

Scaleable Intelligent Video Server System

| | |
|----------------------------------|--|
| <i>Title</i> | SATA and Tape Drive Quality Monitoring Software – Publishable version |
| <i>Revision</i> | C |
| <i>Deliverable #</i> | D9.1 |
| <i>Author</i> | Laurent Taurines et al. |
| <i>Company</i> | Hi-Stor Technologies |
| <i>Date</i> | 19/04/2007 |
| <i>Filename</i> | SIVSS_D9.1_V3.doc |
| <i>Dissemination²</i> | PU |

| REVISION | DATE | DESCRIPTION |
|----------|------------|----------------|
| A | 12/8/2005 | Created by LT |
| B | 30/10/2006 | Updated by ns |
| C | 19/04/2007 | Public version |
| | | |
| | | |
| | | |

² **CO** = Confidential (only for members of the consortium + EC); **RE** = Restricted to a stated circulation list (+ EC)
[replace this footnote with the list]; **PP** = Restricted to other FP6 participants (+ EC); **PU** = Public

TABLE OF CONTENTS

| | | |
|-----------|--|-----------|
| 1 | INTRODUCTION | 4 |
| 2 | HARD DISK DRIVE MONITORING FEASABILITY STUDY | 5 |
| 2.1 | SMART OVERVIEW | 5 |
| 2.1.1 | <i>On-line Data Collection</i> | 6 |
| 2.1.2 | <i>Off-line Data Collection</i> | 6 |
| 2.1.3 | <i>Overall-health Status</i> | 6 |
| 2.1.4 | <i>Short Self-Test.....</i> | 6 |
| 2.1.5 | <i>Extended Self-Test</i> | 7 |
| 2.1.6 | <i>Conveyance Self-Test.....</i> | 7 |
| 2.1.7 | <i>Selective Self-Test</i> | 7 |
| 2.1.8 | <i>SMART Error Log</i> | 7 |
| 2.1.9 | <i>Standard & Feature Summary.....</i> | 8 |
| 2.2 | SMART TOOLS ON THE MARKET | 8 |
| 2.2.1 | <i>Market Applications.....</i> | 8 |
| 2.2.1.1 | <i>Third-Party Applications.....</i> | 8 |
| 2.2.1.2 | <i>Drive Manufacturer Applications.....</i> | 9 |
| 2.2.2 | <i>Key Features.....</i> | 12 |
| 2.2.2.1 | <i>Notification and Action Control Centre.....</i> | 12 |
| 2.2.2.2 | <i>statistical analysis algorithms.....</i> | 12 |
| 2.2.2.2.1 | <i>Linear trend.....</i> | 12 |
| 2.2.2.2.2 | <i>Non-linear trend</i> | 13 |
| 2.3 | LIMITATIONS AND CONSTRAINTS | 13 |
| 2.3.1 | <i>Few data from non-intrusive tests</i> | 13 |
| 2.3.2 | <i>Sending commands through RAID Controllers</i> | 13 |
| 2.3.3 | <i>Attribute IDs are Vendor Specific.....</i> | 13 |
| 2.3.4 | <i>Attributes list is no longer mandatory</i> | 13 |
| 2.3.5 | <i>Off-line Data Collection is obsolete</i> | 14 |
| 2.3.6 | <i>SMART Error Log Limitations</i> | 14 |
| 3 | MAGNETIC TAPE & DRIVE QUALITY MONITORING | 15 |
| 3.1 | STATISTICAL ERROR RATE INFORMATION | 15 |
| 3.1.1 | <i>Instant errors</i> | 15 |
| 3.1.2 | <i>Cumulative errors.....</i> | 15 |
| 3.2 | TAPE ALERTS | 15 |
| 3.3 | INFORMATION AVAILABILITY..... | 16 |
| 4 | HI-STOR TECHNOLOGIES IMPLEMENTATION : STORSENTRY..... | 16 |
| 4.1 | MONITORING SOFTWARE ARCHITECTURE | 16 |
| 4.2 | REAL-TIME ANALYSIS | 16 |
| 4.3 | DETAILED FUNCTIONS OF STORSENTRY | 17 |
| 4.4 | STORSENTRY USER INTERFACE | 17 |
| 4.4.1 | <i>Tracing the use of error correction algorithms</i> | 18 |
| 4.4.2 | <i>Implementing alert devices</i> | 19 |
| 4.5 | SELF-HEALING CAPABILITY..... | 20 |
| 4.6 | FINAL INTEGRATION ON SIVSS PLATFORM..... | 20 |
| 4.6.1 | <i>Introduction</i> | 20 |

| | | |
|-----------|--|-----------|
| 4.6.2 | <i>Operation and performance</i> | 21 |
| 4.6.3 | <i>Tape drives</i> | 22 |
| 4.6.3.1 | Identification | 22 |
| 4.6.3.2 | Drive usage..... | 22 |
| 4.6.4 | <i>Gant Diagram</i> | 24 |
| 4.6.5 | <i>Bit Error Rate</i> | 26 |
| 5 | APPENDIX | 26 |
| 5.1.1 | <i>SATA Drive Experiment Results</i> | 26 |
| 5.1.1.1 | Drive Description..... | 26 |
| 5.1.1.2 | Attribute List..... | 27 |
| 5.1.1.3 | Result Analysis..... | 28 |
| 5.1.1.3.1 | 194_Temperature_Celsius | 28 |
| 5.1.1.3.2 | 1_Raw_Read_Error_Rate | 28 |
| 5.1.1.3.3 | 7_Seek_Error_Rate..... | 29 |
| 5.1.2 | <i>Product-moment correlation coefficient r</i> | 29 |

1 INTRODUCTION

This deliverable presents disk drive and tape drive quality monitoring.

In collaboration with the hardware manufacturers partners, Hi-Stor has to study the feasibility of a new SATA quality monitoring method and create a high speed tape drive real time quality monitoring software that provides overall reliability performances of the video system storage resources. All quality information has to be consolidated in a central powerful database, quality indicators will be created through defect mechanism understanding brought by the hardware manufacturers. Quality monitoring results could be made available through LAN or web browsing access. The quality monitoring software will be connected to the storage enclosure services to inform them of proactive operations to be taken, it shall also be connected to the data mover system in order to initiate data migration operations between storage resources.

The quality monitoring software will then be integrated in the SIVSS platform and the application server. Failure mechanisms will be simulated to demonstrate the efficiency of the quality monitoring software.

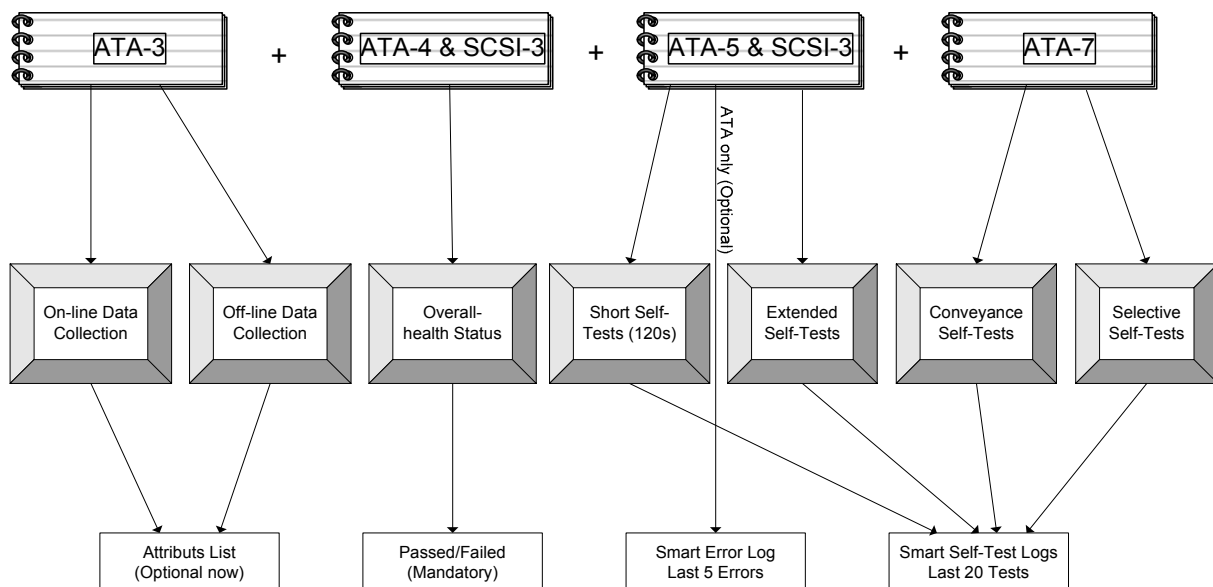
The following chapters detail Hi-Stor work to achieve these objectives.

2 HARD DISK DRIVE MONITORING FEASIBILITY STUDY

2.1 SMART Overview

S.M.A.R.T (Self-Monitoring Analysis and Reporting Technology) is a reliability-prediction technology for ATA/IDE and SCSI disc drives. It is a way to anticipate the failure of a disc drive with sufficient notice to allow a system, or user, to back up data prior to a drive's failure. The original SMART specification (SFF-8035i) was first written by drives manufactures in 1995. Parts of it were merged into ATA-3 standard. Its second revision written in 1996 specified that a list of attributes whose each of them describes a drive's performance or reliability measure, was internally maintain.

Some attribute values are continuously collected while drives are running operating system read and write commands whereas other attribute values are the result of off-line tests built into the drive's firmware. Other built-in tests specified in ATA-5 describes two levels of diagnostic tests that the host can instruct the drive to execute: the short test and the extended test. The short test takes less than 120 seconds. The extended test takes approximately one minute for every gigabyte of disk space. According to Seagate short tests detect whether the drive had failed in 60% to 70% of the cases and extended tests distinguish good and bad devices in 95% of the cases. ATA-5 also introduced an error log whereas conveyance and selective tests were introduced to the ATA_7 standard.



The ATA standard specifies two test modes; off-line mode and captive mode.

In off-line mode the priority is given to the normal system operations. It means that the device shall complete the current commands before executing any self-tests. Devices shall use “off-line” mode for data collection and self-test routines that have an impact on performance if the

device required to respond to commands from the host while performing that data collection.

In captive mode the self-test is executed after receipt of the command until completion or an error occurs.

2.1.1 On-line Data Collection

In order to report predictable failures which are characterized by degradation of an attribute over time, SMART monitors a range of attributes while drives are running operating system read and write commands. Every single attribute is represented by a current, worst and threshold values within the range from 1 to 253 and a raw value coded on 6 bytes. The disc drive sets the appropriate values that the host can retrieve by polling it on a regular basis. A current value below the threshold value indicates an imminent failure or that the device has reached the end of its design life.

some typical characteristics are:

- Head flying height
- Data throughput performance
- Spin-up time
- ECC Error Rate
- Re-allocated sector count
- Seek error rate
- Seek time performance
- Spin try recount
- Drive calibration retry count

Collection of SMART data in “on-line” mode shall have no impact on device performance.

2.1.2 Off-line Data Collection

The ATA standard is a bit confusing with off-line tests and the off-line mode. In this document off-line data collection refers to the off-line tests specified in ATA-3. Those tests shall only be performed in off-line mode. Some of them update attribute values whereas others update the SMART self-test log. Some devices support automatic data collection even though this feature was dropped from the ATA-4 standard.

2.1.3 Overall-health Status

The overall-health status test reports “passed” or “failed”. It uses the on-line and off-line data collection as well as any other results from self-tests to determine the disc health status. It is recommended to back up your data immediately after detecting a failing status.

2.1.4 Short Self-Test

The Short self-test may be performed in either the captive or off-line mode. It should take on the order of ones of minutes to complete. This test are different than the immediate or automatic off-line tests. It check the electrical and mechanical performance as well as the read performance of the disk. Its results are reported in the self-test log.

2.1.5 Extended Self-Test

The Extended self-test may be performed in either the captive or off-line mode. It is a more thorough version of the short self-tests. Its results are reported in the self-test log. Since most of drive manufacturers implement them by scanning the entire disc capacity they may take on the order of ones of hours to complete.

2.1.6 Conveyance Self-Test

Conveyance self-tests may be performed in either the captive or off-line mode. This self-test is intended to identify damage incurred during transporting of the device. It should take on the order of minutes to complete. Its results are reported in the self-test log.

2.1.7 Selective Self-Test

The selective self-test is optional. It may be performed in either the captive or off-line mode. Even though the ATA-7 standard says that selective self-test should include the initial tests performed by the extended self-test plus a selectable read scan, it seems that drive manufacturers use it only to test a specific disc area in order to reduce the execution time. The selective self-test log can be written in order to specify the one or several range of disc logical block addresses (LBA) to be tested. Its results are reported in the self-test log.

2.1.8 SMART Error Log

It is not clear from the ATA standard what should update the SMART error log. It seems that drive manufacturers use it to store up to 5 of the most recent major errors that occurred during on-line and off-line data collection, any type of self-tests as well as normal system read and write operations. The ATA Specification ATA-5 says: "Error log structures shall include UNC errors, IDNF errors for which the address requested was valid, servo errors, write fault errors, etc. Error log data structures shall not include errors attributed to the receipt of faulty commands such as command codes not implemented by the device or requests with invalid parameters or invalid addresses." The definitions of these terms are:

UNC (UNCorrectable): data is uncorrectable. This refers to data which has been read from the disk, but for which the Error Checking and Correction (ECC) codes are inconsistent. In effect, this means that the data can not be read.

IDNF (ID Not Found): user-accessible address could not be found.

For each of these errors, the disk power-on lifetime at which the error occurred is recorded, as is the device status (idle, standby, etc) at the time of the error. For some common types of errors, the Error Register (ER) and Status Register (SR) values can also be retrieved. The meanings of these are:

- ABRT:** Command **AB**oRTed
- AMNF:** Address **M**ark **N**ot **F**ound
- CCTO:** Command **C**ompletion **T**imed **O**ut
- EOM:** End **O**f **M**edia
- ICRC:** Interface **C**yclic **R**edundancy **C**ode (CRC) error
- IDNF:** **I**Dentity **N**ot **F**ound
- ILI:** (packet command-set specific)
- MC:** **M**edia **C**hanged

MCR: Media Change Request
NM: No Media
obs: obsolete
TK0NF: Track 0 Not Found
UNC: UNCorrectable Error in Data
WP: Media is Write Protected

In addition, up to the last five commands that preceded the error are listed, along with a timestamp measured from the start of the corresponding power cycle. The key ATA disc registers are also recorded in the log. The final column of the error log is a text-string description of the ATA command defined by the Command Register (CR) and Feature Register (FR) values. If the command that caused the error was a READ or WRITE command, then the Logical Block Address (LBA) at which the error occurred can be retrieved.

2.1.9 Standard & Feature Summary

| Standards | Tests | Modes | | | Results | | | |
|-------------------------|--------------------------|-----------------------|-------------|---------|------------------------------|---------------|--|---|
| | | Mandatory Optional | Off line | Captive | Attribut List (option) | Passed/Failed | SMART Error Log last 5 errors | SMART Self-test Log last 20 tests |
| ATA-3 | On-line Data Collection | ? | x | NA | x | | | |
| | Off line Data Collection | ? | x | no | x | | | |
| ATA-4 & SCSI | Overall-health Tests | ? | ? | ? | | x | | |
| ATA-5 & SCSI | Short Self-tests (120s) | ? | x | x | | | | x |
| | Extended Self-tests | ? | x | x | | | | x |
| ATA-7 | Conveyance Self-tests | ? | x | x | | | | x |
| | Selective Self-tests | optional | x | x | | | | x |

2.2 SMART tools on the market

The experiment is using DriveSitter and Smarmonetools. A few drive manufacturer applications will be used later on.

2.2.1 Market Applications

2.2.1.1 Third-Party Applications

The application list below describes third-party programs.

[DriveSitter](#) is a full-featured HDD analysis, health diagnostic and background monitoring tool for IDE hard disk drives based upon modern S.M.A.R.T. technology. It reliably detects and forecasts up to 70% of all sudden HDD failures before they happen: You will be alerted of any unhealthy condition by various notification methods including emails, network messages

and execution of user-defined files. Furthermore, DriveSitter collects a stock of past health attributes and estimates the remaining lifetime of each health related attribute (T.E.C. dates) by advanced statistical analysis that provide a minimum of false alarms.

Smartmontools: The smartmontools package contains two utility programs (**smartctl** and **smartd**) to control and monitor storage systems using the Self-Monitoring, Analysis and Reporting Technology System (SMART) built into most modern ATA and SCSI hard disks. In many cases, these utilities will provide advanced warning of disk degradation and failure. It is derived from the [smartsuite package](#), and includes support for ATA/ATAPI-3 to -7 disks and SCSI disk and tape devices.

S.M.A.R.T. Disk Monitor (from SanTool): Monitors local and SAN-mounted storage devices and provides predictive-failure notification using industry-standard S.M.A.R.T technology.

DriveHealth (from Helexis Software Development): This tool allows for predicting possible HDD failure and prevents losing the critical data using S.M.A.R.T. technology that is supported by the most of hard disk manufacturers. It eats minimum CPU and memory resources and can be run as a NT service.

Intelli-Smart (from SoftJam Software): Based on the S.M.A.R.T hard drives and RAID Arrays can be continuously monitored for performance and the user or administrator alerted prior to a disk drive failure. The reporting engine can locally alert the user, send an e-mail, write to the event log or use an SNMP hook. Intelli-SMART is compatible with Windows 95/98/Me/NT/2000 and XP.

ActiveSmart (from Ariolic Software): Various types of hard disks support different numbers of the S.M.A.R.T attributes. In any case, Active SMART will show you and will use all of the S.M.A.R.T. attributes, supported by your disks.

SiGuardian uses technology included in all modern hard disks - S.M.A.R.T. [Self-Monitoring, Analysis and Reporting Technology]. it can estimate your hard drive life span and prevent failures before they occur! You can check your drives at system startup or do it continuously at defined time intervals.

2.2.1.2 Drive Manufacturer Applications

The application list below (also at http://www.benchmarkhq.ru/english.html?be_hdd2.html) describes drive manufacturer diagnostic tools. Most of them have SMART capability amongst other features. The question is: Do those new software offer a higher level of precision in diagnosis?

Fujitsu ATA Diagnostic Tool

OS: Win9x/Me/NT/2k/XP [Download](#) [Homepage](#) 159Kb

Freeware

This is a diagnostic tool to help customers speedily verify whether their Fujitsu hard drive is operating correctly. FJDT can diagnose your suspected faulty hard drive by checking the S.M.A.R.T. data and also by scanning the entire surface of the drive, sector by sector, to verify media integrity. FJDT can be run from a bootable MSDOS diskette, which enables

diagnosis even when the OS is not booting.

Fujitsu SCSI Diagnostics

OS: Win9x/Me/NT/2k/XP [Download](#) [Homepage](#) **737Kb**

Freeware

SCSI Diagnostics is a simple and reliable SCSI diagnostic tool developed by Fujitsu verifying the condition of your Fujitsu SCSI Hard Disk Drive. It checks S.M.A.R.T. information and performs a series of tests.

IBM-Hitachi Drive Fitness Test

OS: Win9x/Me/NT/2k/XP [Download](#) [Homepage](#) **1.9Mb**

Freeware

The Drive Fitness Test (DFT) provides a quick, reliable method to test SCSI and IDE hard disk drives. It analyzes drive fitness, restores drive fitness (contains the following utilities: Erase Bootsector, Low-level format, Filesystem-based Corrupted Sector Repair), displays drive information, etc.

Maxtor Powermax 4.09

OS: Win9x/Me/NT/2k/XP [Download](#) [Homepage](#) **936Kb**

Freeware

The Powermax utility is designed to perform diagnostic read/write verifications on Maxtor/Quantum hard drives. These tests will determine hard drive integrity. The Powermax utility is effective on all ATA (IDE) hard drives with a capacity greater than or equal to 500 MB. All data will be lost when the "Write Disk Pack (low level format)" test is performed.

Maxtor SCSIMax

OS: Win9x/Me/NT/2k/XP, DOS [Download](#) [Homepage](#) **70Kb**

Freeware

SCSIMax is a diagnostic utility for all Maxtor/Quantum SCSI hard disk drives supporting Self-Monitoring, Analysis, and Reporting Technology (S.M.A.R.T). This test will determine hard drive integrity in a short period of time, with a high degree of confidence.

Quantum Data Protection System

OS: Win9x/Me/NT/2k/XP [Download](#) [Homepage](#) **108Kb**

Freeware

QDPS is a diagnostic utility for Quantum SCSI hard disk drives supporting S.M.A.R.T. technology.

Samsung Diagnostic (SHDiag)

OS: DOS [Download](#) [Homepage](#)

Freeware

This program is used to diagnose the disk when the SAMSUNG hard disk ([supported models](#)) is suspected to have failures. It is strongly recommended to back up your data before using this program!

Samsung Drive Diagnostic Utility (Hutil)

OS: DOS [Download](#) [Homepage](#) 187Kb

Freeware

Hutil (The Drive Diagnostic Utility) is made with the aim of testing a Samsung hard disk drive while it is installed inside a PC, regardless of the status of user's operating system. Hutil can test a drive solely manufactured by Samsung ([supported models](#)). It is strongly recommended to back up the user's significant data in advance because Hutil has a Write operation that can erase it.

Seagate SeaTools

OS: Win9x/Me/NT/2k/XP/DOS/Linux [Download](#) [Homepage](#)

Freeware

SeaTools Suite is Seagate's exclusive disc drive diagnostic software designed to troubleshoot most Seagate hard drive issues. It consists of 3 versions: Online, Desktop and Enterprise. Online is a browser based application that will check ATA and SCSI hard drives without having to turn off the system. Desktop edition works with most ATA or SCSI drives in desktop systems and has a 98% accuracy rate. It creates a bootable diskette. Enterprise is ideal for SCSI or Fiber Channel drives in servers and workstations. Tests multiple drives simultaneously and sequentially.

SMART Defender

OS: Win9x/Me/NT/2k/XP [Download](#) [Homepage](#) 1.9Mb

Freeware

Hitachi's SMART Defender is an easy to use Windows program that monitors SMART-capable IDE and SCSI hard disk drives. The program reduces the risk of system down time by assessing the reliability and predicting hard disk drive failures.

Western Digital Data Lifeguard

OS: Win9x/Me/NT/2k/XP [Download](#) [Homepage](#) 1.5Mb

Freeware

Western Digital Data Lifeguard was written specifically for Western Digital EIDE hard drives. Its purpose is to examine the drive and test its functionality. It scans the drive and looks for errors. It even has the ability to correct certain types of errors that it may encounter. The dlgmaker.exe program must be executed on a Windows 95/98/ME system to create the Data Lifeguard Tools disk.

Western Digital DLG for Windows

OS: Win9x/Me/NT/2k/XP [Download](#) [Homepage](#) 3.4Mb

Freeware

This is a Windows version of the Data Lifeguard Diagnostics utility. The utility can perform drive identification, diagnostics, and repairs on a Western Digital FireWire, EIDE, or USB drive. In addition, it can provide you with the drive's serial and model numbers.

2.2.2 Key Features

2.2.2.1 Notification and Action Control Centre

(Most of the applications)

When a Threshold Exceeded Condition (TEC) is reached there are several ways to signal it; email, sound, message, system log as well as an application or script can be triggered on a per-device basis.

2.2.2.2 statistical analysis algorithms

(DriveSitter only)

A **T.E.C. date** is the estimated date when the attribute value will probably exceed its threshold. The estimation is based upon the S.M.A.R.T. attribute values collected in the past under the assumption that the observed trend will continue in the future. The following chapters discuss the T.E.C. date forecast in detail.

alpha or type I error: False Alert

beta or type II error: Unraised Alert

2.2.2.2.1 Linear trend

(Implemented by DriveSitter. See annex for further details)

The linear trend is the best approximated linear representation of the interrelation of two variables (the S.M.A.R.T. attribute values and the time, in our case). Moreover, it can be described by a single line which can in turn be defined by an intercept (y-offset from the x-axis) and a slope. Once we find this line, we can easily compute the estimated time the trend line will probably cross the threshold value.

The DriveSitter's underlying algorithm utilizes the Pearson product-moment correlation coefficient to calculate the best approximation of the linear trend. The dataset consists of all observations that are stored within the S.M.A.R.T. attributes stock (up to 999 observations per drive and per attribute). There are several settings to fine-tune DriveSitter's T.E.C. date forecast and avoid alpha and beta errors as far as possible. Particularly, the user can specify the maximum forecast time span, the minimum number of observations to calculate a trend, the minimum required coefficient of determination ("prediction quality"/portion of explained variance) and the alpha level for significance testing. Moreover the user can specify whether a virtual T.E.C. date should be calculated for S.M.A.R.T. attributes that do not define an explicit threshold (threshold equals 0, so called "performance parameters") and can enable or disable the appraisal of forecast T.E.C. dates for each individual attribute on a per-drive basis.

2.2.2.2.2 *Non-linear trend*

when dealing with nonlinear trend approximation, the distribution within the dataset must be analyzed. Since we do not know the real shape of the distribution, the only way to find a good approximation is to fit a variety of theoretical distributions to the dataset and compare the so-called "goodness of fit". Provided that we find a distribution that sufficiently approximates the data at the best, we just have to find a way to estimate the T.E.C. date which might be the easiest part. Do not get hold of the wrong end of the stick, there are several known statistical techniques to appraise the goodness of fit for almost every imaginable distribution. Nonetheless these techniques are complex and costly. While a non-linear trend approach would be likely to yield pretty accurate results, it is at the same time very complex to implement and very expensive in means of load on the computer memory and/or the system performance.

2.3 **Limitations and Constraints**

2.3.1 **Few data from non-intrusive tests**

Only the on-line data collection has no impact on the system performance. So far the on-going experimentation on 2 Seagate drives (ST380013AS) shows that only 2 attributes are varying (see specific section for further details). Moreover most of the attributes indicate that the drive has reached its end of life and therefore are not very relevant for predicting failures.

2.3.2 **Sending commands through RAID Controllers**

In order to access the disks behind a controller, a pass-through command must be sent of to the controller. Those types of commands are specific to the RAID controller vendors and allows to send only basic comands to the disk system.

2.3.3 **Attribute IDs are Vendor Specific**

The attribute names and meanings as well as the interpretation of the raw value are not specified in any standard. Therefore it could happen that different drive manufacturers use the same attribute ID to describe different drive characteristics or measures. So far it seems that the attribute meaning is a de facto standard whereas the RAW value meaning is vendor specific for some of the attributes. For instance, the attribute ID 194 is the temperature attribute for all drive manufacturers but the associated raw value might be a Celsius, Kelvin, Fahrenheit or any other temperature scale.

Therefore the key point in handling attributs is to know how disc vendor interpret the 6 bytes of the raw value.

2.3.4 **Attributes list is no longer mandatory**

The attributes list was first introduced in the ATA-3 standard. Starting with the ATA-4 standard, the requirement that disks maintain an internal attribute table was dropped. Instead, the disks simply return an OK or NOT OK response to an inquiry about their health. However it is still implemented and used by many vendors.

2.3.5 Off-line Data Collection is obsolete

SMART automatic off-line data collection command is listed as "Obsolete" in every version of the ATA and ATA/ATAPI Specifications. It was originally part of the SFF-8035i Revision 2.0 specification, but was never part of any ATA specification. However it seems to be still implemented and used by many vendors.

2.3.6 SMART Error Log Limitations

Because of the limitations of the SMART error log, if the LBA is greater than 0xffffffff, then either no error log entry will be made, or the error log entry will have an incorrect LBA. This may happen for drives with a capacity greater than 128 GiB or 137 GB.

Moreover some manufacturers ignore the ATA specifications, and make entries in the error log even though the device receives a command which is not implemented or is not valid.

As a conclusion, the disk failure was judged too uncertain to predict. As a consequence, it was decided to report monitoring efforts on tape monitoring where monitoring feasibility has been proved. In the annex, the test results that have been performed on a single disk are displayed showing the difficulty to predict a failure with a disk drive.

3 MAGNETIC TAPE & DRIVE QUALITY MONITORING

3.1 Statistical Error Rate Information

Errors both recovered by the internal drive code correction or recovery process, or non recovered are available for most of the tape drives technologies through the control interface, like SCSI (*Small Computer System Interface*) or Fiber Channel interfaces.

In order to protect data, tape drives add additional bytes to each user data block to be written and the quantity depends on the correction techniques implemented by the drive. These additional bytes include the CRC (Cyclic Redundancy Check) and the ECC (Error Correction Code). The CRC is made up of bytes calculated with a polynomial; it ensures that an error in the block is detected when rereading the data. The drive also reads after write to check the data integrity (check with the error code correction (ECC) activated) and writes it to another location when required (retry operation). The bytes in the ECC are generated using Reed-Solomon algorithms, allowing a certain number of error bytes to be corrected in the block, called BER (Bit or Block Error Rate depending on the ECC used). If the ECC and the “retry” operations fail after multiple trials (number function of the drive manufacturer), the drive reports an “unrecovered error”.

Depending on the error type, the number of occurrences, and defect location repeatability, media aging, dust issues, or physical damages can be sorted into 2 different categories : persistent or volatile.

3.1.1 Instant errors

Some information are returned by the drive following an issued SCSI command. This information is sorted in 2 types :

- Sense key : indicates in what general area the problem that has just been experienced falls.
- Additional sense code : gives a clearer indication of the nature of the problem.

This information is cleared once read by a host.

3.1.2 Cumulative errors

Some information are available in dedicated pages of the drive memory and are not related to a specific issued command. These information are updated after each related information and are cleared either at the cartridge ejection or following cartridge load.

3.2 Tape alerts

This section has been removed to produce a public version.

3.3 Information availability

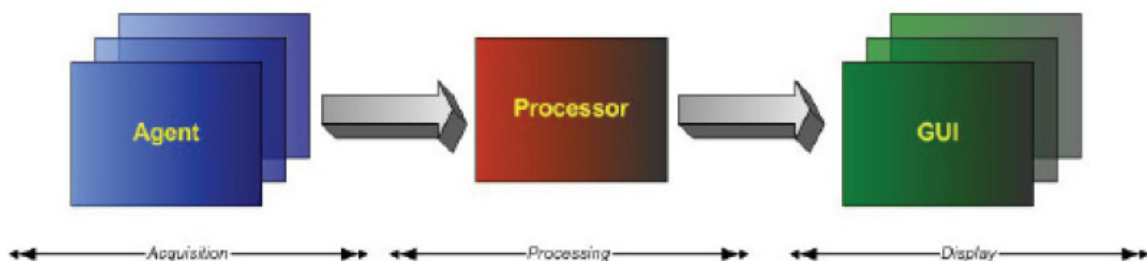
This section has been removed to produce a public version.

4 HI-STOR TECHNOLOGIES IMPLEMENTATION : STORSENTRY

Because Hi-Stor Technologies planned to have the quality monitoring software to be running in parallel of the data mover operations, the idea of getting the instant information through the sense keys was not viable as, the data mover was also able to catch such information, and there was possibilities to crash one or the other components through such implementation. In order to avoid such situation, the software was designed to collect cumulative quality data instead.

4.1 Monitoring Software Architecture

The software is based on a client/server architecture with a module in charge of polling the drive for requested information, and sending these information through the network to another module in charge of processing the data and give the right actions to be taken to insure the infrastructure correct operation.



Agent : this module acquires cumulative data available from the drive log pages.

Processor : this module gathers the data acquired from multiple agents, records these data into a database, and process them in order to give quality/performance diagnostics.

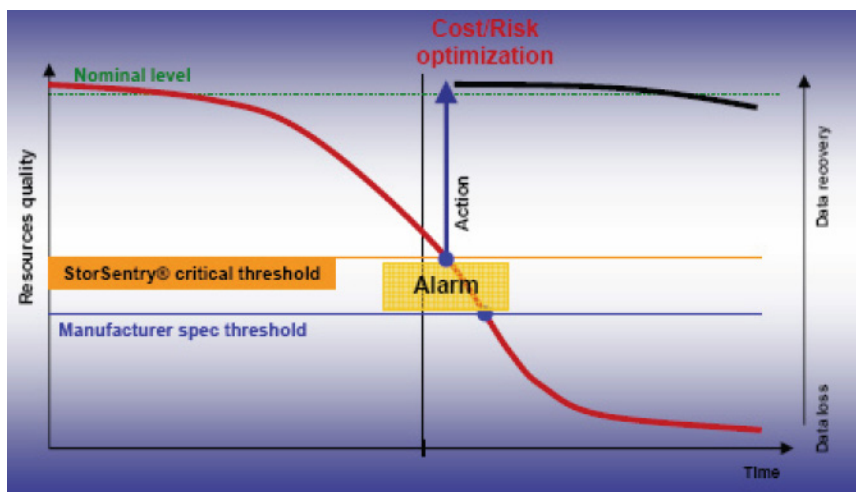
GUI : intuitive interface displaying drive/media quality/performance information through synthetic graphs and tables. Multiple GUI can be connected to a single processor.

4.2 Real-Time analysis

The agent is designed to collect information in parallel to the data mover operating the drive, so that the user has real time status of its infrastructure. This way there is no need to remove a drive or media from a production environment, using manufacturing test tools and reference drive / media to get a quality assessment.

As a media or drive quality and usage history is kept into the application database, StorSentry can compute the trends and compare the different values like the BER (Bit Error Rate), number of loads, degradation rate,... to thresholds defined by the manufacturers. Moreover, keeping trace of the history of each devices (drive or media) enables to discriminate which device is faulty when an error occurs. Obviously, this is possible when enough information has been collected on the drives (after 5 load cycles).

This enables the infrastructure administrator to better to secure the data filed and stored on tapes, through automatic centralized analysis of the media quality over time, providing better planning of retesting/retensioning of externalized tape batches. The regeneration of media is also carried out at the best time (cost/risk trade-off).



4.3 Detailed functions of StorSentry

On-going monitoring of media status and proactive action:

The main functions of StorSentry are shown below:

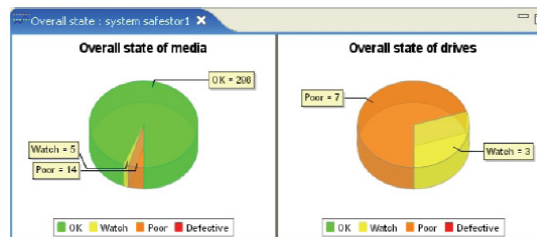
- Analysis of data quality related to drives and media to provide preventive diagnosis of any possible damage and guarantee the storage system works at a high operational level.
- It uses preset modifiable thresholds depending on the value of the contents, which may be different to the manufacturer's thresholds.
- Detailed diagnosis of breakdowns and anomalies and proposals for corrective action such as tape cleaning or replacement. The aim is to ensure service continuity for the architecture or repair of the system in the event of a breakdown. In particular, problems due to faulty media and those due to faulty drives are automatically differentiated by StorSentry.

4.4 StorSentry User Interface

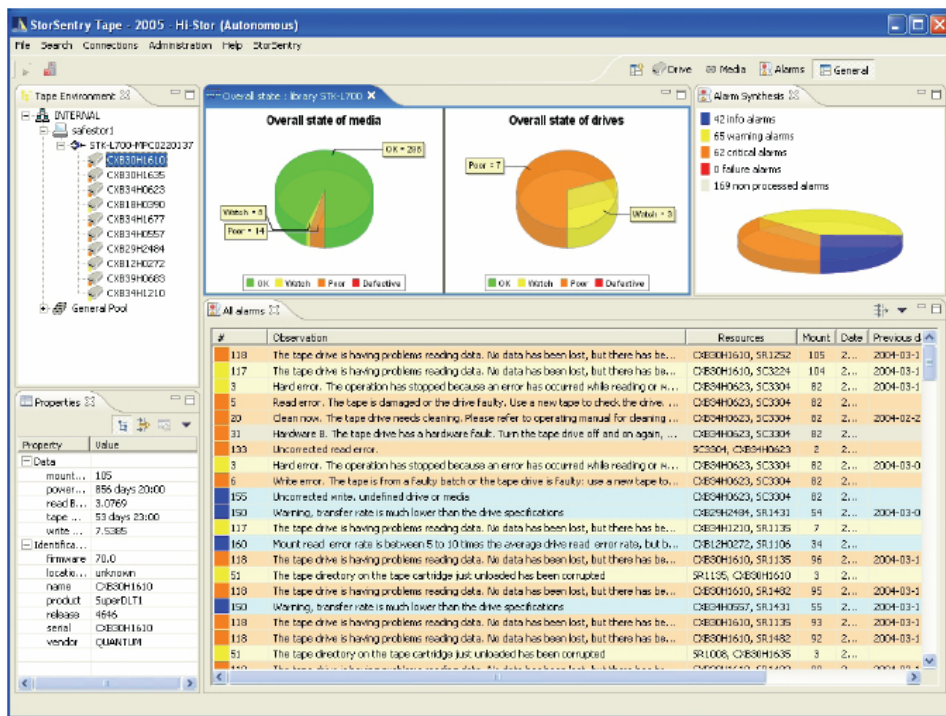
The user interface ensures centralized monitoring reliability and performance levels of the tape storage resources:

- Overall visualization of the quality level of the distributed storage resources.
- A simple search for information using the tree view, provides a physical (resource attached to a server) or a logical view (media belonging to a pool) of the infrastructure. A list of alerts generated, their meaning, and their severity level, is also displayed.

The StorSentry overview allows drives and cartridges to be identified according to their condition: good, to be monitored, poor or faulty. The list of media by condition is directly accessible.



Overview media and drives status



Overview of storage resource quality

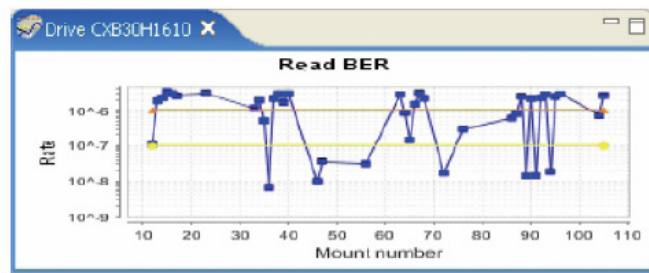
4.4.1 Tracing the use of error correction algorithms

In order to protect data, tape drives add additional bytes to each user data block to be written and the quantity depends on the correction techniques implemented by the drive. These additional bytes include the CRC (Cyclic Redundancy Check) and the ECC (Error Correction Code). The CRC is made up of bytes calculated with a polynomial; it ensures that an error in the block is detected when rereading the data.

The bytes in the ECC are generated using Reed-Solomon algorithms, allowing a certain number of error bytes to be corrected in the block, called BER (Bit or Block Error Rate

depending on the ECC used). The number of corrected bits is calculated by StorSentry. The graph shown below, available on the StorSentry interface, shows the change in BER and the ECC error correction algorithm use.

This graph is available on the interface that shows the change in BER (Bit Error Rate or corrected errors) in the reading of a drive. The yellow line shows the monitoring threshold and the orange line shows the critical threshold. These thresholds can be configured by media pool depending on the customer's needs.



Based on database information StorSentry analyses the drive/media combination to identify causes of failure and to ease the decision making process.

4.4.2 Implementing alert devices

The alerts view allows all the problems detected by StorSentry to be referenced with a message, the name of the resources concerned (drive, cartridge), the date and the severity.

This ensures that preventive decisions can be taken to guarantee data integrity.

Media error alerts

These alerts are triggered by the StorSentry critical threshold defined by pool basis.

For each mount, concerning corrected error rates in both writing and in reading, StorSentry generates an alert depending on the threshold reached. Trend analysis is also carried out and a tape for which no mount has exceeded the threshold of corrected error rates can be subject to another alert, for example, if this rate has increased by a factor of 10 for a significant number of mounts.

End of life alerts

There are alerts for the media expressed in operating hours (real use) or number of mounts

Format recycling alerts

Format recycling alerts can be generated due to the following:

- The corrected error rate is too high
- The media's life has expired
- Etc...

Media retensioning alert

When media has been kept on the shelf for too long, StorSentry generates an alert to retension the tape. The retensioning time can be configured.

4.5 Self-healing capability

In order to take advantage of the diagnosis given by StorSentry, both the data mover and the monitoring application have been coupled through an API.

This has been done for the media only as a first step.

When a media is suspected to be poor, the data mover takes the action of migrating all data onto another media, and blacklist the poor media to avoid anymore use of it.

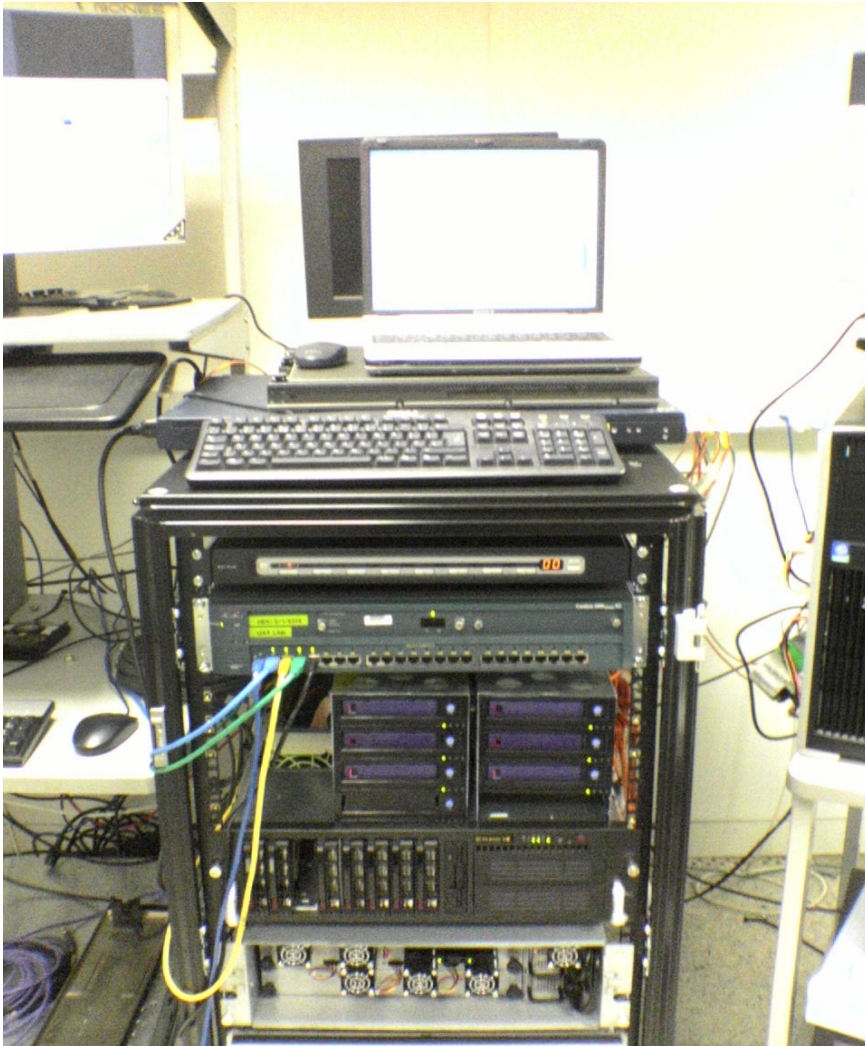
4.6 Final integration on SIVSS platform

4.6.1 Introduction

The tape monitoring software database review described in this document is based on the SIVSS archive activities during SIVSS integration. During the integration process the monitoring software logged data while performance tests on the Data Mover were performed. This analysis is not a real production system analysis but the analysis of an archive system being tested and tuned.

The tape monitoring application, including the agent, the processor and the Admin GUI, has run without failure, nor impact on SIVSS infrastructure. The SIVSS architecture was composed of 6 Tandberg LTO2 Tape Drives.

Based on the collected information, monitoring software allows an analysis of the system performance and usage, to prevent failures.



SIVSS platform during the final integration. The tape system composed of 6 tape drives is above the Data Mover node. Monitoring software has been installed on the Data Mover node.

4.6.2 Operation and performance

The drives usage (Mount/Read/Write) is globally homogeneous; there is no noticeable disparity over the monitored archive infrastructure.

Overall drives transfer rates are good, regarding the O'Mass LTO2 technology. The few alarms related to low transfer rates are due to media errors. Some of the used tapes were not good enough or were old.

The drive quality is globally of good quality. However, 2 of them (RBD42H0205 & PKB38H0671) have an increasing read BER (Bit Error Rate). Even if the recorded BER values are not critical yet, those drives should be monitored to prevent any further permanent errors.

4 cartridges have an increasing number of alarms and should be replaced as soon as possible, to prevent any data loss. The cartridges that were used have become old with the length of the project and were not carefully treated.

4.6.3 Tape drives

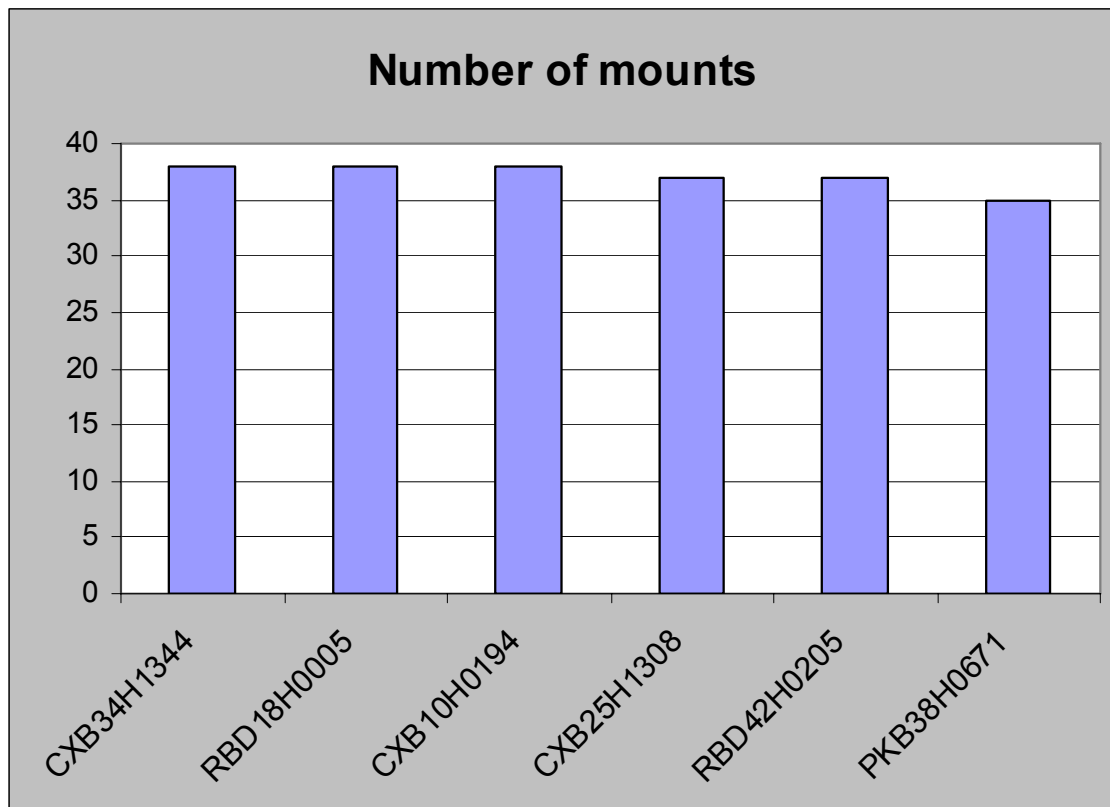
4.6.3.1 Identification

Table 1. Drive identification

| Name | Serial Number | Vendor | Type | Firmware Level | POH | Pos. |
|------------|---------------|----------|------|----------------|--------|------|
| CXB34H1344 | CXB34H1344 | Tandberg | LTO2 | 85.0 | 19,764 | 0 |
| RBD18H0005 | RBD18H0005 | | | | 18,987 | 1 |
| CXB10H0194 | CXB10H0194 | | | | 17,234 | 2 |
| CXB25H1308 | CXB25H1308 | | | 17,567 | 3 | |
| RBD42H0205 | RBD42H0205 | | | 82.0 | 18,145 | 4 |
| PKB38H0671 | PKB38H0671 | | | | 17,876 | 5 |

4.6.3.2 Drive usage

Figure 2. Drive: Number of mounts



The distribution of the total number of mounts over all monitored drives is good:

- Total number of mounts : 223; Average 37
- Max value: 38 (Average + 2.7 %) - Drives CXB34H1344, RBD18H0005 and CXB10H0194
- Min value: 35 (Average – 7.8%) - Drives PKB38H0671

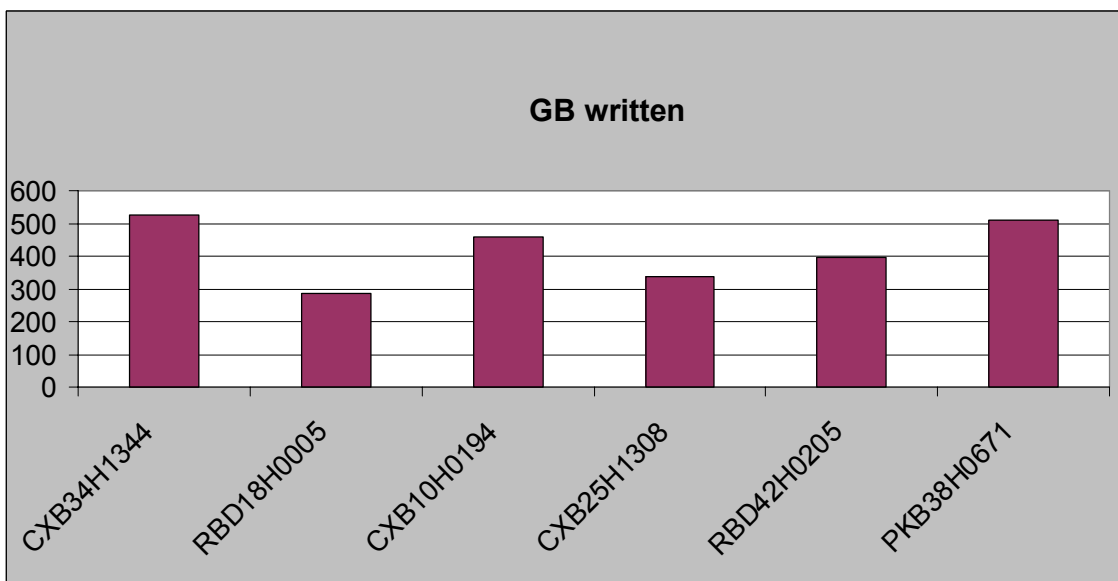
Figure 3. Drive: Read bytes (GB)



The number of read bytes is globally well distributed over all the drives :

- Total number of read bytes: 1,560 GB; Average 260 GB.
- Max Value: 397 GB (Average + 52 %) - Drive RBD18H0005
- Min Value: 202 GB (Average – 22%) - Drive CXB25H1308

Figure 4. Drive: Write bytes (GB)



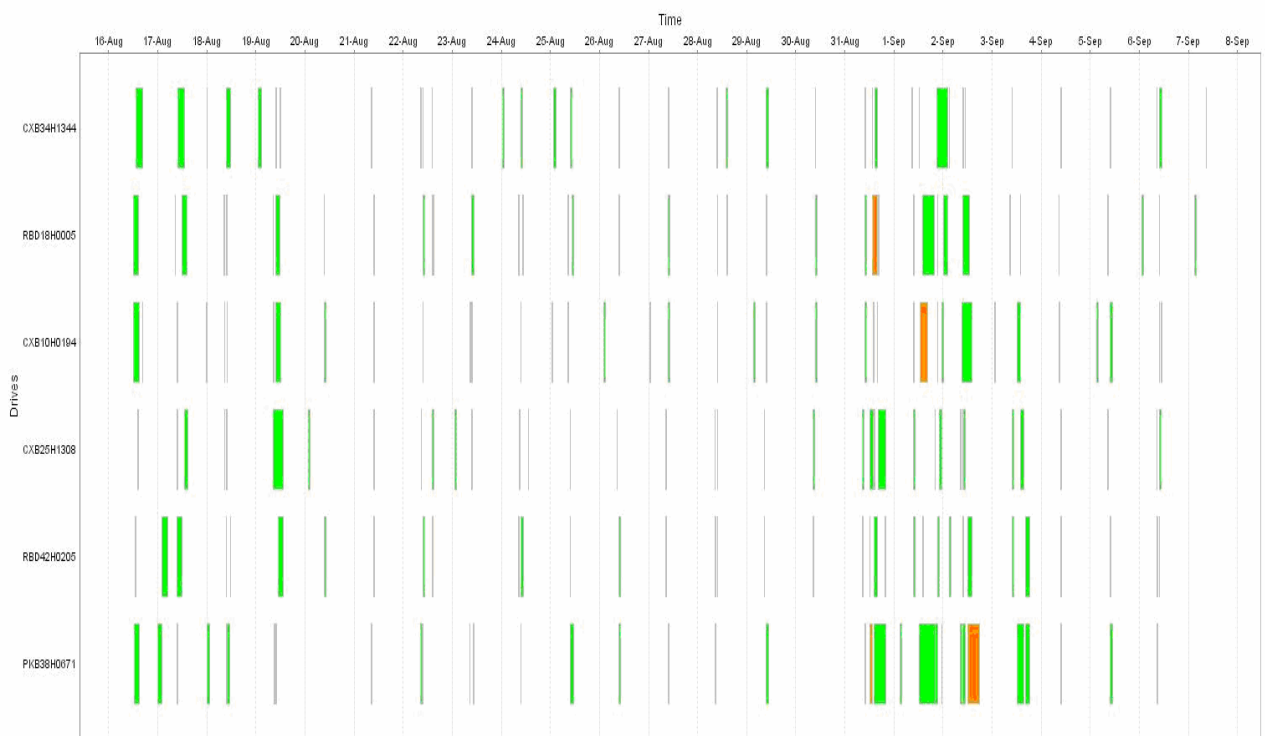
The number of written bytes is globally well distributed other all the drives:

- Total number of written bytes: 2.513 GB; Average 419 GB
- Max Value: 526 GB (Average + 25%) – Drive CXB34H1344
- Min Value: 286 GB (Average – 31%) – Drive RBD18H0005

4.6.4 Gant Diagram

Below is the Gant diagram, during SIVSS platform integration:

Figure 5. Drive: Gant diagram

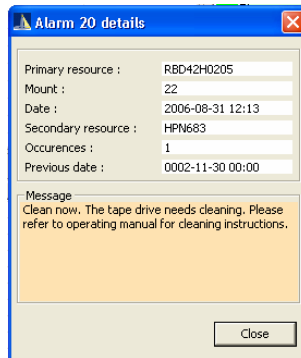
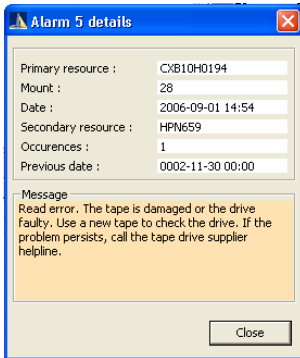


Each strip represents the time spent by a media in a tape drive.

The beginning and the end of each strip match respectively of the use of a tape drive.

Green strips represent nominal transfer rates (24 MB/s), whereas yellow and orange ones represents transfers with low and very low transfer rate (below drive minimal speed)

Figure 1. Drive: Samples of drive related alarms



4.6.5 Bit Error Rate

Despite a few errors encountered on some medias, the read and write BER mark are good.

Figure 2. Drive: Read BER mark

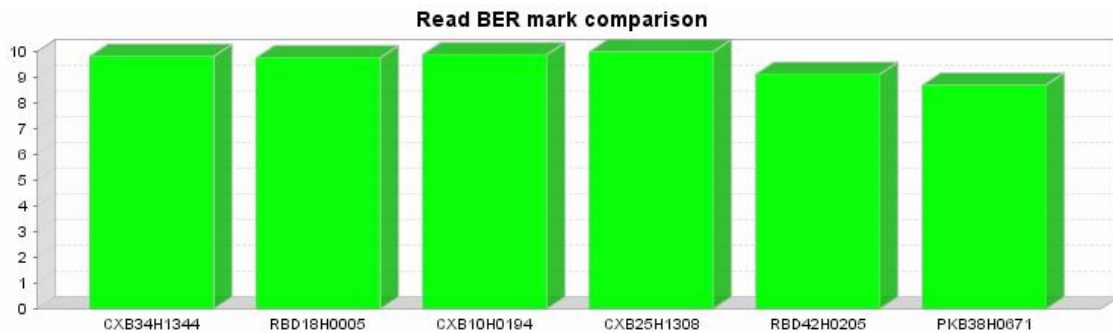
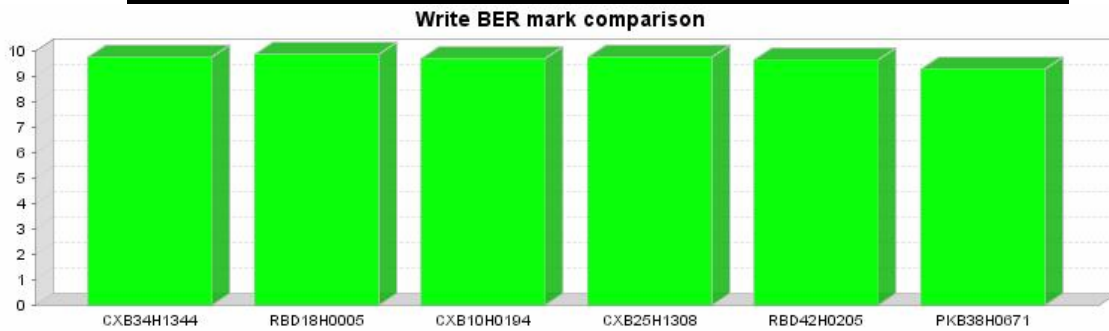


Figure 3. Drive: Write BER mark



5 APPENDIX

5.1.1 SATA Drive Experiment Results

The Seagate SATA drives (ST380013AS) are being exercised in a thermal chamber. While running read and write commands using IOMeter, the SMART attributes are monitored on a regular basis. The temperature of the thermal chamber goes up of 5 degree Celsius every day. Short and extended self-tests are also run every day.

5.1.1.1 Drive Description

Device Model: ST380013AS

Serial Number: 5JV9M2AT

Firmware Version: 3.18

ATA Version is: 6

ATA Standard is: ATA/ATAPI-6 T13 1410D revision 2

Local Time is: Mon Jul 19 11:15:22 2004 Paris, Madrid

SMART support is: Available - device has SMART capability.

SMART support is: Enabled

5.1.1.2 Attribute List

| ID# | Attribut Name | Flag | Value | Worst | Threshold | Type | Updated | Raw_value |
|-----|------------------------|--------|-------|-------|-----------|----------|---------|-----------|
| 1 | Raw_Read_Error_Rate | wpr_o | 60 | 56 | 6 | Pre-fail | Always | 122702364 |
| 3 | Spin_Up_Time | w_o | 100 | 100 | 0 | Pre-fail | Always | 0 |
| 4 | Start_Stop_Count | cos | 100 | 100 | 20 | Old_age | Always | 0 |
| 5 | Reallocated_Sector_Ct | w_cos | 100 | 100 | 36 | Pre-fail | Always | 0 |
| 7 | Seek_Error_Rate | wpr_o | 74 | 60 | 30 | Pre-fail | Always | 29188582 |
| 9 | Power_On_Hours | cos | 100 | 100 | 0 | Old_age | Always | 135 |
| 10 | Spin_Retry_Count | w_co | 100 | 100 | 97 | Pre-fail | Always | 0 |
| 12 | Power_Cycle_Count | cos | 100 | 100 | 20 | Old_age | Always | 3 |
| 194 | Temperature_Celsius | os | 67 | 67 | 0 | Old_age | Always | 67 |
| 195 | Hardware_ECC_Recovered | rco | 60 | 56 | 0 | Old_age | Always | 122702364 |
| 197 | Current_Pending_Sector | co | 100 | 100 | 0 | Old_age | Always | 0 |
| 198 | Offline_Uncorrectable | c | 100 | 100 | 0 | Old_age | Offline | 0 |
| 199 | UDMA_CRC_Error_Count | _prcos | 200 | 200 | 0 | Old_age | Always | 0 |
| 200 | Multi_Zone_Error_Rate | | 100 | 100 | 0 | Old_age | Offline | 0 |
| 202 | TA_Increase_Count | cos | 100 | 100 | 0 | Old_age | Always | 0 |

ID#: Attribute number.

Attribut_Name: Attribute name.

Flag: Every S.M.A.R.T. attribute has some flags defined by the HDD manufacturer. A flag is a data entity that can be either set or cleared. The flags have been abbreviates by one lowercase letter and displays only the flags that are "set" for the attribute.

Warranty: Indicates attributes that are considered to be life critical for the HDD and covered by the drive warranty. If a S.M.A.R.T. attribute with the warranty flag set exceeds its corresponding threshold and the drive is still within the guarantee period, the manufacturer will most likely replace the drive.

Performance: Indicates attributes that represent a performance measure. Performance attributes are generally not considered to be life critical in the eyes of the manufacturer. In one or another case, you might think different.

error Rate: Indicates attributes that represent an error rate.

Count of occurrences: Indicates attributes that represent the number of occurrences.

Online test: If set, indicates that the attribute is only updated by an on-line test. If cleared, the attribute is only updated by an off-line test.

Self preserving: Indicates attributes that can be collected or saved even if the S.M.A.R.T. function of the HDD is disabled.

Value: The current value of the attribute.

Worst: This is the smallest (closest to failure) value that the disk has recorded at any time during its lifetime when SMART was enabled.

Threshold: If the current value is less than or equal to the Threshold value, then the Attribute is said to have failed. If the Attribute is a pre-failure attribute, then disc failure is imminent.

Type: Attributes are one of two possible types: Pre-failure or Old age. Pre-failure Attributes are ones which, if less than or equal to their threshold values, indicate pending disk failure. Old age, or usage Attributes, are ones which indicate end-of-product life from old-age or normal aging and wear out, if the Attribute value is less than or equal to the threshold.

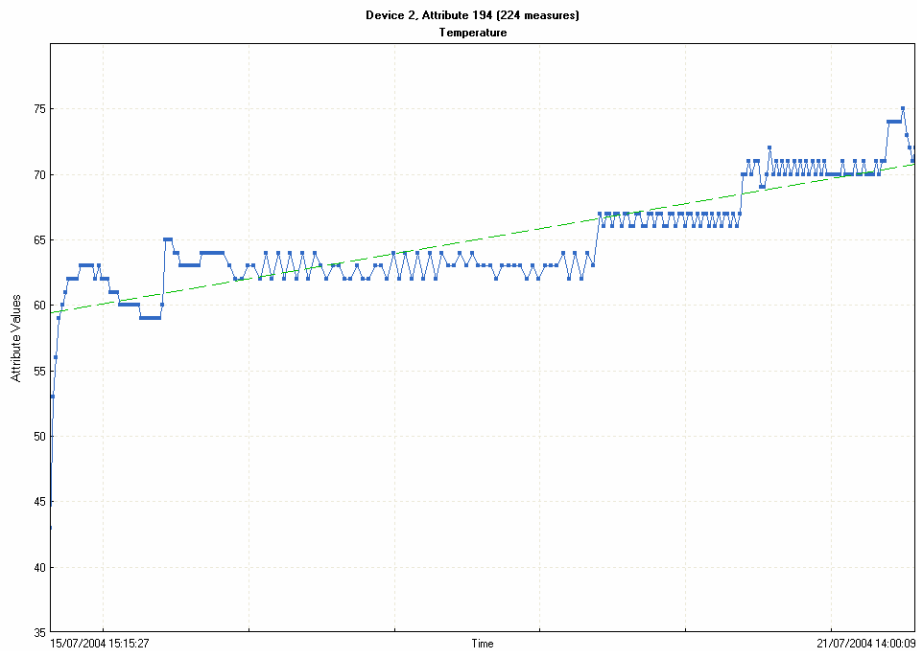
Updated: Shows if the SMART Attribute values are updated during both normal operation and off-line testing, or only during offline testing. The former are labeled "Always" and the latter are labeled "Offline".

Raw_value: Each vendor uses their own algorithm to convert the "Raw" value to the current value in the range from 1 to 253.

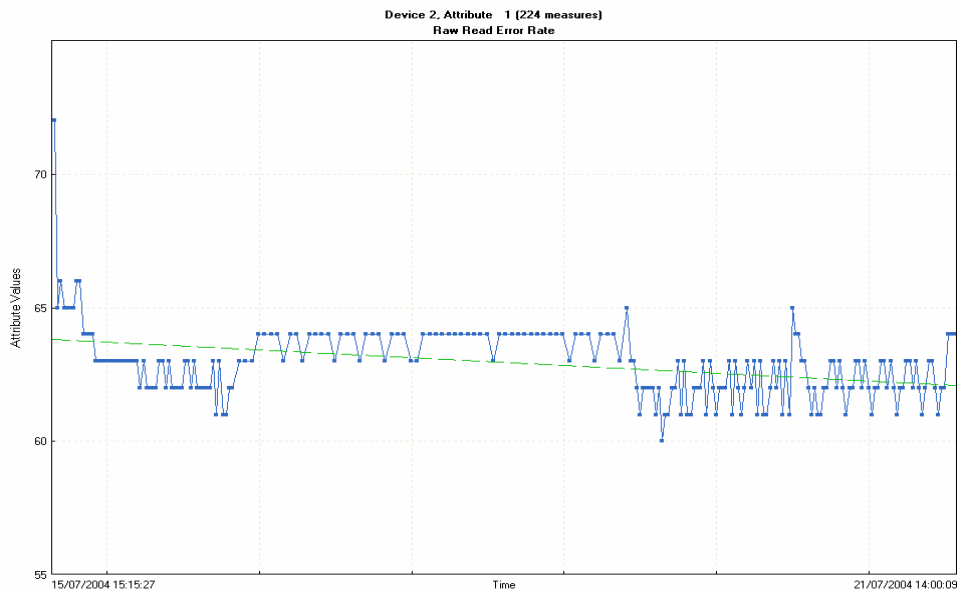
5.1.1.3 Result Analysis

As shown in the attribute list, amongst 15 attributes, 5 are used to predict failures whereas 10 indicate end-of-product life.

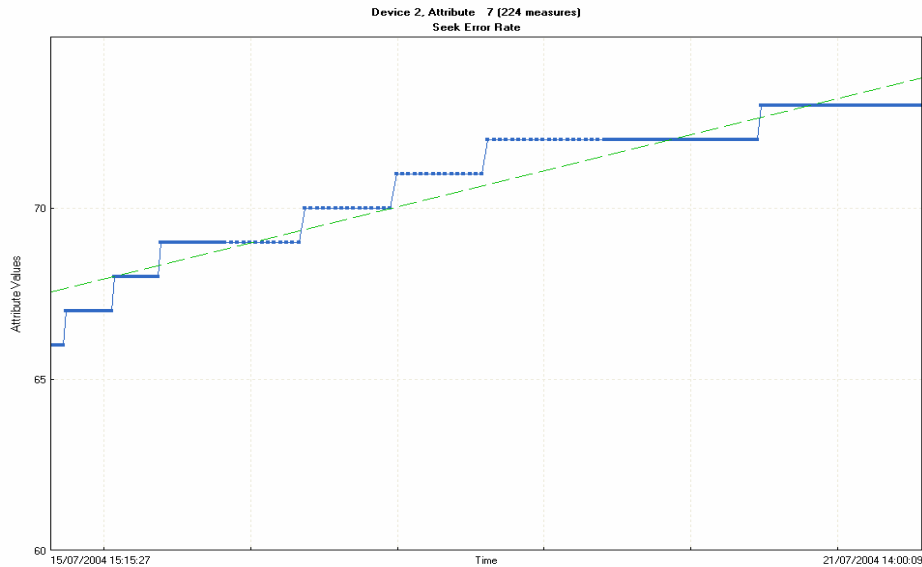
5.1.1.3.1 194_Temperature_Celsius



5.1.1.3.2 1_Raw_Read_Error_Rate



5.1.1.3.3 7_Seek_Error_Rate



5.1.2 Product-moment correlation coefficient r

(DriveSitter implementation)

It was Karl Pearson (1857-1936) at the very beginning of the 20th century who developed a method for finding the best approximated linear representation of the interrelation of two variables in an attempt to help his friend Sir Francis Galton with his research on genetic inheritance: The so-called **product-moment correlation coefficient r** .

To cut a long story short, r ranges between -1 and +1 whereas negative values describe an descending trend, positive values describe an ascending trend and a null value would best be imagined as a straight horizontal trend line. The amount of variation that is accounted for by the trend line can be found by squaring r and is termed the **coefficient of determination r^2** (referred to as "**prediction quality**"). Since r ranges between -1 and 1, r^2 ranges from 0 to +1. The higher r^2 , the better the approximated linear representation; put another way, the higher r^2 , the better the linear approximation fits to the data of the more or less interrelated variables. Note that (theoretically with our understanding of the subject by now) r^2 can be 1 (perfect representation) while r equals 0 (no correlation) but it could not be the other way around.

In addition to this, methods of inference statistics provide means to judge the significance of the correlation coefficient. A correlation coefficient is considered to be significant if it is significantly smaller or greater than 0. Without digging too deep, you should know some simple facts about statistical inference and significance testing. First of all, we collected some data, but usually do not collect all data that exists: There are millions of moments between each pair of measures independent how quick they follow each other. Since we now know that we only have a subset of the whole available data, we do not know whether this subset is a good representation of the whole data, or in other words whether we would find much the same "image" (representation) if we would have collected all data.

In our case, the **level of significance p** is basically the probability that the collected dataset is as extreme or more extreme as the one under consideration if there is in fact **no** correlation. By this, it is judged - based upon our data - whether **no** linear relationship of the two variables is in fact likely. Since p is a probability, it ranges between 0 to 1 whereas smaller values indicate that it is very unlikely that there is in fact no correlation (no linear relationship of the two variables).

To make a decision based upon the level of significance p , statisticians define a so-called **alpha level** and consider every correlation coefficient that falls below this level to be statistically significant. Maybe you remember that we already used the Greek letter *alpha* when talking about statistical errors. To recall it, an alpha error was a "false alarm". And really, the probability of an alpha error (false alarm) is directly related to the defined alpha level: The smaller the defined alpha level, the more rare are alpha errors. On the other hand, the likelihood of a beta error (an unreported malicious condition) increases at the same time the alpha level is lowered. It is an arbitrary convention that the alpha level is set to $\alpha=.05$ as long as there are no other indications.

Summary

As you might have already noticed, statistical calculations and especially statistical inference and significance testing is a somewhat advanced topic. Lets summarize what you should have heard about in this section:

- An easy approximation to the interrelation of two variables is to find the linear "trend". This is most often a complicated task.
- Karl Pearson developed the product-moment correlation coefficient r that describes the interrelation of two variables and includes finding the best linear approximation of their interrelationship. r ranges between -1 and +1 whereas negative values indicate an descending trend, positive values indicate an ascending trend and null indicates
- The coefficient of determination r^2 describes the "quality" of the representation of the two variables by a single straight line. It ranges from 0 to +1 whereas higher values indicate better quality.
- The level of significance p can be used to judge whether the correlation that was found within the analyzed dataset is probably due to sampling errors. p is a probability and thus ranges between 0 and 1 whereas lower values indicate little likelihood that the discovered relationship can be ascribed to sampling errors. Put another way, a low p indicates a high probability that the empirically found correlation represents a real linear relationship of the two variables.
- The level of significance is tested against an alpha level. A level of significance smaller than the alpha level is considered to indicate statistical significance. The alpha level is commonly set to $\alpha=.05$ by convention as long as nothing else is indicated. It has direct impact on the number of alpha and beta errors whereas lowering the alpha level leads to less alpha but more beta errors and vice versa.