7] and other principles of disk failure prediction, we first create parameter health degree model and disk health degree model. By analyzing the number of parameters whose DT has changed and the weight of parameters whose DT has changed, the disk health degree early-warning threshold can be confirmed, we compare disk health degree with disk health degree early-warning threshold, thereby the potential predictable disk failures can be predicted. Then, it takes data migration technology to protect the data [8]. For disk early-warning, REM 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, REM can repair part of disk media errors while recover the data from fault zone. REM effectively avoids lengthy process of data rebuilding after disk failure. In addition, the system's faults not really happen in the process of data migration. So it can effectively avoid the complex checksum computation of the rebuild process and greatly reduce the impact on system performances. REM can dynamically adjust the speed of data migrate according to the system I/O load to reduce the impact on entire system's performance.

II. THE KEY TECHNOLOGY OF REM

A. Early-warning technology

REM introduces early-warning technology into RAID. It predicts disk and array failure through disk health degree model and array health model.

The disk health degree model is based on SMART technology. There are two health degrees: Parameter Health Degree (PHD) and Disk Health Degree (DHD). The PHD is calculated by the data (DT), threshold (TH) and attribute value (AV) of disk SMART parameters. AV has been set to the maximum normal value as default. TH is the fault limit value set by the manufacturers. Users monitor the DT of the SMART parameter in regular time. The formula of parameter health degree is PHD=100*(DT-TH)/(AV-TH). For this formula, the value of AV and TH does not change, but the value of DT change all the time, meanwhile the value of AV and DT can compare with the value of TH at the same time, so AV and TH as the denominator, DT and TH as numerator.

DHD is calculated by the weight of a single parameter and PHD. According to analysing the weight of SMART parameters and the test results of actual data have changed of many disks (It cannot be listed here due to limited space), the five parameters of RSC (Reallocated Sectors Count), SRC (Spin-up Retry Count), RER (Raw Read Error Rate), PCC (Power Cycle Count) and SER (Seek Error Rate) have greater influence to the disk failure and their DT can be compared with the TH. We confirm the RSC and SRC weight are 20%, RER weight is 40%, PCC weight is 10% and SER weight is 10%. So we have DHD=0.2RSC+0.2SRC+0.4RER+0.1PCC+0.1SER.

In order to evaluate the reliability of disk, we compare the DHD with Disk health degree Early-warning Threshold (DET). It determined by two factors:

1) In five SMART parameters, the number of parameters whose DT has changed.

If the DT of disk SMART parameter has changed, it indicates this parameter has negative impact on disk health degree, if the DT of a few disk SMART parameters has simultaneous changed, it indicates these parameters have greater negative impact on disk health degree.

2) The weight of parameters whose DT has changed.

The DT change of parameters has impact on DHD, and the weight of DT changed parameters also has impact on DHD. When the weight is greater and DT of parameter has changed, it indicates this parameter has greater negative impact on DHD, because of the weight of RER parameter is great, the DT change of RER parameter leads to obvious reducing of DHD.

According to the number of DT changed parameters and DHD calculation formula (the weight of DT changed parameters); the DET can be calculated. We take only one parameter change as an example to introduce the calculation of DET. When only the DT of RSC has changed, and the PHD of RSC reduces from 100 to 0, others are 100 (Table I). According to the calculation formula of DHD, when the DT of RSC has changed from 100 to 36, others are 100. So, when the weight is 20, the DET of RSC and SRC are 0.2*0+(0.2+0.4+0.1+0.1)*100=80; when the weight is 10, the DET of PCC and SER are 0.1*0+(0.2+0.2+0.4+0.1)*100=90; when the weight is 40, the DET of RER is 0.4*0+(0.2+0.2+0.1+0.1)*100=60.

The array's health condition is based on the state of the disk health, the state of array, temperature and the fan speed in the array. According to the state of array and environment, if the temperature is high, the fan speed is fast, or the voltage is higher than normal, then the early-warning information should be given immediately and check whether the data migrate should be started or not. If the situation above is not happened, the system will check the DHD. If only one DHD is lower than DET, then the system will find out whether there is a spare disk, if spare disk exists, then there is no need to start array migrate, otherwise the array migrate should be started.

B. Data Migration Technology

When the early-warning information has been given, REM will copy data directly to an appropriate node and location. Data migration is divided into the following three levels: sector level, disk level, array level. When detecting these informations, REM can utilize disk self-healing; disk migration and array self-healing to protect system, as shown in Figure 1.

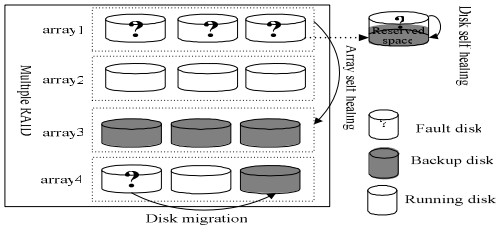


Figure 1. Data migration technology.

Disk self-healing tries to repair part of disk media errors while recovering the data. REM reserves a part of space for each disk to repair part of media defective. When the disk has bad sectors, the part of sectors from the reserved space can be allocated to replace the bad sectors and store the mapping information in the address mapping table. When following disk access bad sectors, the operation will be redirected to the reserved sectors. The technology is shown as Figure 2.

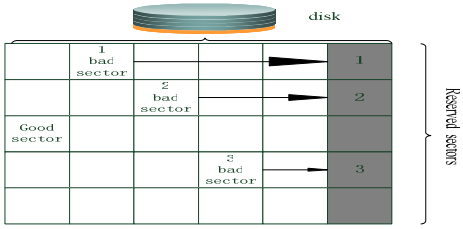


Figure 2. Disk self-healing technology.

When DHD can't meet the requirements, REM will copy the data of fault disks directly to other spare disks. While data migration is going on, system can quickly copy the whole fault disk to other free disks. For the data of sectors whose media error occurs, the way of data checking can be used to recover the original data, as shown in Figure 3.

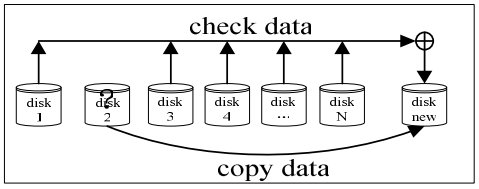


Figure 3. Disk migration technology.

When the health of more than one disk array is poor, data errors may occur at any time, it must commit data migration as soon as fast. If hot spare disks of safety and healthy in multiple arrays storage system is enough, the array self-healing can be completed, which use the remaining hot spare disks to copy the early-warning fault array completely and modify the array mapping table and update the configuration information. The actual data migration theory of array self-healing technology is exactly the same as disk migration technology, except that it involves multiple disks.

III. EVALUATION METHODOLOGY

A. *Experimental settings*

The prototype of REM system is installed on an iSCSI storage server. Storage clients are connected to the storage server using the Cisco 3750 Gb Ethernet. The hardware and software details of REM system are listed in Table I.

TABLE I. HARDWARE AND SOFTWARE OF REM 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

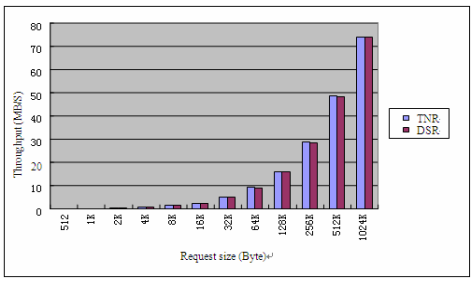
B. *Numerical results and discussions*

The results are listed in Table II, which presents the testing results of nine disks in the disk array. The results indicate the case of RER and SER changed (the DT changed of PCC is only 1, so PCC is out of consider), and disk health degree early-warning threshold is 50. So, the 9 disks are normal and fulfill the actual usage. But the health of disk 1 is the lowest and has less reliability; the health of disk 3 and 6 are higher and have better reliability. From the testing results in table 5, if the disk health degree approaches to or less than 50, the fault can be predicted and the early-warning information should be given immediately. According to early-warning, system can migrate data from imminent faulty location to safety area, thus corresponding data migration can be carried on.

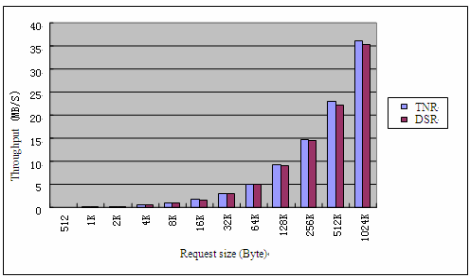
TABLE II. THE TEST RESULTS OF DHD

Disks number	RSC			SRC			RER			PCC			SER			
	DT	TH	AV	DT	TH	AV	DT	TH	AV	DT	TH	AV	DT	TH	AV	**DHD**
D_1	100	36	100	100	97	100	40	6	200	99	20	100	47	30	100	**59**
D_2	100	36	100	100	97	100	117	6	200	99	20	100	51	30	100	**75**
D_3	100	36	100	100	97	100	117	6	200	99	20	100	85	30	100	**80**
D_4	100	36	100	100	97	100	114	6	200	100	20	100	84	30	100	**79**
D_5	100	36	100	100	97	100	105	6	200	99	20	100	58	30	100	**74**
D_6	100	36	100	100	97	100	118	6	200	99	20	100	90	30	100	**81**
D_7	100	36	100	100	97	100	113	6	200	99	20	100	78	30	100	**78**
D_8	100	36	100	100	97	100	117	6	200	99	20	100	67	30	100	**77**
D_9	100	36	100	100	97	100	53	6	200	99	20	100	65	30	100	**64**

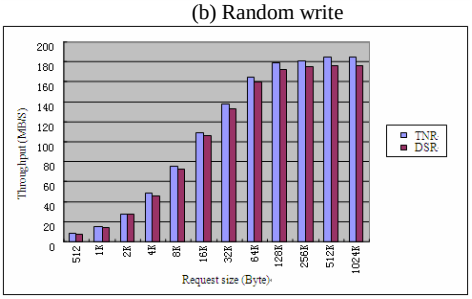
The evaluation of early-warning provides ground for data migration. Our evaluation test, by contrast, gives four state of the RAID5 system: Traditional Normal RAID (TNR), Traditional Rebuild RAID (TRR), Disk Self-healing RAID (DSR), and Disk Migration RAID (DMR). We give the throuthput perfermance of this four structure using Iometer from the view of the request size.



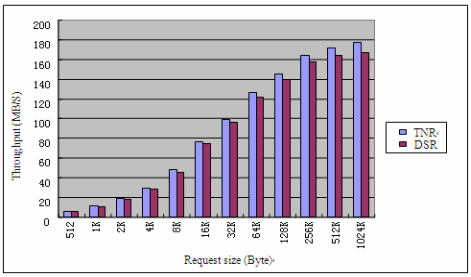
(a) Random read



(b) Random write



(c) Sequential read



(d) Sequential write

Figure 4. The throughput of TNR and DSR systems with Iometer.

Figure 4 shows the throughput of the two states of TNR and DSR, which measure many type of request size with random and sequential access. As we have expected, the throughput of DSR is always close to the TNR, which means there is little or no impact on the system performance in disk self-healing, The main reason is added some I/O agents, the execution of a I/O request to carry more instructions, and each request need the circulation find out whether exist bad sectors, but the influence is within acceptable limits.

IO performance: Figure 5 shows the throughput of the two states of TNR and DMR systems, which measure many type of request size with sequential read access.

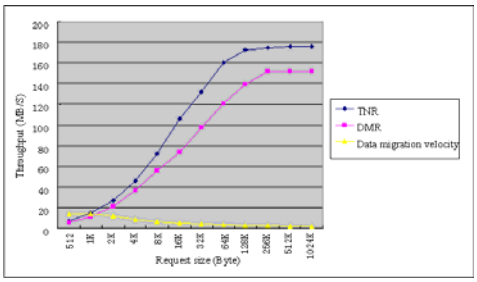


Figure 5. The throughput of TNR and DMR systems with Iometer.

Disk migration leads to lots of disk I/O, meanwhile migration threads are started from background. Its execution speed is closely related with the CPU occupancy rate, when it performs too fast, which make the CPU become the bottleneck of system performance, if the I/O speed of the disk is too fast, it might seriously influence performances. Dynamic adjusting migration velocity is to reduce the impact on system performances. From the yellow line can be seen as system I/O workload rise, migration velocity would also automatically reduce the impact on system normally I/O performance. Although data migration has some impact on system normally I/O performance, compared with the TRR system (Table III), the influence of system are much smaller. The analysis above is based on the sequential tests, the random test results are not given, which are the same as the sequential tests.

TABLE III. 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

The throughput of TNR, DSR, DMR and TRR are compared, it use sequential read and request size is set to 1024K. TNR throughput can approach to or exceed 180MB/s, DSR throughput can approach to or exceed 160MB/s, DMR throughput can approach to or exceed 140MB/s, and TRR throughput is only 50MB/s. The throughput of DSR、DMR and TRR respectively reduced 11%、22% and 72% than TNR, and the throughput of DSR and DMR respectively increased 220% and 180% than TRR. So rebuild RAID system performance is poor than traditional normal RAID system, but the performance of disk self-healing and disk migration system are almost similar with the performance of normal RAID system, and it obviously higher than the performance of rebuild RAID system. The sequential write, random read-write and the results of other request size are not mentioned here, which are the same as the situation above.

IV. CONCLUSION

This paper proposes a new RAID system to enhance the reliability of the RAID storage system, which called REM using early-warning and data migration technologies, according to the running state of RAID, early-warning technology can effectively predict system fault in advance, on one hand, REM system can reduce unnecessary long-term backup time and shorten backup window; on the other hand, it can greatly reduce the risk of data loss. Once the fault has been predicted, the early-warning information would be given immediately, REM can utilize disk self-healing technology, disk migration technology and array self-healing technology to remove data from imminent faulty location to an appropriate node and location. No matter which kind of data migration technology, the performance of REM system substantially outperforms traditional rebuild RAID system.

ACKNOWLEDGMENT

This work is sponsored in part by 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, and the and Youth Chenguang Program of Wuhan under Grant No.20 1050231073, and the HUST Fund under Grant No.2011QN053 and No.2011QN032, and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] D. A. Patterson, G. Gibson, and R. H. Katz, “A case for redundant arrays of inexpensive disks (RAID),” Proc of the 1988 ACM SIGMOD international conference on Management of data, Chicago, Illinois, United States: ACM, 1988, pp.109-116.

[2] J. Zhang, “Method for backing up and recovering data in a hard disk,” IEEE. T. COMPUT, vol. 3, pp. 112–115, 2005.

[3] Q. Xin, E. L. Miller, T. Schwarz, D. D. E. Long, S. A. Brandt, and W. Litwin, “Reliability Mechanisms for Very Large Storage Systems,” Proc of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies, San Diego, CA: IEEE, 2003, pp. 146-156.

[4] G. F. Hughes, J. F. Murray, K. Kreutz-Delgado, and C. Elkan, “Improved disk-drive failure warnings,” IEEE. T. RELIAB, Vol. 51, pp. 350–357, 2002.

[5] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, “Hard drive failure prediction using non-parametric statistical methods,” Proc. International

Conference on Artificial Neural Networks (ICANN/ICONIP 03), Springer, 2003.

[6] G. Hamerly, and C. Elkan, “Bayesian approaches to failure prediction for disk drives,” Proc. International Conference on Machine Learning (ICML 01), IMLS Press, 2001, pp. 202-209.

[7] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, “Machine learning methods for predicting failures in hard drives: A multiple-instance application.” J. MACH. LEARN RES, Vol. 6, pp. 783–816, 2005.

[8] J. Wilkes, R. Golding, C. Staelin, and T. Sullivan, “The HP AutoRAID hierarchical storage system,” ACM. T. COMPUT. SYST , vol. 14, pp. 108–136, 1996.