

## A More Reliable Way to Predict Potential Disk Failure



### Introduction

Information drives business. Every business activity—profit, non-profit, production, service, mechanical, technical, professional, agricultural, education, and government—relies on information to function. Access to the most accurate, current, and reliable data is critical for business survival, which is why businesses spend so much money on gathering data, storing data, and ensuring the integrity of their information systems. According to Pinheiro, Weber, and Barroso (2007), “It is estimated that over 90% of all new information is being stored on magnetic media, most of it on hard disk drives.”

Ensuring and maintaining the reliability of the data on disks is a huge expenditure for an enterprise in time, money, and human resources. Some estimates are that a third of an information technology budget should be invested in disaster recovery and business continuity activities such as backup, redundancy, and failover to ensure data access and reliability and to prevent or recover from the loss of business-critical data stored on disks. According to Seagate, 60 percent of drive failures are mechanical. Mechanical failures usually happen over a period of time and therefore can often be anticipated. The Self-Monitoring, Analysis, and Reporting Technology (SMART) integrated into most modern disk drives was designed to provide adequate warning.

Business have been basing their spending decisions on well-established concepts of disk performance and reliability. Two recent papers, “Failure Trends in a Large Disk Drive Population,” by Pinheiro, E., Weber, W., and Barroso, L.A. (2007), of Google, Inc., (the Google study), and “Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?” by Shroeder, B., and Gibson, G., (2007) of Carnegie Mellon University (the Carnegie Mellon study), appear to refute or modify some of these accepted axioms about disk reliability.

This paper provides a more reliable perspective of disk failure trends and performance that is predicated on metrics and customer experience with a greater population, a larger variety of disk types, and a wider range of environments and uses. The paper also describes how LSI’s Proactive Drive Health Monitoring solution mitigates issues with disk drive failures and helps prevent unexpected drive failures. LSI’s drive PDHM solution is a complement to SMART that works across drive types, manufacturers, and models.

## Disk Reliability, Business Decisions, and Strategic Planning

In any organization, the most important consideration of data reliability metrics is how the information affects the bottom line—revenue and expenses, profit and loss, and market positioning. Disk reliability data informs decisions and strategies about

information back-up, asset management, and equipment purchases. The information is critical to management in making sound business and strategic planning decisions. To make and justify their decisions, managers need accurate, reliable, and timely information.

The Pinheiro, Weber, and Barroso (2007) study and the Shroeder and Gibson (2007) study appear to refute or modify some long-standing, accepted axioms about disk reliability.

Disk vendors publish expected reliability data for their disks. But the findings of Pinheiro, Weber, and Barroso (2007) and Shroeder and Gibson (2007) from actual field populations provide evidence that actual reliability and failure rates differ from published datasheet rates, and many factors can influence the actual reliability and failure rates that a storage administrator might experience. These factors include workload, environmental conditions, failure policies, and average disk age. These studies highlight that a simple Mean Time to Failure (MTTF) number may not be sufficient to predict actual disk failure rates.

## What is a Disk Failure?

Organizations compile and interpret data differently, so they arrive at different findings and publish different statistics about disk reliability and failure rates. Differences in disk reliability and failure statistics can also be partially attributed to differences in how the term “disk failure” is defined.

A disk is considered failed when the data stored on the disk cannot be accessed, either partially or fully, without human or mechanical intervention. Disks are manufactured mechanical devices, and as mechanical devices, they are susceptible to failure. A variety of reasons, from environmental factors such as fire and water, to internal causes such as firmware failure and damaged sectors, contribute to disk failures. Effective and timely preventive actions such as backup and redundancy can mitigate the potential impact of data loss.

In “Failure Trends in a large Disk Drive Population,” Pinheiro, Weber, and Barroso considered a disk failed when the disk was replaced as part of a repairs procedure. They acknowledge, “It is not always

clear when exactly a drive failed,” and “replacement can sometimes be a few days after the observed failure event.” Pinheiro, Weber, and Barroso state that “Most available data are either based on extrapolation [of lab study data]...or...relatively modest sized field studies.” In the lab, the results are based on small samples in controlled environments and projected onto real populations used in the field, under real-world conditions, and in real-world environments. Field study results are based on empirical data from small field samples, and the results are projected onto large populations.

Customers apply their own experience and criteria to determine when disk failure occurs, and that determination is quite different from how vendors determine disk failure. Vendors publish datasheets with their product MTTF statistics compiled from accelerated lifecycle testing under controlled laboratory conditions. These conditions

include specific usage and duty cycles, controlled temperatures, and other stable environmental standards. The results from the smaller samples assume constant reliability and failure rates when they are projected onto larger populations. However, both lab and field results often fail to account for infant mortality, maintenance, and other normal wear-and-tear factors.

Either way, the results are skewed and do not reflect actual user experience in the real world. To disk manufacturers and vendors, the most important consideration should be real-life customer experiences, rather than controlled laboratory experiences. LSI's concern for field quality, integrity, and reputation are just some of the things that set it apart from its competition. Customer issues are quickly prioritized for resolution with communication back to the customer being the cornerstone of the customer support team's focus.

## Data Reliability, Availability, and Accessibility

Having reliable, highly available, and instantly accessible data is the ultimate goal of storage in the real world. As one of the first and leading suppliers of highly available enterprise disk arrays, LSI considers a disk “failed” when the health monitoring system identifies a real or potential degradation in performance that jeopardizes the health of an array to an extent that threatens data integrity or availability. LSI has tracked actual disk failures for several years and has worked with disk vendors to understand and minimize differences between specified and actual failure rates.

LSI focuses on the customer experience by proactively tracking the manufacturing yields and field qualities of its products. Upper management reviews the metrics monthly, and any anomalies are quickly investigated. Technical task teams analyze the root cause, develop and implement corrective actions, and resolve customer concerns.

### Data Study Populations

Both the Google study and the Carnegie Mellon study looked at approximately 100,000 disks, and the tracking duration was up to five years. Garth Gibson, associate professor of computer science at Carnegie Mellon University and co-author of the study, was careful to point out that the study did not necessarily track actual disk failures but cases in which a customer decided a disk had failed and needed replacement. Gibson also said he has no vendor-specific failure information, and that his goal is not “choosing the best and the worst vendors” but to help them to improve disk design and testing.

The LSI research methodology, metrics, findings, and conclusions are based on more than 900,000 disks— nine times the population in either the Google study or the Carnegie Mellon study. The LSI failure policy is more useful and relevant, and it more accurately reflects real-world concerns and requirements. LSI derives its disk failure data from a larger sample size, a much larger pool of

vendors and customers from more real-world environments, and a wide range of workloads, usages, and temperatures,

rather than controlled laboratory conditions. LSI also compiles its data by separate disk type, rather than combining all disk types. Therefore the data and the results are more relevant.

### Data Review and Analysis

The LSI Customer Quality Engineer Group reviews the disk return data monthly. The returns are evaluated to determine the reason for the return and whether to include the returned in the data used to calculate quality trends. Anomalies in the data are reviewed with the returns coordinator, and a decision is made based on the review to either include or exclude the data in the quality trends calculations. LSI uses the following data collection and analysis methods for calculating quality trends:

- Establish a single, consistent baseline of failure rates. LSI combines the information from the different datasheets from multiple vendors.
- Maintain the integrity of the data and ensure the reliability of the results. LSI identifies and eliminates false data points, that is, data that is not indicative of quality trends. An example of a false data point is a disk returned for a non-warranty reason such as the customer placed an incorrect order. False data points are flagged and excluded from the data used to calculate quality trends.
- Prevent data about field retrofits from distorting MTTF statistics. Field retrofits are disks that were replaced but have not failed and might never fail, so including them would distort MTTF data. Field retrofit data is flagged and removed from returns. A field retrofit occurs when a management or contractual decision is made to proactively replace a disk with a suspected problem.
- Ensure that returned disks are properly diagnosed and categorized. LSI analyzes a statistically valid sample of all returned disks to identify the failure. When a returned disks shows no sign of failing, or if no failure was evidenced in the data the disk reported, the disk is diagnosed as No Fault Found (NFF). Disks diagnosed as NFF returns are included in the Annual Return Rate (ARR) data. LSI uses the NFF rate to closely estimate the true AFR of customers returning disks and uses the data when calculating quality trends.

### Classes of Disks in the Study Populations

The most accurate and useful way to study disk reliability is by compiling and maintaining separate data for each class of disk, which neither the Google study nor the Carnegie Mellon study did.

Both studies included lower-end Serial Advanced Technology Attachment (SATA) disks, Parallel Advanced Technology Attachment (PATA) disks, and Small Computer System Interface (SCSI) disks. The Carnegie Mellon study included only four Fibre Channel (FC) disks. The findings were not broken down based on the different classes of disks under different operating conditions, tracking durations, disk capacities, or usage requirements, all criteria that LSI tracks meticulously.

Conversely, the LSI study used primarily FC disks along with three different SATA disks: consumer, nearline, and offline. LSI tracked disks currently in use in its customer base. The larger population of tracked disks provides more empirical and statistical evidence and stronger support for the data. LSI categorized the data by disk type and even disk vintage. The LSI metrics are more focused, and the findings are more authoritative, reliable, and relevant because they more accurately reflect user needs and experiences.

- Enterprise FC disks are used for networks and systems where data availability, reliability, and integrity supersede—or should supersede—all other considerations. The integrity, availability, and speed of transferring information among network components are critical to the business. Connectivity, capacity, scalability, reliability, security, and performance are paramount.

- Consumer SATA disks cost less than FC disks, so SATA disks are deployed when cost is more of an issue than performance and data availability. SATA disks are appropriate for entry-level servers and low-end servers, in relatively low-usage environments, and as secondary storage devices. SATA disks are not intended for use as primary storage in high-availability, complex enterprise environments. Misusing and abusing consumer SATA disks by deploying them in enterprise environments as a cost-saving measure results in high failure rates, low data reliability, and lost data.
- Nearline and offline SATA disks are intended to be used as secondary storage, where cost is a significant factor, and speed and performance are not as critical. In an enterprise, SATA disks are appropriate for data retention and protection, such as disk-to-disk backup; disaster recovery and business continuity; and archiving, storing, and retrieving data.

DISK PARAMETERS	DISK COUNT	TRACKING DURATION
73GB 10K FC	34,957	September 2002 - Present
146GB 10K FC	135,983	September 2002 - Present
300GB 10K FC	64,210	November 2004 -Present
36GB 10K FC	94,706	December 2002 -Present
73GB 15K FC	302,871	January 2003 - Present
146GB 15K FC	55,086	January 2006 - Present
300 GB 15K FC	2,867	December 2006 -Present
250GB 7200 SATA	117,921	January 2004 - Present
400 GB 7200 SATA	66,384	June 2005 - Present
500GB 7200 SATA	26,005	June 2006 - Present
<b>TOTAL DRIVES TRACKED</b>	<b>900,990</b>	

Table 1 Disk Types, Counts, and Tracking Duration

Table 1 shows the different types of disks and their specifications, the quantity of each type of disk, and the time windows for tracking the reliability of each type of disk.

## Disk Reliability Metrics

The underlying assumption behind reliability statistics is that the devices are used properly, for the purposes intended, under conditions and in environments for which they are intended, within vendor-recommended specifications, and with failure policies similar to those used in the vendor's reliability testing.

### MTTF as a Measure of Disk Reliability

The standard metric published by disk vendors to show disk reliability is an MTTF number, usually stated in hours. The MTTF is the average time projected for how long a set of devices should last until the first failure of a disk when the usage metrics are based on the underlying assumption. LSI's real-world experience has shown that the Annual Failure Rate (AFR) and the ARR are usually higher than would be expected from the vendor's MTTF prediction.

### Computing MTTF Values

MTTF values are computed based on the historical record of actual disk failures in a large number of disks over a large period of time in an actual field- or lab-tested array. A basic MTTF formula is the number of disks times the number of hours divided by the number of disk failures. Figure 1 shows an example of an MTTF calculation.

$$\frac{1000 \text{ drives times } 2400 \text{ hours (100 days)}}{2 \text{ failures}} = \text{MTTF of } 1,200,000$$

Figure 1 Example MTTF Computation

### Defining Mean Time Between Failures

According to the Storage Network Industry Association, the Mean Time Between Failures (MTBF) is the average time from the start of use to the first failure in a large population of identical systems, components, or devices, and is one measure of how reliable a product is. MTBF is usually expressed as hours, and the higher the MBTF, the more reliable the product. The underlying assumption for computing MTBF is that components experience constant failure rates that follow an exponential law of distribution. MTBF calculations also assume that a component is repaired after a failure and immediately returned to service.

Figure 2 shows an example calculation of an MTBF of 1,200,000 hours MTBF.

$$\frac{17520 \text{ hours (2 years) times 68 sample units}}{1 \text{ failed unit}} = 1,191,360 (\sim 1,200,000) \text{ MTBF}$$

Figure 2 Example MTBF Calculation

### Annual Failure Rate as a Measure of Disk Reliability

The AFR is more useful as a measure of disk reliability than is the MTBF. The AFR is derived from the MTBF and is defined as the reciprocal of the MTBF. The AFR is expressed in years and percent. Figure 3 shows an example calculation of a 0.73% AFR. Figure 4 shows an easier calculation for arriving at the same result.

$$\frac{1,200,000 \text{ hours MTBF}}{8760 \text{ hours (per year)}} = 136.9863 \text{ years}$$
$$\frac{1 \text{ failure}}{136.9863 \text{ years}} \times 100\% = 0.73\%$$

Figure 3 Example AFR Calculation

$$(8760 \text{ hours per year} \div 1,200,000 \text{ hours MTBF}) \times 100\% = 0.73\%$$

Figure 4 Simplified Example AFR Calculation

### Annual Return Rate as a Measure of Disk Reliability

The ARR is the number of units returned for a year divided by the number of units shipped for a year. A large number of returned disks show no problem, and are diagnosed as no fault found (NFF). However, the customer's perceived or actual problem that caused the disk return was real.

Disk AFR data is widely available on manufacturer and vendor web sites and datasheets. However, actual ARR data is proprietary and confidential, so ARRs appear only as "less-than" rates, such as < 1%. Figure 5 shows how an example <1% ARR result is derived.

$$\frac{1000 \text{ units returned per year}}{100,001 \text{ units shipped per year}} = < 1\%$$

Figure 5 Example ARR Computation

## Why Actual ARR Data is Higher than Published AFR and MTTF Data

Shroeder and Gibson (2007) identify “a significant discrepancy between the observed ARR and the datasheet AFR for all data sets. While the datasheet AFRs are between 0.58% and 0.88%, the observed ARRs range from 0.5% to as high as 13.5%. That is, the observed ARRs by data set and type are by up to a factor of 15 higher than datasheet AFRs.” This observation agrees with LSI’s real-world experience that the AFR and the ARR are usually higher than would be expected from the vendor’s MTTF prediction.

Because actually running a new disk model long enough to approach the expected MTTF would take longer than vendors sell a disk model, most vendors publish MTTF statistics that have been validated through accelerated life testing. This is accomplished by running a large set of disks for some period of time and then extrapolating the observed failures out to a much longer time period. This requires making assumptions about how the failure rate may change as the disks age. This accelerated life testing is also performed in environments, under workloads, and with disk error recovery times that can be different from actual field usage. From this testing, MTTF hours are often specified at over 1,000,000 hours for disks that LSI finds through actual return rates over several years to be closer to a 500,000 hour MTTF. This translates to an actual AFR of 1.75%.

Like the Google study and the Carnegie Mellon study, LSI identified significant differences between vendor-specified and actual ARRs. In 2006, based on the differences between actual and vendor-specified AFR data, LSI worked with its largest disk vendor to understand the details of their Reliability Demonstration Testing (RDT) procedures, the accelerated life testing that provides the basis of the vendor’s advertised MTTF. LSI identified three key factors that can cause different results in vendor tests and real-world use:

- Differences between controlled RDT laboratory environments and real-world customer data centers
- Differences in disk replacement policies
- Disk damage caused by electrostatic discharge and by damage from disk assembly, integration, and customer usage. These factors are not encountered in a controlled RDT laboratory.

## AFR and ARR as Indicators of Failure Rates

Just as the MTBF and AFR metrics are related, the AFR and ARR metrics are related. The AFR is the percentage of confirmed disk failures in a population in a twelve-month period. The ARR is the percentage of disks returned by the user to the vendor from a population in a twelve-month period. The difference between these two metrics is the confirmation of the disk failure.

The number of disk failures that cannot be confirmed or re-created comprises no more than 1% of the total field population of disks. LSI has found the percentage of NFF disks to be as high as approximately 40% of all disks returned. Therefore, LSI’s AFR

expectation of 1.75% combined with the NFF rate of 40%, means that the ARR would need to be 2.9% to exceed the AFR goal of 1.75%. Figure 6 shows the ARR chart of a specific enterprise model of disk that has been in the LSI customer base for approximately four years.

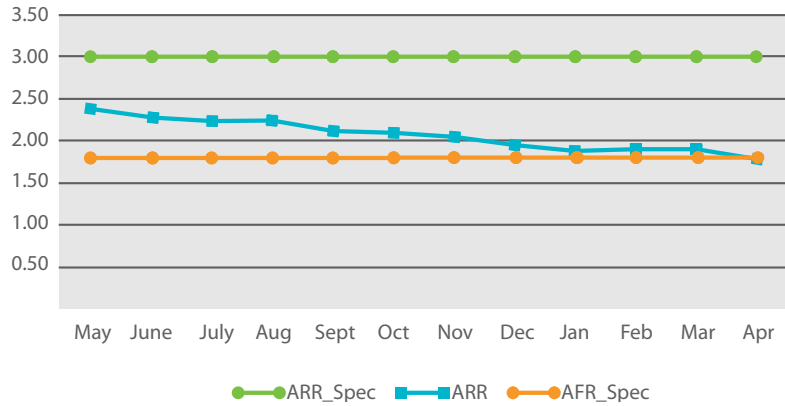


Figure 6 Actual ARR Chart of a Specific Enterprise Model Disk

The graph in Figure 6 shows three metrics for the set of disks. The disk had a vendor-specified MTTF of 1,000,000 hours, which converts to an AFR of 0.73%. LSI actually measured a somewhat steady return rate at approximately 2% over most of the disk’s product life.

Based on an MTTF of 1,000,000, the data demonstrates that the 1.75% failure rate LSI predicted more accurately represents the actual customer experience than does the 0.73% failure rate the vendor predicted. The gap can be attributed to several factors:

- Differences between controlled RDT laboratory environments and real-world customer data centers
- Differences in disk replacement policies
- Disk damage caused by electrostatic discharge or handling during the disk’s assembly, integration, and customer usage. These factors contribute to the increased actual failure rates in the field.
- Differences in the accelerated life testing by the vendor under lab conditions and actual customer usage in the field

### Impact of a Bad Vintage

LSI has observed that the impact of bad vintages on disk reliability from different vendors can be severe. To identify a bad vintage, LSI sorts the failed disks as seen in the field according to the date of manufacture of the disk family, analyzes the failure rate metrics for the disk family, and then isolates the cause of the failure. Potential failure causes could include mechanical issues such as media errors, fabrication line changes, and lubrication breakdown (Schroeder & Gibson, 2007).

Figure 7 shows an example of the impact on the ARR of a disk vintage with a known manufacturing defect.

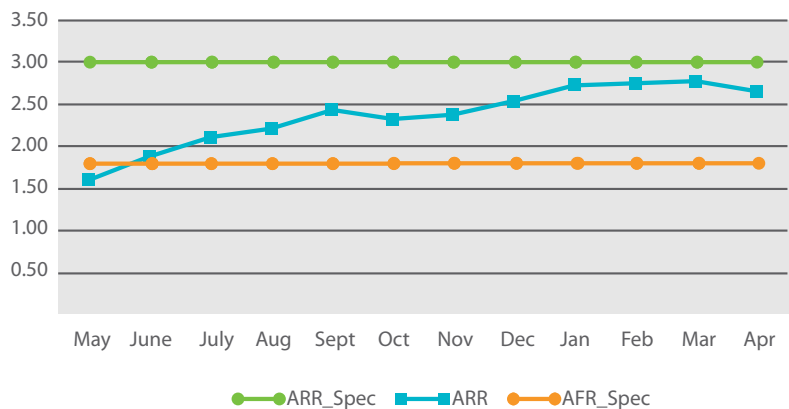


Figure 7 Impact on the ARR of a Disk Vintage with a Known Manufacturing Defect

This FC enterprise disk model has been in the LSI customer base for approximately three years. The disk vendor spent several months investigating the root cause of the problem and was able to make corrective actions just as the disk was reaching the end of its production life.

The return rate shown in Figure 7 crested at 2.78% in November 2006, compared with the approximately 2% ARR shown in Figure 6 for an FC enterprise disk model. The return rate is slowly decreasing as the disks age, probably as a result of the “survival of the fittest” theory described by Pinheiro, Weber, and Barroso (2007). As disks that are vulnerable to this failure mode fail, they are replaced with disks on which the vendor has implemented final corrective action. Although this particular disk was an enterprise FC disk, LSI has seen the bad vintage phenomenon in both enterprise FC disks and in SATA disks.

LSI works hard to detect, resolve, and mitigate the effects of bad disk vintages by constantly monitoring failure rates and by investigating the root cause of failures in the field. LSI then prompts the suppliers to perform corrective actions to their product and processes. LSI also contacts its OEM customers to identify any bad disk vintages and resolve their field issues.

### SATA Disk Reliability

SATA disks are just as reliable as FC disks when used in the proper environment, for the purposes intended, under conditions for which they are intended, and within vendor-recommended specifications. Primary storage systems store business-critical information. Primary storage data requires continuous availability and entails a high number of transactions per second. FC, or enterprise-level, disks meet these specifications. Secondary storage contains business-important information that needs to be online, but the data is accessed sporadically and has sequential large block I/O requirements. Less expensive SATA disks meet these specifications.

LSI has found that the return rates of FC disks and SATA disks are similar. Figure 8 shows the ARR of a SATA disk model, and Figure 9 shows the ARR of an enterprise-level FC disk model. Both models have been in the field population for approximately 2 years.

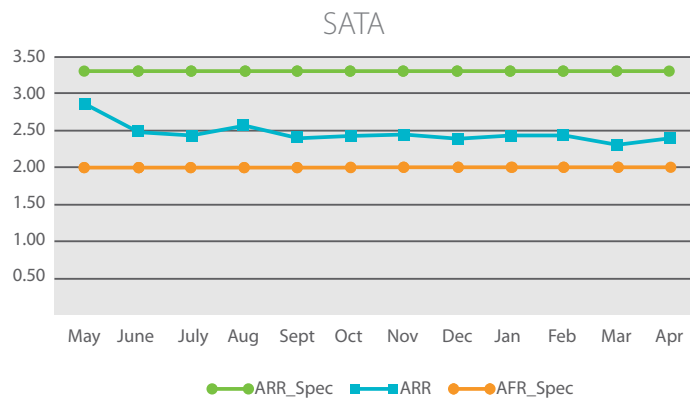


Figure 8 Actual ARR for a SATA Disk Model

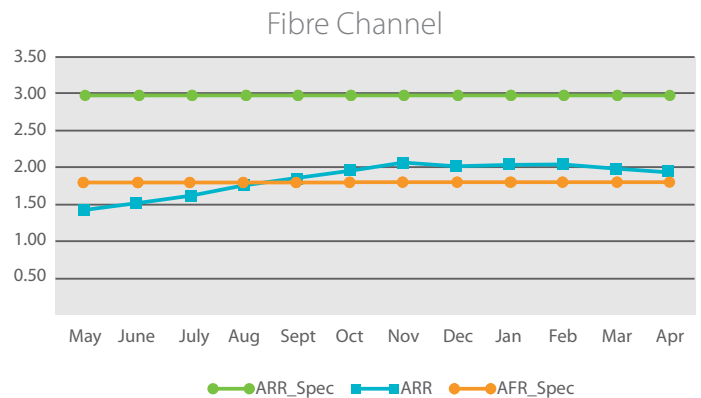


Figure 9 Actual ARR for an FC Disk Model

Both the SATA disk and the FC disk show a fairly steady return rate of approximately 2%. Two significant factors can radically influence disk reliability experiences and have a significant impact on disk return rates.

- Environmental differences such as temperature at customer sites, which Schroeder and Gibson also noted.
- The improper use of SATA disks in non-SATA environments. Figure 10 illustrates how the improper use of SATA disks contribute to an excessive failure rate.

Figure 8 shows a return rate of approximately 2.4% for a SATA disk model used in an appropriate SATA environment. Figure 10 shows a return rate of more than 4% for the same disk model used in an environment not appropriate for SATA. The contrasting return rates underscore the importance of using disks appropriately.

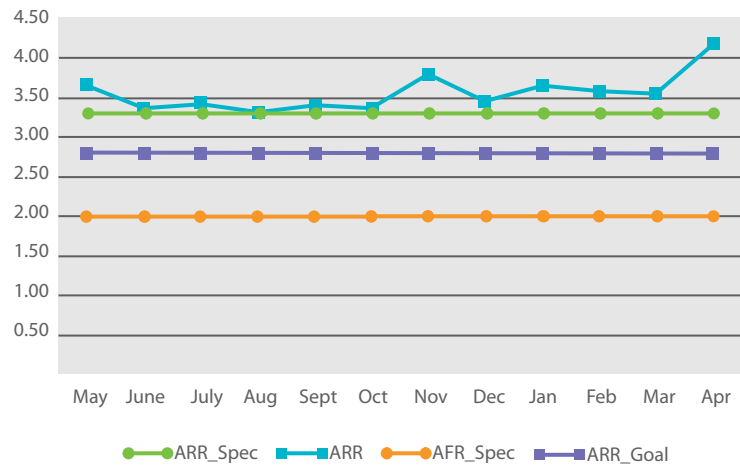


Figure 10 Actual Return Rates for SATA Disks Used in Improper Environments

### Mitigating the Effects of Disk Failure

As with all mechanical devices, disks can eventually fail. To help end-users avoid or mitigate the potential impact of data loss resulting from disk failures, disk manufacturers and vendors maintain extensive databases of data and statistics, such as the MTBF. Disk manufacturers and vendors publish the most current and accurate information available to them about the expected life cycle, reliability, and performance of their disks so end-users can make rational, informed business decisions about data availability and reliability and preventing data loss resulting from disk failure. Business decisions should be based on accurate, comprehensive, and current information about disk performance and reliability. The availability and reliability of business critical information should be the primary and overarching concerns.

## Proactive Drive Health Monitoring

### How PDHM Improves Drive Reliability, Integrity, and Availability

LSI's PDHM solution provides a strong defense against unexpected drive failures. Proactive Drive Health Monitoring is a comprehensive, efficient, and timely way to manage array controller storage data availability. With little or no adverse effect on normal performance, PDHM can significantly reduce the potential of reliability problems developing simultaneously on one or more drives in a RAID group. This capability enhances the data availability of any storage system.

Without PDHM's real-time detection and notification, support personnel might not be aware that a drive is about to fail. With PDHM, support personnel can make informed decisions about replacing a drive before it fails. Proactive Drive Health Monitoring improves data availability, integrity, performance, and productivity for IT staff and array end-users.

- **Enhanced Data Availability** – Multiple drive failures can cause significant data loss, but PDHM prevents reliability issues that could lead to multiple drive failures developing in the same RAID volume.
- **Improved Data Integrity** – Repeated and exhaustive internal drive error recovery procedures can degrade performance and possibly compromise data integrity. Proactive Drive Health Monitoring prevents the continued, and sometimes silent, impact of internal drive reliability problems.
- **Better Array Performance** – PDHM quickly identifies sluggish drives that exhibit consistent performance degradation.
- **Greater Productivity** – PDHM automatically identifies reliability issues before serious problems develop, which means that support personnel can proactively manage RAID arrays and thereby reduce disk downtime.

### PDHM Distinguishes Between Types of Error and Exception Conditions

The controller firmware distinguishes between two types of error and exception conditions:

- **Drive-reported error and exception conditions**, whereby the drive itself reports internal reliability problems
- **Synthesized error and exception conditions**, whereby the controller firmware declares a drive as unreliable based on the error or exception rates the controller firmware counts for each drive

The synthesized error and exception condition and the drive-reported error and exception condition are managed as two separate conditions for each drive; that is, two separate indicators are maintained for each drive. The indicators are reported to administrative clients, such as SANtricity and Simplicity, through a common mechanism. Synthesized PDHM is a rate-based detection method, in which errors are counted within a specified time interval, or window. When the error count exceeds the specified threshold, PDHM concludes that the drive is unreliable.

#### Errors and Exceptions for FC and SAS Drives

Drives that use the SCSI protocol, such as FC drives and SAS drives, return error and exception conditions in the form of sense keys. The controller firmware monitors the rates of three critical drive sense keys for each drive. The ANSI T10 document entitled "SCSI Primary Commands - 3 (SPC-3), T10/1416-D," defines these keys as follows:

- **Recovered Error** – The command completed successfully, with some retry and recovery action performed by the [drive].
- **Medium Error** – The command terminated with a non-recovered error condition that might have been caused by a flaw in the medium or an error in the recorded data. This sense key also can be returned if the [drive] is unable to distinguish between a flaw in the medium and a specific hardware failure.

- Hardware Error – The [drive] detected a non-recoverable hardware failure, such as controller failure, device failure, or parity error, while performing the command or during a self test.

In addition to the sense keys, the number of I/O requests whose response time exceeded a threshold value is also counted. The Excessively Long I/O Time counter increments for only drive I/O requests that exceed a user-defined maximum time to complete.

### Errors and Exceptions for SATA Drives

Serial ATA (SATA) drives also provide error and exception information. The SATA drive-reported error codes follow the ATA interface specification, so the SATA drive-reported error codes are different from the SCSI error codes. Because most of the RAID arrays communicate with host systems and RAID controllers through the SCSI protocol, all of the ATA error codes are mapped to SCSI error codes. After the error codes are mapped, the same counter and threshold mechanisms monitor errors and exceptions reported by SATA drives.

The following SATA error codes are mapped to the SCSI Medium Error:

- IDNF – IDentity Not Found
- UNC – UNCorrectable Error in Data

The following SATA error codes are mapped to the SCSI Hardware Error:

- ICRC – Interface Cyclic Redundancy Code (CRC) error
- ABRT – Command ABORTed
- MC – Media Changed
- MCR – Media Change Request
- NM – No Media

When SATA drives are in the same enclosure as FC drives, the SATA interface must be converted to an FC interface. The conversion process includes mapping ATA status values to corresponding SCSI sense keys.

The LSI custom interposer chip has been designed to perform the following functions, as well as others.

- Polling for internal drive reliability status; that is, drive internal SMART reporting
- SATA to FC conversion
- ATA status to SCSI sense key mapping

When SATA drives are in the same enclosure with SAS drives, the I/O controller's application-specific integrated circuit (ASIC) performs these functions.

SATA drives do not report recovered errors, so no Recovered Error counter is maintained for these drives. Lengthy internal recovered-error processing on SATA drives is detected by the Excessively Long I/O Time counter and threshold. The drive PDHM feature does not count every reported SATA error because some errors indicate a problem with the interposer chip and not the SATA drive. Expected error and exception rates associated with nearline SATA drives are higher than expected error and exception rates for enterprise class FC and SAS drives. Therefore, thresholds for SATA drives are usually set to a higher value than thresholds for FC and SAS drives.

# Drive Counters and Threshold Values in a PDHM Environment

The controller maintains separate sets of configurable threshold values:

- Four threshold values for FC and SAS drives
- Three threshold values for SATA drives

For each drive error and exception counter, the controller firmware maintains a configurable time window value that specifies how long the PDHM count is retained. If the drive counter value does not exceed the threshold within the time window, the counter is cleared and reset to zero. The counting process begins again and overwrites all previously collected data.

If the drive counter value exceeds the threshold within the time window, the drive state changes to reflect the PDHM condition. The controller declares a “Synthesized PDHM” condition for the drive and sets a state flag and reason value. This process calculates the rate of errors and provides real-time detection of problematic drives. Combining the threshold with the time window results in a rate-based detection mechanism for identifying drive reliability problems.

Through its normal event notification mechanism in the controller firmware, SANtricity detects the change in the drive-rate and updates the drive-rate representation in the object graph. SANtricity also sends you an email or SNMP trap, based on how the Enterprise Management Window is configured.

Time windows are measures in hours. If the time window is set to zero for a counter, that counter and threshold is not evaluated or monitored by the controller firmware. Monitoring for that type of drive error or exception is disabled.

For a drive I/O request, only the initial drive-reported I/O error or exception is counted. Subsequent errors or exceptions during retry or the recovery actions are not counted. This process limits counted errors or exceptions to one per drive I/O request.

When one controller detects a change in a drive’s error and exception status, that status is immediately reported to and adopted by the other controller. Then after a controller either resets or reboots, the controllers communicate with each other to ensure that any existing PDHM status on the alternate controller is reported to and adopted by the controller that is resetting or rebooting.

Drive PDHM reporting counters are not persistent when both controllers reboot, such as during a storage system power cycle. However, a synthesized drive PDHM state may be quickly reinstated if the drive continues to report an excessive rate of errors or exceptions after the power-up or the simultaneous reboot of both controllers. When a drive is failed because of a synthesized drive PDHM, the failed drive state is persistent.

## Errors and Threshold Values for FC Drives and SAS Drives

This section describes the error causes, types, levels, and related threshold values for FC drives and SAS drives.

- Recovered Error – This threshold must be set conservatively based on drive specification and array platform performance limits for worst case I/O workload profiles. Some drive-reported recovered errors are normal and can be expected.
- Medium Error – A medium error threshold is set conservatively based on drive specifications of expected medium errors and performance limits for the array platform.
- Hardware Error – Drive reported hardware errors are critical, so the threshold is set to 1.

- Excessively Long I/O Time – An Excessively Long I/O Time is any drive I/O response time that is greater than the specified Excessively Long I/O Time threshold and is less than the array controller’s maximum drive I/O timeout.

Determining the response time is especially important when distinguishing a drive channel problem from an internal drive problem. If a request is successful but took a long time, the problem is probably caused by the drive, not the channel. In the case of an internal drive problem, it is appropriate and prudent to signal a synthesized PDHM condition when the Excessively Long I/O Time counter is exceeded.

However, if a request takes so long that it times out and therefore fails, it is possible that the channel is the problem. Drive channel problems are handled by a controller firmware feature for monitoring and managing drive channel errors. The Excessively Long I/O Time counter should not increment in this case.

In addition to detecting drives that are exhibiting reliability problems or are about to fail, the Excessively Long I/O Time counter helps you quickly identify “slow” drives that can significantly affect performance. An array controller eventually fails a consistently “bad” drive when repeated I/O timeouts prevent the completion of a specific drive I/O request. Therefore, the Excessively Long I/O Time threshold is set lower than the array controller’s gross I/O timeout threshold and higher than the normal expected drive I/O response time. Excessively Long I/O Time detection is triggered only by too many slow drive I/O response times. A few slow drive I/O response times are not sufficient to trigger the detection mechanism if the threshold is set appropriately.

**This type of failure is not a new feature of synthesized PDHM, but has been a part of the drive health monitoring system for some time. The failure occurs before and after PDHM is implemented.**

### Errors and Threshold Values for SATA Drives

The thresholds for SATA drives are independent and different from thresholds for FC and SAS drives. This is necessary because the typical error rates can vary between SATA drives and FC and SAS drives.

- Recovered Error – SATA drives do not report recovered errors, so this threshold is not maintained. However, an Excessively Long I/O Times threshold can be used to detect an excessive rate of internally recovered drive errors.
- Medium Error – A medium error threshold is set conservatively based on drive specifications of expected medium errors and performance limits for the array platform.
- Hardware Error – Drive reported hardware errors are critical, so the threshold is set to 1.
- Excessively Long I/O Time – Controller firmware uses different time-out settings for SATA drives and for FC and SAS drives. SATA drives can exhibit differences in performance characteristics requiring different Excessively Long I/O Time thresholds and counters from those used for FS/SAS drives.

## Drive Failure Options for PDHM

You can set RAID array firmware to automatically fail the drive when error and exception thresholds are exceeded, which starts the reconstruction of the drive’s data On a global hot spare<sup>1</sup>. As you might expect, notification of this action is recorded in the appropriate logs as well as in SANtricity or Simplicity. Drive PDHM alerts can also be automatically forwarded via email.

Alternatively, you can set array firmware to simply use logs, email, and either SANtricity or Simplicity to notify support personnel that a drive has exceeded a drive PDHM threshold. This notification allows support personnel to plan for replacing the drive at an appropriate time or to perform another suitable repair or recovery procedure.

<sup>1</sup>“Global hot spare” is not a standard, universal term. As used here, it refers to an extra drive that can be used when an active drive fails.

Proactive Drive Health Monitoring fails a drive under two conditions: when it has too many errors, and when it is too slow. Other controller functions fail a drive when it has a hard failure. The synthesized drive PDHM feature prevents the continued use of a drive that is clearly exhibiting unreliable behavior. Without PDHM drive reporting, a drive can be sending numerous errors to the array controller, but unless it is a hard failure, the controller keeps the drive going, although usually with slower performance. Not only is performance degraded when an unreliable drive continues to be used, but data can be lost when two drives in a volume have been reporting numerous errors and they both abruptly fail at the same time. Proactive Drive Health Monitoring gives support personnel an opportunity to replace the drives one at a time, and thereby avoid data loss.

The automatic drive failure option should be disabled during the initial installation. Disabling the automatic drive failure option allows the feature to perform an automatic survey of the health of all of the drives in the array storage system, which allows support personnel to establish a safe and controlled plan for replacing drives with reliability problems. In cases where multiple drives in the same RAID volume have reliability problems, you should mirror or back up the data before replacing the first drive that has the most serious reliability problem. After the initial process of identifying and replacing drives with reliability problems is completed, then it is appropriate to turn on the option for automatic drive failure.

**Array controllers do not fail a drive, even when a PDHM threshold is exceeded, if failing the drive causes a loss of data availability, such as when a volume is already degraded or a volume initially has no redundant data. Regardless of the volume type and state, a critical alert is always issued when a drive PDHM threshold is exceeded, so you are always notified about a drive that is developing reliability problems.**

## PDHM and Event Logging

A critical event is logged by the controller when a synthesized or drive-reported PDHM condition is declared for the drive. The synthesized drive log entry identifies which threshold triggered the alert and includes the exception counter value or slow I/O response time counter value for the drive that triggered the event. Support personnel have to manually monitor the log to determine when a drive need to be replaced. Now, LSI's PDHM provides a more-automated way of monitoring the logs for developing reliability problems and eliminates the need for support personnel to manually monitor the logs.

## Calculating I/O Times

It is important to ensure that only the actual duration of each I/O operation is calculated for the Excessively Long I/O Time counter and avoid including the time it spends queued on the drive when the drive is busy. Consider Figure 1, where  $S(R_i)$  is the start time for I/O request  $R_i$ , and  $E(R_i)$  is the end time.

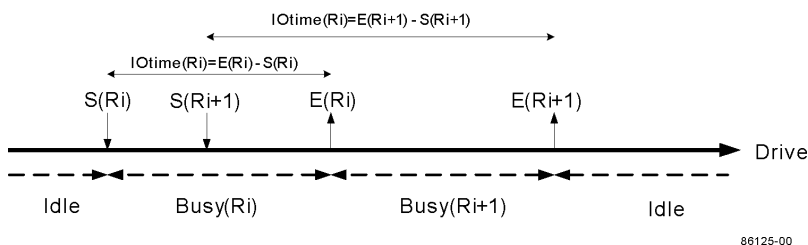


Figure 11 Simple I/O time calculation

A simple  $IOtime(R_{i+1}) = E(R_{i+1}) - S(R_{i+1})$  calculation is not sufficient, because the calculation includes the time the request spent waiting for  $R_i$  to complete in addition to the  $R_{i+1}$  processing time. Instead, for a busy drive, the difference in end times is calculated, for example,  $IOtime(R_{i+1}) = E(R_{i+1}) - E(R_i)$ . This calculation is performed in the RAID controller rather than in SANtricity or Simplicity.

## PDHM and SANtricity

SANtricity is the storage management software available for the high-end RAID arrays from LSI. The software provides a common graphical user interface across all host platforms.

### Needs Attention Icon

SANtricity displays a Needs Attention icon in the Array Management Window for drives that have an error or exception condition that exceeds thresholds. You can select the drive in the physical view to display the drive properties, which shows the cause of the attention indicator. The drive properties also have an indication of whether this condition is synthesized or drive-reported.

### Drive Diagnostic Data

SANtricity supports the menu-driven collection of drive diagnostic data (for all drives in the array) into a file. Each drive entry in the file includes the synthesized PDHM counters for that drive. The file can then be transferred to appropriate support personnel for evaluation.

Additionally, SANtricity supports the menu driven collection of entire array diagnostic data into a file. As you might expect, drive diagnostic data along with PDHM counters are included in this file, which can then be transferred to appropriate support personnel for evaluation.

### Recovery Guru

The Recovery Guru feature of SANtricity has three entries for PDHM conditions indicating impending drive failures:

- High Data Availability Risk – When a PDHM condition exists for a drive assigned to a volume group that either does not have data redundancy (such as a RAID 0 volume), which has lost data redundancy through another drive failing, or through a similar circumstance.
- Medium Data Availability Risk – When a PDHM condition exists for a drive assigned to a volume that has data redundancy.
- Low Data Availability Risk – When a PDHM condition exists for a drive that is unassigned or is a hot spare.

Recovery Guru includes an option for automatically failing drives with PDHM conditions. However, a drive will not be failed if its failure might lead to a loss in data availability, so data availability is not compromised. For example, if the volume is a RAID 0 volume, and the Recovered Error threshold has been exceeded for a drive, the drive is not failed, because that would cause data loss. Instead, as with all synthesized PDHM conditions, SANtricity notifies you of the PDHM state for the drive.

## PDHM and Simplicity

Simplicity is the storage management software available for the entry level RAID arrays from LSI. The software provides a common graphical user interface across all host platforms.

Simplicity reports an error or exception condition as a textual status in the profile for the drive itself, but Simplicity does not display an iconic representation of the status. The drive profile also indicates whether the condition is synthesized or drive-reported.

Simplicity also gives you the ability to collect the entire array diagnostic data into a file, but there is no direct access to just the drive diagnostic data collection.

## Conclusion

The reliability data that disk vendors publish in their datasheets for their disks is different from the real-life experiences in the field. The differences are expected and justifiable because lab testing and the real world use different criteria and metrics for determining what constitutes a failed disk. Some of the differences result from different definitions of what disk failure is. Other differences are between empirical real-life data and controlled laboratory data. Another factor in disk failure statistics is the bad vintage phenomenon. LSI constantly monitors failure trends to identify and resolve bad vintage issues.

A simple MTTF calculation might not be sufficient to predict actual disk failure rates. The ARR, which is derived from MTBF data, is a better predictor of disk reliability. Observed actual AFRs can be significantly higher than are datasheet ARRs compiled from laboratory tests. Datasheet claims are based on accelerated life testing under static laboratory conditions that do not replicate dynamic, real-life conditions.

Actually running a new disk model long enough to approach the expected MTTF takes longer than vendors have a disk model on the market, so most vendors publish MTTF statistics that have been validated through accelerated life testing. Laboratory testing is based on several assumptions, including failure rate changes as disks age, disk-error recovery times, workload fluctuations, and environmental changes. Real-world experiences do not permit assumptions.

The SMART tool integrated into most modern disk drives is an adequate tool for collecting operational drive metrics and statistics. However, SMART is not a reliable tool for predicting when a drive is about to fail. Synthesized PDHM monitors the rate of drive-reported error and detects drive performance degradation often associated with unreported internal drive issues. LSI's drive PDHM reporting tool provides predictive metrics, identifies drives that clearly display behavior that indicates developing reliability problems, and predicts actual drive failure. LSI's SANtricity and Simplicity tools equipped with PDHM proactively notify you when a drive exceeds user-defined thresholds. Proactive Drive Health Monitoring automatically issues a critical alert notification and can optionally fail the drive when a specified error rate or degraded performance threshold is exceeded. Proactive Drive Health Monitoring can significantly reduce the potential of reliability problems developing on one or more drives in a RAID group at the same time. The drive PDHM capability provides a comprehensive, efficient, and timely way to manage array controller storage data availability, and enhances the reliability, integrity, and data availability of any storage system, yet has little or no impact on normal performance,

To learn more about the issues associated with disk failures, to determine the most effective preventive measures, and to find out more about how LSI can help you implement the most advanced drive failure predictive measures for your RAID array, contact your local LSI representative.

## References

Computerworld Inc. // Disk drive failures 15 times what vendors say, study says //Robert L. Scheier;

<http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9012066&source=Quigohttp%3A%2F%2Fwww.computerworld.com%2Faction%2F%2Farticle.do%3Fcommand%3DviewArticleBasic%26articleId%3D9012066>

Pinheiro, E., Weber, W., and Barroso, L.A., 2007. "Failure Trends in a Large Disk Drive Population," 5th USENIX Conference on File and Storage Technologies (FAST'07), February 2007. USENIX Association.

Shroeder, B., and Gibson, G., 2007. "Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?" 5th USENIX Conference on File and Storage Technologies (FAST'07), February 2007. USENIX Association.

## Ownership of Materials

The Document is provided as a courtesy to customers and potential customers of LSI Corporation ("LSI"). LSI assumes no obligation to correct any errors contained herein or to advise any user of liability for the accuracy or correctness of information provided herein to a user. LSI makes no commitment to update the Document. LSI reserves the right to change these legal terms and conditions from time to time at its sole discretion. In the case of any violation of these rules and regulations, LSI reserves the right to seek all remedies available by law and in equity for such violations. Except as expressly provided herein, LSI and its suppliers do not grant any express or implied right to you under any patents, copyrights, trademarks, or trade secret information. Other rights may be granted to you by LSI in writing or incorporated elsewhere in the Document.

## Trademark Acknowledgments

Engenio, the Engenio design, MegaRAID, HotScale, SANtricity, and SANshare are trademarks or registered trademarks of LSI Corporation. All other brand and product names may be trademarks of their respective companies.

## Performance Information

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of LSI products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they want to purchase.

## Disclaimer

LSI has provided this Document to enable a user to gain an understanding of the LSI Proactive Drive Health Monitoring (PDHM) tool ("the Tool") when used in conjunction with LSI Storage Systems. This Document and the Tool referenced in it are designed to assist a user in making a general decision as to whether an LSI Storage System configuration is appropriate for such user's objectives. Neither this Document nor the Tool are designed or intended to guarantee that the configuration a user chooses will work in a specific manner. While the guidance provided by this Document and the Tool can help a user to choose an appropriate configuration (or avoid a configuration that is not appropriate), there is no way LSI can guarantee the exact performance and/or results of the Information contained in this Document. Accordingly, LSI assumes no obligation whatsoever for the use of the Information provided in this Document or the Tool, AND UNDER NO CIRCUMSTANCES WILL LSI OR ITS AFFILIATES BE LIABLE UNDER ANY CONTRACT, STRICT LIABILITY, NEGLIGENCE OR OTHER LEGAL OR EQUITABLE THEORY, FOR ANY SPECIAL, INDIRECT, INCIDENTAL, PUNITIVE OR CONSEQUENTIAL DAMAGES OR LOST PROFITS IN CONNECTION WITH THIS DOCUMENT OR THE TOOL.

THE INFORMATION AND MATERIALS PROVIDED IN THIS DOCUMENT AND THE TOOL ARE PROVIDED "AS IS" AND LSI MAKES NO WARRANTIES EXPRESS, IMPLIED OR STATUTORY, INCLUDING, WITHOUT LIMITATION, THE IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE WITH RESPECT TO THE SAME. LSI EXPRESSLY DISCLAIMS ANY WARRANTY WITH RESPECT TO ANY TITLE OR NON-INFRINGEMENT OF ANY THIRD PARTY INTELLECTUAL PROPERTY RIGHTS, OR AS TO THE ABSENCE OF COMPETING CLAIMS, OR AS TO INTERFERENCE WITH USER'S QUIET ENJOYMENT.

For more information and sales office locations, please visit the LSI web sites at: [lsi.com](http://lsi.com) [lsi.com/contacts](http://lsi.com/contacts)

### North American Headquarters

Milpitas, CA  
T: +1.866.574.5741 (within U.S.)  
T: +1.408.954.3108 (outside U.S.)

### LSI Europe Ltd.

European Headquarters  
United Kingdom  
T: [+44] 1344.413200

### LSI KK Headquarters

Tokyo, Japan  
Tel: [+81] 3.5463.7165

LSI Corporation and the LSI logo design are trademarks or registered trademarks of LSI Corporation. All other brand and product names may be trademarks of their respective companies.

LSI Corporation reserves the right to make changes to any products and services herein at any time without notice. LSI does not assume any responsibility or liability arising out of the application or use of any product or service described herein, except as expressly agreed to in writing by LSI; nor does the purchase, lease, or use of a product or service from LSI convey a license under any patent rights, copyrights, trademark rights, or any other of the intellectual property rights of LSI or of third parties.

Copyright ©2008 by LSI Corporation. All rights reserved. > 1208

