Methodology for Usage of Emerging Disk to Ameliorate Hybrid Storage Clouds

Sandeep R Patil, Riyazahamad M Shiraguppi, Bhushan P. Jain, Sasikanth Eda

IBM India Software Labs, Pune, India

{sandeep.patil, riyaz.shiraguppi,sasikanth.eda}@in.ibm.com

bpjain@cs.stonybrook.edu

***Abstract*--With the dramatic evolution of various greener disk and memory technologies, helped in rapid establishment of Hybrid storage environment comprising of heterogeneous storage units. In order to provide an efficient and optimised storage solution there exists a necessity of adapting smarter changes in the management stack that enables compact sensing, processing, decision making capability based on the importance of data and disk life span predicted on operational workloads. This paper reviews system architecture for two faces of RAS cloud features namely disaster recovery planning, disk breakage prediction independent of disk technology and host aware data tier based on disk life span.**

***Index Terms*— Hybrid, disk, life span, prediction, weight, curve fitting, S.M.A.R.T parameters, SSD, disaster recovery.**

I. INTRODUCTION

Storage in general and disks in particular are the driving force for all Information Technology (IT) enabled business. Since the storage system hosts the entire business data, it is vital to have these systems with state of the art storage characteristics. Further, characteristics like performance, reliability, backup, availability and data protection of storage subsystems like disks directly impact on the effective execution of an IT enabled business. Hence generally, significant IT budget is allocated for ever growing storage needs of a business. Along with others, this includes cost involved in;

- Upgrade or replacement of storage subsystems like disk replacement to newer disk technology.
- Improved disaster recovery technologies.
- Adherence to changing business and legal requirements.

Newer Disk Technologies viz., Solid State Disks (SSD) and memory technologies like Phase Change and Race Track [1] offer great promise and unique characteristics typical for their use in Storage Cloud. Moreover since clouds span across different geographies, it adds newer elements to be considered in cost efficiency. This paper discusses a new method for disk breakage prediction independent of disk technology on operating workloads and host aware data tiering of storage systems. This will help to optimize the existing methodologies by taking into consideration the advantages of various disk technologies as well as different cost factors.

II. HYBRID STORAGE AND STORAGE CLOUDS

Cloud computing refers to a computing platform that is able to dynamically provide, configure, and reconfigure technology infrastructure to address a wide range of dynamic needs. There are various clouds addressing specific requirements like high performance computing, image or video processing. The amount of unstructured data is and will continue to increase exponentially due to astronomical data generated from video, audio, graphics and web applications. Storage Clouds are one of the most popular storage systems in the enterprise world.

The new evolving disk technologies like Phase Change Memory, Racetrack memory and Flash Class storage memory will potentially be used in providing more innovative solutions. In the light of so much research and advances in the disk technology, a complete restructuring of the backend storage to replace the older technology disks with newer technology ones is neither economical nor is practically feasible. It is imperative that multiple generation of disk technologies co-exist in the same storage system for effective operation and low cost of ownership, giving us the ground for Hybrid Storage. Thus we have a hybrid storage structure formed in which all the different disk & memory technologies are available and we must leverage this hybrid architecture.

Hybrid Storage is predicted to be very apparent with coming times. Building systems with the use of hybrid storage systems collectively is known as storage cloud with hybrid storage devices.

III. RELATED WORK: NESCIENT OVER HYBRID DISK TECHNOLOGY

Current disaster recovery systems [2] and disk replacement methodologies do not consider the advantage of the underlying disk technology and cost budgeting based on data stored during the operational process. There is a strong need for such systems to do so to derive the much needed growth in performance and to meet the challenges of being greener.

A. Pending Disaster Recovery

There are various backup techniques for disaster recovery like Hot Backup which maintains an additional site, operating in parallel to the main installation; Warm Backup which maintains another inactive site ready to become operational

978-1-4673-4422-7/12/$31.00 ©2012 IEEE

within a matter of hours; Split Site where two installations are used and in case of emergency one centre can keep the organization running by performing only jobs with high priority. Other techniques include Cold Backup which maintains empty computer premises ("empty shell") with the relevant support ready to accept the immediate installation of appropriate hardware.

In a disaster prone environment or system, data in danger is the data on storage/systems which have been recently processed at the site and whose replication to the remote backup site has not yet taken place. There are various techniques to tackle pending disaster which automatically trigger replication/backup activity of systems and storages on receiving an event of possible disaster, either through fire, smoke, earthquake, water sensors attached to the system or via external systems like forecast news etc. But these existing methods lack on how effectively the data in danger can be backed up/replicated such that the efforts to backup maximum amount of data is at its best taking under consideration disks of various technologies available.

B. *Disk Replacement*

Traditionally, industry follows the disk replacement strategy of replacing problematic disks. The schemes phase out old disks and replace them with newer technology disk. The tools that are used typically rely on failure factors reported by disk's S.M.A.R.T reports. While this is an acceptable approach, it is not a smart and business effective approach. The existing methods and tools used to strategize disk replacement polices for data centres have started to appear primitive with the advent of newer disk technologies like Flash, Phase Change technology and Racetrack memory using spintronic science which will continue its advent.

Ideal solution is to replace all the disks with newer technology disks, but the cost expense prohibits this. Since regular performance of disk directly impacts on the business costs (thru power consumption/ heat dissipation/Carbon credit utilization), its vital to have a method that can be deployed in tools which will help to strategize and identify right disks for replacement in a geographically spread cloud environment where data is replicated across disks which are located in different parts of the world.

IV. PROPOSED TECHNIQUE FOR AMELIOLERATED BEHAVIOR

The following methodologies explain how we can utilize the various evolving disk technologies to ensure a smarter and cost effective solution.

A. *Proposed Pending Disaster Recovery*

Hybrid Storage is very apparent due to various disk technologies. Based on the above facts, the paper proposes a mathematical model to help design and develop methodologies that can be used to enhance the disaster recovery systems using hybrid storage systems.

When a storage system receives an event of possible disaster, it immediately initiates the switch to its replicated server located in a different site. This dictates that the server is configured to do replication with appropriate RPO (Recovery Point Objective) at an acceptable RTO (Recovery Time Objective). This replication/backing up process is enhanced with the following:

1) Disk technology based disaster recovery: The offshore replication/backup system takes into account the technology of the underlying disk on which the data resides and its characteristics during the backup operation especially during a disaster. Since typically SSD tends to have a better data access rate (data seek time) than HDD, the amount of data that can be read at a given time will be much more from a SSD as compared to HDD. Thus scheduling data on SSD for an early backup helps backup maximum amount of data, especially when an unplanned emergency or disaster occurs and when the total amount of available time is unknown. In hybrid storage environment, the system identifies the non backed up data residing on SSD disk and schedule its backup before data on HDD during a disaster. Thus we minimize the risk impact by following this greedy method for backing up data. Similarly the backup storage system at the remote site where the data is being backed-up, automatically takes into consideration the technology of the underlying disk (on which the data is being backed up) and its characteristics into account during backing up data from a site which has hit a disaster. It ensures that the disk with the fastest write cycle is selected for backing up the data. This helps confirm that best efforts are made to backup maximum data from the disaster hit site.

2) Disaster recovery by Change page replication: Change-page-threshold [3] is a parameter that indicates percent of pages that are dirty after which the dirty pages be will be flushed to disk from RAM. Higher the value of this parameter, more are the pages in main memory which are not yet externalized to the remote storage. Hence in case of natural disaster (Cyclone/fire/earthquake) there is a great chance that this data will be lost. The system automatically changes the Change-page-threshold parameter to a lower value so that data in main memory (RAM) is saved and makes sure that the disk it is being committed to is an SSD in case of hybrid storage as shown in Fig. 1. Subsequently the above proposed steps will help towards greedy backup to minimize the risk.

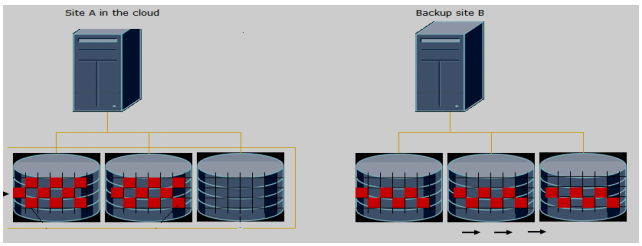

Fig. 1 pictorial representation of disaster recovery between two sites A and B

The paper proposes the following steps to maximize the atomic data reads and pedigree of physical apparatus considering the constraints of total electricity left and the time for disaster.

- Calculate total electricity left and total data left to be read. Then read data from disks such as SSD's first thereby consuming less electricity and catering to faster needs. Avoid reading huge files or contiguous data blocks since they would be copied only partially, it is better to copy other files which can be copied in full than partial data. The above process takes into consideration hours left for electricity supply to end while starting the copy operation to maximise atomic reads and full file backup.
- Read from more robust disks since while backing up data in emergency due to heat generated disk is susceptible to fail.
- Read and write to physical apparatus which is lubricated well
- Calculate which disks and other physical apparatus are more worn out and copy from them at end.

B. *Proposed Disk Replacement*

This paper proposes a methodology for disk replacement with the use of calculating the devices which needs to be replaced in priority based on the mathematical model which will be derived from device characteristics. The proposed technique analyses and mathematically forms weight to each operational disk with the factors and presents a sorted view of most probable disk to replace with a newer technology disk to ensure a positive influence on the business ROI , efficiency and aid in smarter, realistic disk replacement strategy.

The mechanism is to aid for smarter disk replacement strategy which is more relevant to the upcoming trends of Storage Clouds spread across geographies, Carbon Credit management with its impact on business and the changing trend in underlying storage technology (from Tape to Flash and flash transformed to Phase Change or Racetrack – in future). The methodology can be incorporated as a consumable tool for storages.

Assuming that measurement of the stated factors like power consumption, head dissipation, carbon credit utilization, etc per disk can be easily determined; detailed mathematical calculation procedure of the disk replacement result is explained as follows:

1) *Forming a mathematical variable for individual disk behaviour:* Consider vendors specifications that are provided in the data sheet as the initial observations of a disk. Each disk's S.M.A.R.T parameters [4] are equated to a set variable. Suppose the provided disk is a Hard Disk Drive, vendor specifications are collected and equated to the variable "SP" (Note all the S.M.A.R.T specifications that are provided by the vendor for a particular HDD as the initial reading).
SP= {Head flying flight, Data throughput performance, Spin up time, Reallocated sector count, Seek error rate, Seek time performance, Spin try recount, Drive calibrations retry count }

Similarly if the provided disk is a Solid Disk Drive, vendor specifications [4] are collected and equated to a variable "SD" (Note all the S.M.A.R.T specifications that are provided by the vendor for a particular SDD as the initial reading).

SP= {Power management, Latency specifications, Random Read/Write Input/output operations per second, Electrical characteristics, Altitude, Electromagnetic immunity, Shock and vibration}

In a heterogeneous cloud data centre, the S.M.A.R.T parameters of individual disks can be equated to variable Xi (i=0, 1, 2, 3… n).

After installing the disks in a production environment, the disk behaviour differs from the ideal specifications provided by disk manufacturer. The S.M.A.R.T parameter differences exhibited by the examining each individual disk is noted as differential variable (Δx). The value of this may contain either positive or negative tolerances.

2) *Considering the Environmental Disturbances at production environment:* Environmental deviations are noted between the default vendor specified temperatures to the field temperature. Here environmental noise [5] can assumed to be largely an accumulation of Thermal noise.

3) *Forming a mathematical equation for individual disk behaviour:* A minimum of 100 hours of observation time is considered for all the disk parameters. Assumption at this state is that there are no impulse I/O load changes, environmental issues, and electrical characteristics. The noted S.M.A.R.T values per disk (initial, installed, 100 hour readings) are plotted on to a graphical sheet and traditional normalization techniques are applied on each S.M.A.R.T parameter plot. The plotted curve can appear as anyone or combination of the following curves; exponential decay curve ($y=Ae^{-Kx}$), exponential raise curve ($y=Ae^{Kx}$), linear form ($y=Ax+B$), parabolic nature ($y^2=4ax$) and sum of exponential curves ($y=Ae^{ax}+Be^{bx}$). Curve fitting mechanism named "least squares" [6] is applied on the plotted readings and readings are normalised to form a curve.

4) *Example:* consider the values of time in hours vs. SP are noted as (2,1.8) (4, 1.5) (6,1.4) (8,1.1) (10,1.1) (12,0.9) the tabulated values are as shown in Fig: 2

Time (hours)	SP (SMART parameter)
2	1.8
4	1.5
6	1.4
8	1.1
10	1.1
12	0.9

Table-1 : Noted values for Time (hours) on x-axis and S.M.A.R.T parameters on y-axis

The plotted curve for the above tabulated values follow a decay and numerical method technique is applied on them to achieve the curve equation.

The solved curve equation is given as;

$$0.936^{SP} = (T/2.013) \quad (1)$$

where T denotes time in hours and SP denotes SMART parameter value for observed disk. Disk breakage algorithm uses the above calculated equation to measure the SMART parameter values based on the time provided.

5) *Considering the weighted approximation:* The equations that support the weighted parameters like environmental changes and importance of data stored is given below;

$$a0 \sum_{i} wi + a1 \sum xiwi + a2 \sum x^2 iwi = \sum yiwi \quad (2)$$

$$a0 \sum xiwi + a1 \sum x^2 iwi + a2 \sum x^3 iwi = \sum xiyiwi \quad (3)$$

$$a0 \sum xi^2 wi + a1 \sum xi^3 wi + a2 \sum xi^4 wi = \sum xi^2 yiwi \quad (4)$$

Equations 2,3 and 4 are noted from weighted least squares curve fitting normalisation mechanism [7]. And are used in the proposal for considering the weighted factor (w_i) for environmental variations.

6) *Identifying the sudden changes noted via plot:* Impulse response time (Δt) between the two consecutive time points are calculated in order to estimate the sudden changes occurred within the combined S.M.A.R.T parameter plot. The algorithm considers the slope variations derived from impulse response and modifies the weighted approximation accordingly.

7) *Identifying the contour region surrounding the breakage point:* The achieved equation is extrapolated based on the resource procurement period (differs from vendor to vendor). Apply Voronoi principles [8] to identify the closed contour region around the given point plotted on the site as shown in Fig: 3

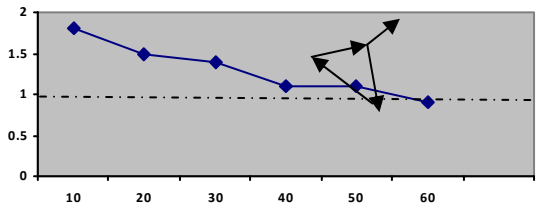

Fig. 2 Plot for Time (hours) on x-axis vs. S.M.A.R.T parameters on y-axis and disk breakage threshold

Assumed here is the disk breakage at SMART parameter value equals to '1'. The plotted curve forms the extrapolation of disk values based on time variations. The contour region circled against the marked break point denotes alarming time gap that triggers an auto notification to the cloud administrator.

The smarter cloud management stack prototype achieved by using the proposal is tabulated as shown in Table-2.

Disk no.	Degrade	Data importance	Time left	Backup	Replace Priority
SATA1	50%	4%	72hr	Yes	3
SSD1	90%	1%	30hr	Yes	6
SATA4	40%	28%	>50hr	No	4
SATA6	10%	50%	>100hr	No	5
SATA8	99%	78%	24hr	No	1
SSD7	2%	10%	>100hr	Yes	7
SSD0	75%	35%	55hr	No	2

Table-2 : Noted various disk parameters, calculated degradation values, time left for replacement and its measured priority

The above table demonstrates the heterogeneous cloud scenario containing variations in disk technologies, degradation ratios, data importance factor, extrapolated time scale and back up provided for individual disk array.

The proposed disk replacement methodology calculates the degradation factor using the SMART parameter variations, extrapolated time left for replacement using curve fitting mechanism and provide the values to the centralised management stack. The centralised management application identifies the data importance stored within the disks,

backup option available at each disk site and forms priorities for disk replacement.

The proposed algorithmic description is framed as a flowchart as shown in Fig: 3.

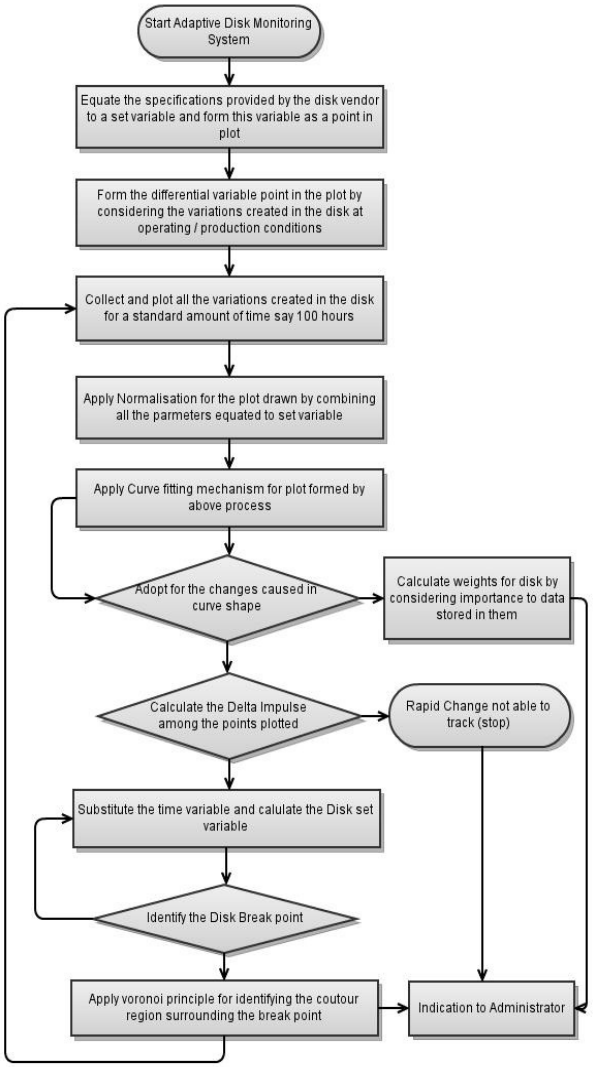

Fig. 3 Flow chart demonstrating the proposed algorithm for disk breakage prediction

V. CONCLUSIONS

Systems and procedures are integrated to have enhanced methodologies for disaster recovery and disk replacement strategies. In the above sections, the device characteristics and important factors discussed are required for enhancement of current disaster recovery systems and disk replacement strategies. Integrating these features in storage subsystems will enhance the systems and will be a step towards making smarter and greener planet. Below are some of the subsystems which might make use of the above findings.

1) Programmers skilled in the art of filesystems or databases can ensure that the foregoing and other changes in form and details may be made therein to achieve the stated proposal.
2) To reconfirm on the read/write speed characteristic of the underlying disks and associated technology and to make sure that the data being moved is on the right disks, the proposed methodology can be deployed with an additional facility which will periodically run read/write test on the disks to reconfirm the disk characteristic to hold high read affinity or high write affinity data.
3) The type and technology of a given disk (required for the proposal) can be identified using the CIM interfaces provided by the disk vendor.

REFERENCES

[1] S. N. Piramanayagam and Tow C. Chong, Developments in Data Storage: Materials perspective, John Wiley & Sons, New Jersey, 2012.
[2] Chris Wolf, *The Definitive Guide to Building Highly Scalable Enterprise File Serving* Statistics. Realtimepublications.com, 2005.
[3] C.G. Rudolph, "Business Continuation Planning/Disaster Recovery: a Marketing Perspective", IEEE Communications Magazine, 1990, Vol. 28, No. 6, pp. 25-28.
[4] Zbigniew Chlondowski. "S.M.A.R.T. Site: attributes reference table". S.M.A.R.T. Linux. Retrieved January 17, 2007.
[5] Intel X18-M/X2 SATA Solid State Drive product manual, May 2009.
[6] Lawrence C. Evans, *An Introduction to Stochastic Differential Equations version,* Department of Mathematics, US Berkely.
[7] R. K. Jain, S.R.K. Iyengar and M. K. Jain*, Numerical Methods*, New Age International, 2009
[8] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, Dr Sung Nok Chiu, *Spatial Tessellations: Concepts and Applications of Vorornoi Diagrams*, John Wiley & Sons, 2009