IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Quality Analysis of Storage Drives

Rakesh S. Surve[1*], Vikas N. Honmane[2]

[1,2] Department of Computer Science, Walchand College of Engineering, Sangli, India

[*]*Corresponding Author: rakeshsurve15@gmail.com, Tel.: +91 9730230327*

DOI: https://doi.org/10.26438/ijcse/v7i6.287290 | Available online at: www.ijcseonline.org

Accepted: 10/Jun/2019, Published: 30/Jun/2019

Abstract— Quality Analysis of Storage Drives is an important thing to be done before actual use of drives in the market. Finding the failure reason of drives and reporting it in an automated way is a challenging work to do. There is no direct solution available to find the actual cause of failure, the cause of failure can be related to drives, firmware, or testing environment. So using various techniques such as drive analysis based upon various parameters of hard disk drive we can predict the cause of failure. Storage drives are important component of any computing devices which needs to be analyzed using different techniques. This paper is an attempt to explore the quality analysis of storage drives.

Keywords— Quality Analysis, Drives Security, Performance, Drive analysis.

I. Introduction

As the amount of data is growing day by day the size of Storage media is also increasing rapidly. Various new technologies like Cloud computing, Flash Memory are used in many advance devices. Storage Media are used everywhere right from compact devices to large Data Centers. Drives are basically used for storing the data, as the devices are getting compact the size of Data storing Drive is getting compact. There is an advance development in Drives and Storing Technology. Drives such as Hard Disk Drive (HDD), Solid State Drives (SSD) and Serial Attached Technology Attachment (SATA) have unique features in terms of Storage, Performance, and Speed. Quality analysis of drives include to test various parameters related to its disk stress, IO operations, compatibility, operating System which needs to test against each kind of drive before it comes in actual use. Each drive has its own advantages and disadvantages. Various machine learning techniques can be applied for failure analysis of drives.

Using machine learning techniques we can analysis different parameters such as disk stress, I/O operations, operating system compatibility. Different techniques can be applied such as Logs analysis.

Heuristic approach can be defined as there is no proper way to do drive analysis. Heuristics include various artificial intelligence techniques, using these techniques we can use to do analysis of drives. Some of the technique has been used and third party software's for analysis of drives but there is limit we can't do full analysis and the results are not accurate. For these there is need to define a new technique which includes machine learning approach, testing approach, Firmware based approach. Using these techniques we can do analysis and the accuracy for drive analysis can be achieved. Reliability demonstration test of drives is an important thing to do it includes Hardware, software test, acceleration factor. Other parameters need to be tested for failure of Drives like Permanent Failure of Drives, Drive Endurance, Read Disturb and Data Retention.

Parameters which can be used for prediction of failure in hard disk drive are S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology). Attributes using this attributes we can predict the failure of Hard disk drive. Test cases can be designed for device, storage, input /output, and endurance which will have different inputs and some dependent on each other and with these test cases we can predict the nature of hard disk drive.

II. Related Work

This section discusses various research papers referenced for quality analysis of drives.

In [1] the author has stated Storage and memory technology of SATA, SAS and Flash Memories. It shows the development structure of how the storage media has grown in past few years. Hard disk drive has been played an important role in mass data storage, as it is cost effective, non-volatile, and compatible with system. Flash memory has been used on large scale in new storage media and in new applications. Response time it is nothing but input/output operations performed at a given single time, the system whose response time is fast are generally known for its resource utilization. Touch Rate it can be defined as the total system capacity which can be assessed in a given interval of time.Using above two parameters we can do quality analysis of the drives. The relation between response time and touch rate is the response time measures the time on a one object while touch rate can be measures on whole dataset as whole.

Issues related to Solid state drive (SSD) stated in [2] SSD contains memories and microcontroller along with direct current to direct current convertors, temperature sensors, filters capacitor, DDR (Double Data Rate) memory for fast data transfer. The increase number of NAND memory for data transfer and how to optimize SSD is an issue. Another issue which is highlighted is SSD data recovery and error classification for reliability and endurance of SSD. In [3] the author has described The Device level validation of Drives which are used for Storage and real time applications. SATA IF protocol test, CAM (Common Access Mode) Test, Feature Test, Performance Test. Connect HBA to System drive for boot Operating System. Drive Master is used as interface for the command and data transfer. Another BusXpert Protocol analyzer is used to track the process whether the command is working properly or not, how the data is transfer from System to Drive. CAM testing includes all the addressing modes connected to devices. Test conducted for 1000 loops for each addressing mode Feature Tests like Hard Reset Test, Soft Reset Test, and Power management. Performance Testing includes Boot Time Simulation Read/write operations, Flush time and Seek time these parameters are used for performance testing of SSD Advantage of the above method is they have test different features. Disadvantage is they have used prosperity tool and the test which are conducted are limited. In [4] the author has stated the approach reliability of SSD. In [5] the author has discussed various security features about SSD. In [6] the author has raised the condition in data centers why the storage media fails. In [7] the author has discussed the vulnerabilities of USB (Universal serial bus) and attacks on USB. In [8] the author has discussed about the security issues of servers.

III. METHODOLOGY

Quality Analysis of Drives and certification needs to be carried out in automated environment. For which we need to design automated environment and attach drives for Quality Analysis of Drives and certification. The automated environment includes all types of tests. The test framework is made up of many framework APIs which perform basic security operations via interface commands. These framework APIs are written in the python language. The test scripts (which are also written in Python), coupled with the framework API's together form the automated Environment. To start executing the test scripts, there are a couple of inputs which must be provided to the test framework in order for it to work properly. Develop an automated framework for finding the failure reason. Another approach which can be followed is Disk reformat, many times normal formatting of the disk doesn't work need to do it using the firmware based method, which needs to have firmware knowledge.

A. Heuristic Approach

Machine learning techniques can be applied on different drives for analysis; Algorithm such as random forest, Convolutional neural network and support vector machine can be used. Random forest algorithm divides the data into number of different forest and the working of random forest it gives the best value after classification using convolutional neural network we can train the machine for drives failure. As we know that Convolutional neural network has input, hidden layers and classification layers. Input will be test cases Drives analysis then in hidden layer a function based on parameters received in input which will be used to classify the failure reason. Many times the drives get failed due to improper formatting or system creates its own partition while testing the drives so therefore we need to follow another approach.

Supervised learning technique can use for drive analysis as we have the input label as the logs which can be used to Learn the pattern of on what condition for what parameters the Drives behaves. Using the supervised and Behavioral analysis we can define the accurate reason of the drive failure and at what condition. Using both techniques we can predict the performance of the drive.

B. Firmware Based Approach

Another approach which can be followed is Disk reformat, many times normal formatting of the disk doesn't work, need to do it using the firmware based method, which needs to have firmware knowledge. It includes Firmware setup which includes a controller, Drive, software like Python, java. Using the above setup we can reformat the drive and do analysis. This approach will clean the drive and make it to default firmware mode. Where no system generated files are created only the drive include some binary files. Disk Flush can be done using the Firmware Based method. It cleans the memory from the drive and system. The command is send from controller to the Firmware based system. The python code can be used to get the analysis of the Disk Flush.

C. Black Box Approach

Black box approach is generally used in software testing. Black box testing is nothing but behavioural testing; this approach can be used for analysis of the drives.

Techniques which can be used under black box approach are as follows: Equivalence partitioning is nothing but an approach to divide the input values into two types that is valid and invalid partitions and testing each parameter of drives. Boundary value analysis is an important approach for verifying whether the test cases fails at particular conditions. This approach can be used to measure Disk Stress of Drives, Disk Stress can be defined in terms of how the disk behaves at particular condition, whether it works properly or not at high stress is important task for quality analysis of drives. Various parameter need to check for disk stress against each drive and against each machine depending upon certain conditions. Disk stress includes the amount of input output operations performed in a given time, for this we can go with the boundary analysis technique for disk stress of each Drive.

The overall process of analysis of drives cannot be done using single method, combining them and applying them together can give us appropriate result. The Analysis of the Hard disk drive can be done effectively by combining above three methods described. Tools like python framework, flask and machine learning algorithms can be used effectively for prediction of disk Failure.

Storage drive such as Solid state drives can be analysed using Heuristic approach where the S.M.A.R.T attributes of Solid state drive can be used for failure analysis. As we know that solid state drive works on flash memory the failure analysis of solid state drive need to be done on the endurance of the flash memory cell and the controller which is used to store the data on solid state drive.

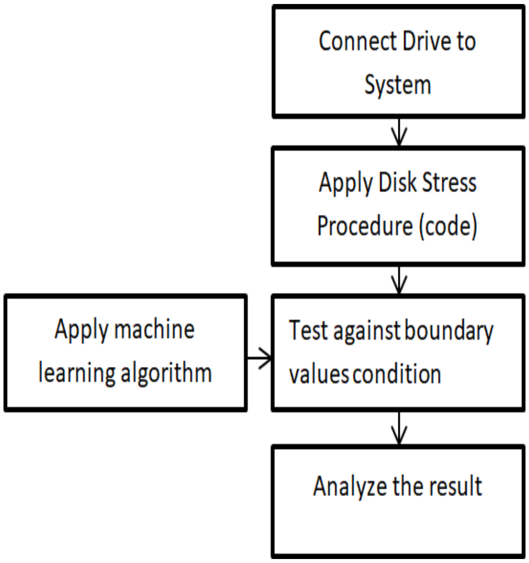
Connect Drive to System
Apply Disk Stress Procedure (code)
Apply machine learning algorithm
Test against boundary values condition
Analyze the result

Figure 1. Disk Stress Analysis

As we can see the Drives are connected to appropriate system based upon the configuration need. As we know that the each drive has its own properties if we take hard disk drive then the size of data storage is less as compared to SSD or SATA. SAS drive has more Flash NAND memories as compared to other drive so the input/output operations performed are faster on SSD. Various servers have different controllers designed so before going to start the process make sure you have connected the controller with the right machine. Disk Verification includes Verifying the disk against various random input output operations. This can be done using python framework and tensor flow API and machine learning algorithm and defining conditions for Disk Verification. Many times the result of Disk Verification changes as per the System and Data configuration, so there is need to analysis the disk on the standard parameters. Figure 1 shows the flowchart of the disk stress analysis. Multiport drive test is used to check the endurance and commands based behaviour on specific port for communications between the system and drives. Combination of Python framework and firmware based hardware we can do analysis of the drives.

IV. Results and Discussion

The Analysis of drives is done on the parameters which are SMART (Self, Monitoring, Analysis, Reporting Technology) Attributes of hard disk drives. This S.M.A.R.T. Attributes are based on Read error rate, Spin up time, Power on hours, Reported uncorrectable errors, Command timeout, and reallocation event count. Figure 2 represents the ROC (Receiver operating characteristic) curve with AUC (Area under curve) 0.83 and Figure 3 represents ROC curve with AUC 0.78.

Using these Parameters we have applied the random forest algorithm on the Data Set Provided by BackBlaze, which includes S.M.A.R.T. values for classification of Hard disk drive failure.

The Data Set is used for prediction of failure of hard disk drive is form BackBlaze. They take data of each Drive present in market in their data center; this data includes S.M.A.R.T Attributes of hard drives. The results which we get after applying random forest algorithm are shown below. S.M.A.R.T. attributes have higher correlation to disk failure. Random forest is an ensemble tool which classifies the data based upon subset of observations and subset of variables to build group of decision trees.

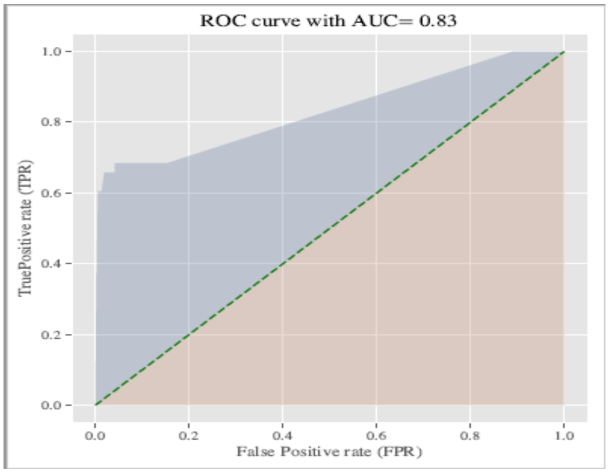

Figure 2. Results after Subsampling

True positive rate: 0.6052631578947368
False positive rate: 0.10849056603773585
Fraction of failed disks after subsampling: 0.4
Fraction of working disks after subsampling: 0.599
Working: 109 | failed: 73

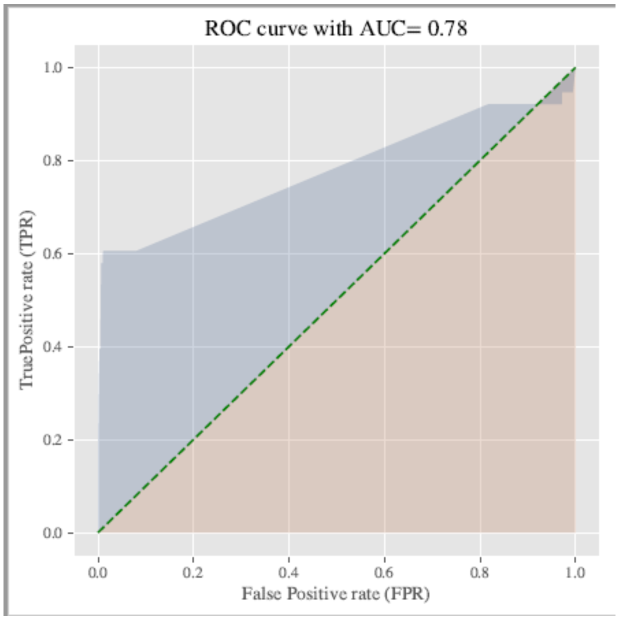

Figure 3. Results after Oversampling

True positive rate: 0.6052631578947368
False positive rate: 0.0942622950819672
Fraction of failed disks after oversampling: 0.4
Fraction of working disks after oversampling: 0.6
Working: 41011 | failed: 27340

Here we can see the number of working disk and failure disk based upon the important parameters which we have used for subsampling and oversampling using random forest technique. Using feature extraction nine important features have been extracted from data set. Number of trees used are 200 maximum features used is 9 maximum depth of the trees used is 10 and minimum samples leaf is 5. The Dataset have data of 68,433 Hard disk drive data. We have applied subsampling and oversampling to the data set as the data is imbalanced, we got the results as 0.83 and 0.78 as data correctly classified.

V. Conclusion and Future Scope

Quality analysis of storage drives needs to be done to predict the failure of hard disk drive. There are issues like Disk Stress, Disk Flush, Boot Test, Disk verification need to be done as this are not related to actual drive rather they are dependent on many aspects which are connected to it. So discussed above approaches can be used to Quality analysis of storage Drives consisting of machine learning technique, black box testing, and firmware testing.

Acknowledgment

We would like to thank all the authors for inspiring research oriented studies by their publications. We are thankful to all the authors whose papers are referred for preparing this article.

References

[1] Tom Coughlin, Roger Hoyt, and Jim Handy ”Digital Storage and Memory Technology (Part 1)” IEEE Advancing Technology for humanity.(2017)

[2] RINO MICHELONI, Senior Member IEEE”SCANNING THE ISSUE Solid-State Drives (SSDs)”(2017)

[3] Raja Subramani, BharathRadhakrishnan, Krishnamurthy Puttaiah(2013) “Complete Device Level Validation of Solid State Flash Drives – An Approach” IEEE 15th International Conference on Computer Modelling and Simulation

[4] Nematollah Bidokhti “SSD Next Gen RDT” Annual Reliability and Maintainability Symposium.(2016)

[5] “Data Security Features for SSDs” A MICRON WHITE PAPER

[6] Iyswarya Narayanan, Di Wang, Myeongjae Jeon, Bikash Sharma, Laura Caulfield, AnandSivasubramaniam, Ben Cutler, Jie Liu, BadriddineKhessib, KushagraVaid.(2016) “SSD Failures in Datacenters: What? When? and Why?” ACM 978-1- 4503-4381-7/16/06.

[7] Dongho Won (2007) ”Vulnerability Analysis of Secure USB Flash Drives” IEEE

[8] Sonali Patra, N C Naveen, Omkar Prabhakar(2016) “An AutomatedApproach For Mitigating Server Security Issues” IEEE International Conference On Recent Trends In Electronics Information Communication Technology

Authors Profile

Mr. Rakesh S. Surve pursuing Master of Technology from Walchand College of Engineering, Sangli (Maharashtra) in Computer Science and Engineering Domain. His research domain is Storage Media and Machine learning.

Mr Vikas N. Honmane pursed Master Of Technology from College of Engineering Pune and currently working as Assistant Professor in Department of Computer Scinece and Engineering Sangli, His research area is Machine Learning and Deep Learning.