A Highly Accurate Method for Assessing Reliability of Redundant Arrays of Inexpensive Disks (RAID)

Jon G. Elerath, *Member*, *IEEE*, and Michael Pecht, *Fellow*, *IEEE*

Abstract—The statistical bases for current models of RAID reliability are reviewed, and a highly accurate alternative is provided and justified. This new model corrects statistical errors associated with the pervasive assumption that system (RAID group) times-to-failure follow a homogeneous Poisson process, and it corrects errors associated with the assumption that the time-to-failure and time-to-restore distributions are exponentially distributed. Statistical justification for the new model uses theories of reliability of repairable systems. Four critical component distributions are developed from field data. These distributions are for times to catastrophic failure, reconstruction and restoration, read errors, and disk data scrubs. Model results have been verified to predict between 2 and 1,500 times as many double disk failures as estimates made using the mean time-to-data-loss (MTTDL) method. Model results are compared to system-level field data for a RAID group of 14 drives and show excellent correlation and greater accuracy than either MTTDL or Markov models.

Index Terms—Monte Carlo simulation, redundant systems, reliability modeling, repairable systems.

1 Introduction

The concept of a redundant array of inexpensive disks for data storage was patented by N. K. Ouchi in 1978 (US Patent 4,092,732). In the late 1980s, it was popularized to increase performance and combat the ever-increasing probability of failure associated with large pools of hard disk drives (HDDs) [1]. Reliability estimates were created assuming that HDD failures were dominated by catastrophic failures in which the HDD simply could no longer serve data. Degradation of the media, termed "bit rot," was the primary suspect for data that simply were no longer properly recorded on the media, but bit rot was determined to be a very low-probability event and inconsequential relative to catastrophic failures. There was little justification for assuming that times-to-failure were not exponentially distributed with constant failure rates [2].

Recently, field data have shown that the mean time-to-data-loss (MTTDL) model for predicting RAID reliability is highly inaccurate [3]. The three causes for this inaccuracy, errors in statistical theory of repairable systems, incomplete consideration of failure modes, and inaccurate time-to-failure distributions, will be discussed in detail.

Markov models appear to promise more accurate estimates of RAID reliability and have been explored numerous times [4], [5], [6], [7]. However, Markov models resolve only one of the three errors; they allow inclusion of additional HDD failure modes. They do not correct errors in statistical theory of renewable systems or input distributions.

This paper combines the theory of known statistical principles for repairable systems with the practical aspect of actual field data to yield a highly accurate model. A statistically defensible model is developed and solved using Monte Carlo simulations. The new model corrects the three errors that render current methods erroneous. The results indicate that MTTDL calculations underestimate the true number of double disk failures (DDFs) by orders of magnitude.

Past-RAID model methods are presented in Section 2. Statistical errors associated with MTTDL and Markov models, as well as the bases for accurately modeling repairable systems, are presented in Section 3. Section 4 presents additional HDD failure mechanisms and modes that are critical to an accurate model, and Section 5 presents the new model. The four input distributions, developed from measured field data, actual system operating rules, and physical limitations of HDDs, are discussed in Section 6, and comparisons of model predictions to field results are presented in Section 7.

2 Previous Work

Numerous papers have been written on RAID reliability, a few of which are [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], and [11]. Most often, the system reliability and number of DDFs, leading to a loss of data, are estimated using the MTTDL [1], [2], [7], [8], [9], [10], [11], a rather straightforward expression that assumes the time-to-failure and time-to-restore distributions are exponentially distributed with constant failure and restoration rates. The system MTTDL is also assumed to follow a homogeneous Poisson process,

- *J.G. Elerath is with NetApp, 495 E. Java Dr., Sunnyvale, CA 94089. E-mail: jon.elerath@netapp.com.*
- *M. Pecht is with the Center for Advanced Life Cycle Engineering (CALCE), University of Maryland, Room 1103, Engineering Laboratory (Building 89), College Park, MD 20742. E-mail: pecht@calce.umd.edu.*

Manuscript received 27 Feb. 2008; revised 15 July 2008; accepted 23 July 2008; published online 4 Sept. 2008.
Recommended for acceptance by C. Bolchini.
For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number TC-2008-02-0093.
Digital Object Identifier no. 10.1109/TC.2008.163.

having a constant failure rate. MTTDL attempts to estimate average time between system failures, defined as the concurrent failure of two HDDs in an $N+1$ RAID group. The expression for MTTDL, in which the HDD failure rate is λ and the restoration rate is μ, is given as

$$\text{MTTDL} = \frac{(2N+1)\lambda + \mu}{N(N+1)\lambda^2}. \quad (1)$$

Since the repair rate is usually much larger than the failure rate, the MTTDL expression can be simplified as

$$\text{MTTDL}_{Indep} = \frac{\mu}{N(N+1)\lambda^2} = \frac{\text{MTTF}_{disk}^2}{N(N+1)\text{MTTR}_{disk}}. \quad (2)$$

The expected number of DDFs is calculated assuming that the system has a constant rate of occurrence of failure, that the rate of occurrence is the inverse of the MTTDL, and that the number of DDFs can be estimated by multiplying the rate of occurrence of failure with the number of hours and systems at risk, as is the case in the renewal theory.

Corrupted data encountered during reconstruction of a catastrophic failure leads to DDFs and data loss. This scenario of events occurs with far greater frequency than two concurrent catastrophic failures, yet it is not included in the MTTDL calculation. Kari [4] first addressed the issue of latent defects but assumed they were caused only by media deterioration (bit rot) and were independent of usage. Although bit rot is not a real concern, there are numerous other causes of data loss in today's HDDs even though data were correctly and accurately written to the HDD initially. Some of the causes of latent defects are both spin-time and usage-time dependent and have been identified in other research [3].

Markov models have been proposed as alternatives to the MTTDL calculation. The potential advantage of a Markov model is the ability to add any number of additional states corresponding to conditions such as degraded operation or latent defects. Although Markov models offer an improvement over the MTTDL in that they can represent additional failure states and can model more than one failure and repair rate, they still require constant transition rates.

Geist and Trivedi [6] used a Markov model with two different failure rates in an attempt to account for correlations that exist between the first and second failures in an $N+1$ RAID group, although he did not include latent media defects. Schwarz et al. [7] used a Markov model for assessing the effectiveness of several data scrubbing schemes in a RAID-10 configuration of an offline archive system, but the analysis did not include large RAID groups.

As with the MTTDL model, the Markov models also assume that failures follow a homogeneous Poisson process and that transition rates for failures and restorations are constant. Studies by Shah and Elerath [12], Pinheiro et al. [13], and Schroeder and Gibson [14] present field data indicating that failure rates are not constant in time. Elerath and Magie [15] show that HDD failure rates are also dependent upon random variation in the manufacturing process, and design and manufacturing changes may significantly alter the time-to-failure distribution for the better or worse. Time-to-failure distributions can be mixtures of multiple distributions because of production vintages, thereby creating nonconstant failure rates, even if the underlying rates of the two mixtures are constant [16].

Practical logic, which will be presented in detail later, suggests that a constant rate for restoration of catastrophic failures is equally incorrect.

In both the MTTDL and Markov models, it is assumed that system failures follow a homogeneous Poisson process and have a constant rate of occurrence. However, even though all the system components (HDDs, in this case) have constant failure rates, there is no statistical basis for the assertion that the repairable systems (RAID groups) will have a constant rate of occurrence of failure [17]. This assumption, to a large degree, calls into question the results of any Markov model, even if latent defects are included.

The errors and inadequacies that affect both the MTTDL and all Markov models raise the question of their accuracy and suggest that an improved model is needed. The new model must account for latent defects, allow rates for component failures, component restorations, latent defects, and data scrubbing to take on any distribution, and most importantly, allow the system failures to follow nonhomogeneous Poisson processes.

3 Modeling Repairable Systems

One of the most significant contributions this paper adds to RAID reliability modeling is the elimination of the assumption that components and the system must be modeled by homogeneous Poisson processes. Field data show that failures rarely have constant failure rates, and logic dictates that restorations have a minimum time to complete greater than zero and a maximum time less than infinity. The most egregious assumption in all models to date is the assumed relationship between component failure (hazard) rates and system rates of occurrence of failure.

Ascher [17] points out that there is little connection between the properties of component hazard rates and the properties of the process that produces a sequence of failures. That is, times between successive system failures can become increasingly larger even if each component hazard rate is increasing [18]. Even if the HDDs have constant failure rates, there is no statistical basis for assuming that the system will be an HPP (have a constant failure rate).

In order to appreciate Ascher's statements above and the impact they have on RAID reliability models, we need to understand two different statistical concepts and how they differ from each other. The first concept is commonly referred to as a hazard rate, $h(t)$, and the second is the rate of occurrence of failure, ROCOF. Note that when the hazard rate is constant, it is commonly called a failure rate and denoted by λ. We will explore these through two hypothetical tests.

3.1 Test 1

We will test 10 components sequentially. All component failures follow the same time-to-failure distribution. When a component fails, it is removed from the test, the next is installed and it is then run to failure. For each component, we record the time from its start to failure and run until all 10

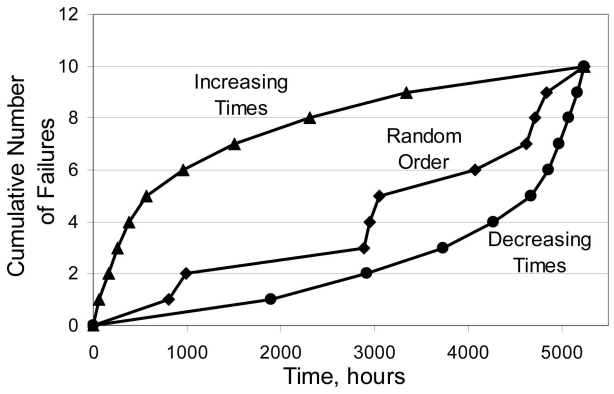

Fig. 1. Interarrival time sequences of the same data.

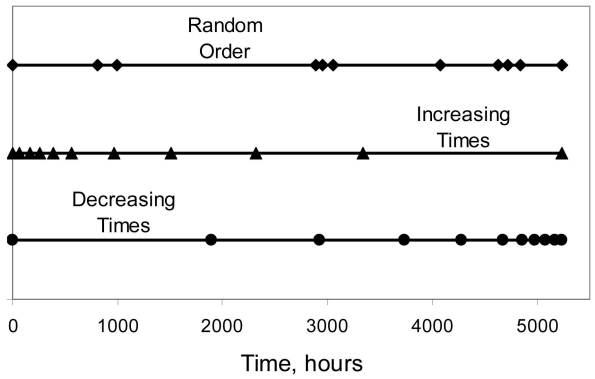

Fig. 2. Cumulative number of failures based on sequence of interarrival times. The sequential order affects the cumulative number of failures seen at any point in time.

have failed. At the end, we have generated 10 independent and identically distributed (iid) times-to-failure.

3.2 Test 2

We create a system made up of 10 components. All component failures follow the same time-to-fail distribution. When a component fails, it is replaced. There are an infinite number of spares for replacement. System failure is defined as two components failed simultaneously. For each system failure, we record the time since the last system failure. There is no end to this test since we have an infinite supply of spares. Note three things: 1) At some point in time, there will be more system failures than components in the system; that is, at some time there may be 25 system failures. 2) The times between system failures are not independent, since when one component is replaced, there are still nine others that have been running for some unknown amount of time. The same nine were in use at the time of two (or more) system failures, so the time to the next system failure depends on the ages of all the components in the system. 3) The total number of system failures is infinite or, practically, unknown.

3.3 Test Data Analysis

Test 1 generated component times-to-failure that are iid. The data can be analyzed by ordering them from shortest to longest time-to-failure and using standard statistical techniques for assessing fit to a distribution. For this data, order of occurrence is unimportant.

Test 2 generated "interarrival" times, t_i, that are dependent. The time-to-failure $N+1$ depends on the ages of all the components in use at that time. As time moves on and numerous replacements occur, the ages of the components can be anything from 0 (the most recent replacement) to T, the full duration of the test, where $T = \Sigma t_i$. For this test, order is critical. The data should NOT be ordered and subjected to standard statistical analyses. The sequence of the interarrival times should first be analyzed, retaining the order in which they occur, to determine whether a trend exists, either increasing or decreasing.

Using the same 10 interarrival times, a graphical example of what happens when repairable system data are treated as component data is shown in Figs. 1 and 2. The upper line in Fig. 1 shows the sequence occurring in a random pattern. The middle line shows an increasing sequence of interarrival times, and the bottom shows a decreasing sequence of interarrival times. The significance of these is observed in Fig. 2, in which the cumulative number of system failures is plotted for the three different sequence assumptions.

Thompson [19] presents a more theoretic commentary on the distinction between the system rate of occurrence of failure and component failure rate. Let $N(t)$ be the cumulative number of failures a system experiences in time and $E[N(t)]$ be the expected number of failures as a function of time. The rate of occurrence of failure is the derivative of the expected number, as given by

$$\text{ROCOF} = \frac{d}{dt} E[N(t)]. \tag{3}$$

Thompson [20] goes on to note that the number of components that fail as a function of time, $N(t)$, is a binomially distributed random variable, i.e.,

$$\Pr[N(t) = k] = \binom{n}{k} F(t)^k [1 - F(t)]^{n-k} \quad for\ k = 0, 1, \ldots n. \tag{4}$$

For a binomial distribution, the expected value, $E[N(t)]$, is equal to $nF(t)$. However, replacing $E[N(t)]$ with $nF(t)$ and completing the differentiation leads to an interesting observation as shown by

$$\text{ROCOF} = \frac{d}{dt}\frac{E[N(t)]}{n} = \frac{d}{dt}F(t) = f(t) \neq \frac{f(t)}{1 - F(t)}. \tag{5}$$

The rate of occurrence of failure of a system is a density function, not a hazard (failure) rate. However, the MTTDL and Markov methods both use the expected value for a binomial distribution to calculate the expected number of system failures as a function of time. MTTDL assumes that 1/MTTDL is the system failure rate (ROCOF), assumes that $\lambda t \approx F(t)$ for small values of λt, and it estimates $E[N(t)] = RG\lambda t$, where RG is the total number of RAID groups. Markov models compute $F(t)$ as an output and, thus, simply multiply $(RG)F(t)$. However, as already noted, the failure rate of a component, $h(t)$, is statistically different from the rate of occurrence of failure (ROCOF) for a system [17], [18], [19], [21], [22], even if the component hazard rate is constant.

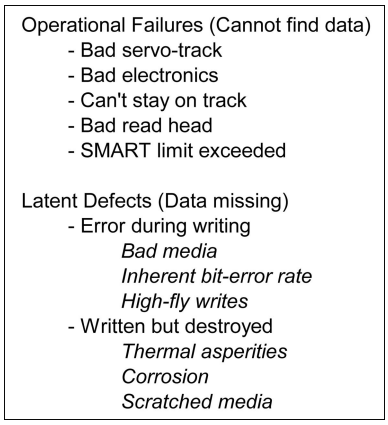

Fig. 3. Breakdown of read error causes.

The model used in these analyses does not rely on component attributes but rather determines the system ROCOF directly through Monte Carlo simulation. This error is therefore eliminated.

4 HDD Latent Media Defects

Latent defects, which refer to data that are unknowingly corrupted or destroyed, occur with greater frequency than operational (catastrophic) HDD failures. Failure modes and mechanisms based on HDD electromechanical and magnetic events are summarized in Fig. 3, grouped by one of two possible consequences: operational failures or latent defects.

Each group has its own failure distribution and consequence at the system level. All read failures can be classified as 1) HDD incapable of finding the data or 2) data missing or corrupted. The failure mechanisms presented here are not novel [3], [23], but neither are they readily available from HDD manufacturers.

4.1 Cannot Find Data

The inability to "find" data is most often caused by "operational" failures, which can occur any time the HDD disks are spinning and the heads are staying on track. Heads must read "servo" wedges that are permanently recorded onto the media during the manufacturing process and cannot be reconstructed with RAID if they are destroyed. These segments contain no user data but provide information used solely to control the positioning of the read/write heads for all movements. If servo-track data are destroyed or corrupted, the head cannot correctly position itself, resulting in loss of access to user data even though the user's data are uncorrupted. Servo tracks can be damaged by scratches or thermal asperities.

Tracks on an HDD are never perfectly circular. The present head position is continuously measured and compared to where it should be and a position error signal is used to properly reposition the head over the track. This repeatable run-out is all part of normal HDD head positioning control.

Nonrepeatable run-out caused by mechanical tolerances from the motor bearings, excessive wear, actuator arm bearings, noise, vibration, and servo-loop response errors can cause the head positioning to take too long to lock onto a track and ultimately produce an error. High rotational speeds exacerbate this mechanism in both ball and fluid-dynamic bearings.

HDDs use self-monitoring analysis reporting technology (SMART) to predict impending failure based on performance data. For example, data reallocations are expected and many spare sectors are available on each HDD, but an excessive number in a specific time interval will exceed the SMART threshold, resulting in a "SMART trip."

Currently, most head failures are due to changes in magnetic properties. Electrostatic discharge (ESD), physical impact with microcontaminants, and high temperatures can accelerate magnetic degradation. ESD-induced degradation is difficult to detect and can propagate to full failure when exposed to localized heat from thermal asperities. The HDD electronics are attached to the outside of the HDD. DRAM and cracked chip capacitors have also been known to cause failure.

4.2 Data Missing

Data are sometimes written poorly initially or can it be corrupted after being correctly written. Unless corrected, missing and corrupted data remain as "latent defects."

4.2.1 Errors during Writing

The bit error rate (BER) is a statistical measure of the effectiveness of all the electrical, mechanical, magnetic, and firmware control systems working together to write (or read) data. Most bit errors occur on a read command and are corrected, but since written data are rarely checked immediately after writing, bit errors can also occur during writes. BER accounts for a fraction of defective data written to the HDD, but a greater source of errors is the magnetic recording media that coats the disks.

Writing on scratched, smeared, or pitted media can result in corrupted data. Scratches can be caused by loose, hard particles (TiW, Si_2O_3, C) becoming lodged between the head and the media surface. Smears, caused by "soft" particles such as stainless steel and aluminum, will also corrupt data. Pits and voids are caused by particles that were originally embedded in the media during the sputtering process and subsequently dislodged during the final processing steps, during the polishing process to remove embedded contaminants, or during field use. Hydrocarbon contamination (machine oil) on the disk surface can result in write errors as well.

A common cause for poorly written data is the "high-fly write." The heads are aerodynamically designed to have a negative pressure and maintain the small, fixed distance above the disk surface at all times. If the aerodynamics are perturbed, the head can fly too high, resulting in weakly (magnetically) written data that cannot be read. All disks have a very thin film of lubricant on them for protection from head-disk contact, but lubrication build-up on the head can increase the flying height.

4.2.2 Data Written but Destroyed

Most RAID reliability models assume that data will remain undestroyed except by degradation of the magnetic properties of the media ("bit rot"). Although it is correct that media can degrade, this failure mechanism is not a significant cause. Data can become corrupted any time the disks are

spinning, even when data are not being written to or read from the disk. Three common causes for erasure are thermal asperities, scratches and smears, and corrosion.

Thermal asperities are instances of high heat for a short duration caused by head-disk contact. This is usually the result of heads hitting small "bumps" created by particles embedded in the media surface during the manufacturing process. The heat generated on a single contact may not be sufficient to thermally erase data but may be sufficient after many contacts.

Heads are designed to push particles away, but contaminants can still become lodged between the head and disk. Hard particles used in the manufacture of an HDD, such as Al_2O_3, TiW, and C, can cause surface scratches and data erasure any time the disk is rotating. Other "soft" materials such as stainless steel can come from assembly tooling. Soft particles tend to smear across the surface of the media, rendering the data unreadable. Corrosion, although carefully controlled, can also cause data erasure and may be accelerated by T/A-generated heat.

5 NHPP-Latent Defect Model

RAID reduces the probability of data loss by grouping multiple inexpensive HDDs into a redundant configuration. Most RAID configurations use a single additional HDD within the RAID group for redundancy and add error correction by using parity, a part of the write process that performs an "exclusive OR" calculation on the user data and saves it on one or more HDDs in the RAID group. Correcting user data by using parity requires that all other HDDs in the RAID stripe are read (including the parity disk) and the user data recreated. This is a relatively slow process as compared to error correcting codes (ECCs) on the HDD, which also improves data integrity.

ECC uses Boolean operations to encode blocks of data on a single HDD, interleaving the data and the ECC bits. On each read command, user data and ECC are read. If a data inconsistency occurs, the data are corrected on the fly (less than one revolution), data integrity is preserved, and performance is not degraded. ECC strength is enhanced by interleaving multiple blocks of data so that errors covering a large physical area (many bits) can be corrected. ECC is faster than data recovery across multiple HDDs, but since ECC is read with every block of user data, excessive ECC use can degrade performance. ECC on the HDD and parity across the HDDs are commonly used together to ensure accurate data recording and transfer.

Since ECC is performed on the HDD, the NHPP model includes only the logic associated with the redundant HDDs in the RAID group. The model is evaluated using sequential Monte Carlo techniques to simulate the time-dependent, or chronological, behavior of the RAID group [3], [24], which allows all input distributions to take on any time-dependent form, not just exponential. For each HDD in the RAID group, each of the four transition distributions is sampled. The operating and failure times are accumulated until a specified mission time is exceeded. This research uses a mission of 87,600 hours (10 years). During that time, the sequence of HDD failures, repairs, latent defects, scrubs, and DDFs is tracked. Each sequence of sampling required to reach the mission is a single simulation and represents one possible system operating chronology. If 10,000 simulations are needed to develop the cumulative failure function, described in [25], it is equivalent to monitoring the number of DDFs for 10,000 systems over the mission life. The times to RAID group failures are then presented as charts showing the cumulative number of DDFs as a function of time.

Four distributions are required for the model: time to operational failure, time to latent defect, time to operational repair, and time to scrub (latent defect repair). An operational failure (Op) is one in which no data on the HDD can be read, even though the data may have no defect. Removal and replacement of the HDD is the only resolution for operational failures. Latent defect (Ld) refers to unknown or undetected data corruption. If only a few blocks of data are corrupted, the reconstructed data are written to another good section of the faulty HDD and the defective section is mapped out to prevent reuse. To correct a latent defect, all other disks in the RAID group must be readable and the associated data in the stripe must be correct and uncorrupted.

System failure occurs when two HDDs fail simultaneously. The order of occurrence of operational and latent defects is significant. If an operational failure occurs after the existence of a latent defect on a different HDD, the data cannot be reconstructed on the replacement HDD because the required redundant data are corrupted or missing. Thus, a latent defect followed by an operational failure results in a DDF. Write errors that occur during reconstruction of an HDD will be corrected the next time the data are read or will remain as latent defects, but their creation during a reconstruction does not constitute a DDF. The probability of suffering a usage-related data corruption in an unread area during the time of reconstruction is small, so DDFs rarely occur during reconstructions. Multiple HDDs with latent defects do not constitute DDF unless they happen to coexist in blocks from a single data stripe across more than one HDD, an extremely rare event that is not modeled.

System designers have realized the potentially devastating impact of latent defects and so eliminate or reduce them by data scrubbing. During scrubbing, data on the HDD is read and checked against its parity bits even though the data are not being requested by the user. The corrupt data are corrected, bad spots on the media are mapped out, and the data are saved to good locations on the HDD. Since this is a background activity, it may be rather slow so as not to impede performance. Depending on the foreground I/O demand, the scrub time may be as short as the maximum HDD and data-bus transfer rates permit, or it may be as long as weeks.

In short, the two scenarios that result in DDF are given as follows: 1) two simultaneous operational failures and 2) an operational failure that occurs after a latent defect has been introduced and before it is corrected. Multiple simultaneous latent defects do not constitute failure.

The NHPP model logic is shown in Fig. 4. In state 1, the data and parity HDDs are good, there are no latent defects, and a spare HDD is available. Failure transitions depend on the number of HDDs available and the distribution of time-to-failure or restoration. A generic functional notation, $g[a; b]$, is used to represent transitions and the critical variables "a"

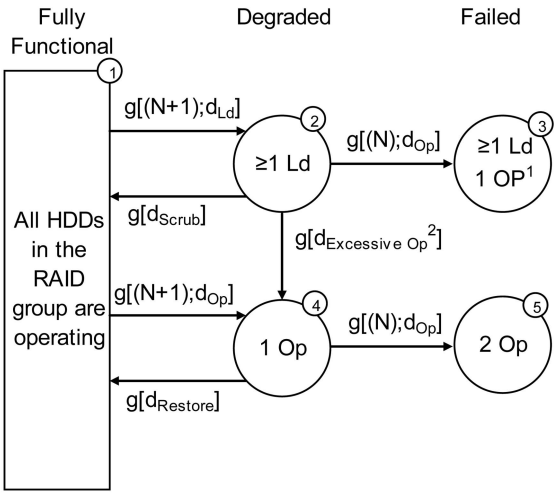

Fig. 4. State diagram for $N+1$ RAID group. Note 1: Op failure must be a different HDD from the one with the Ld. Note 2: This transition does not have an explicit rate. It is included in the measured rate of Op from field data.

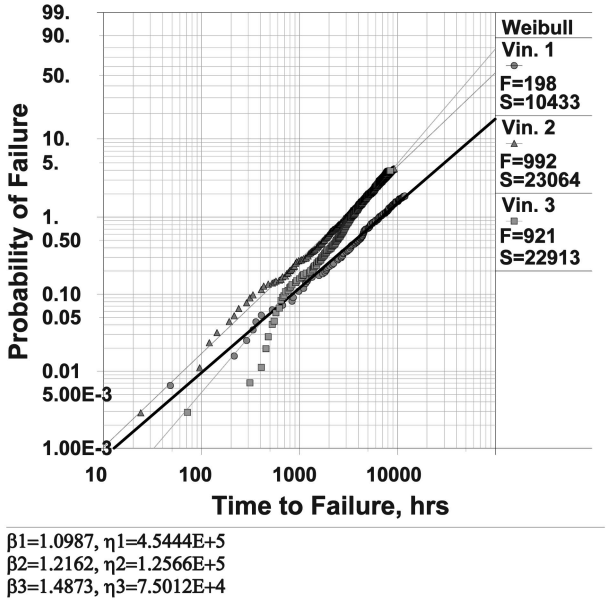

Fig. 5. HDD vintage data: All fit a Weibull time-to-failure distribution, but only vintage 1 has a constant failure rate $(\beta = 1)$.

and "b" without conveying any specific operation. The lowercase letter "d" represents an arbitrary distribution form, and the subscript is the name of the distribution. For example, the transition from state 1 to state 2, denoted $g[(N+1), d_{\mathrm{Ld}}]$, is a function of the $N+1$ HDDs developing a latent defect according to the failure distribution, d_{Ld}.

In state 2, one or more HDDs have latent defects. From state 2, an operational failure in any of the N HDDs other than the one with the latent defect results in state 3, a DDF state. The transition from state 2 to state 3 is governed by the N HDDs and the operational failure distribution, d_{Op}. Transition from state 2 to state 4 occurs because the time to reallocate a sudden burst of media defects on a single HDD exceeds a user-specified threshold. This results in a "time-out" error or SMART trip such as "excessive block reallocations." In this transition, massive media problems render the HDD inoperative, just like any other operational failure, so the frequency of transition from state 2 to state 4 is included in the operational failure distribution, d_{Op}. A third transition from state 2 is back to state 1. This represents repair of latent defects according to the scrubbing distribution, d_{Scrub}.

State 4 represents one operational failure. The transition to state 4 from state 1 is a function of the number of HDDs in the RAID group and the operational failure distribution, d_{Op}. There are two transitions out of state 4. A second simultaneous operational failure results in transition to DDF state 5, the second DDF state. Alternatively, from state 4 the operational failure can be replaced with a new HDD and data reconstructed according to the restoration distribution, d_{Restore}, returning the RAID group back to state 1 with full operability. The restoration distribution, d_{Restore}, includes the delay time to physically incorporate the spare HDD and has a minimum time to reconstruct based on the HDD capacity, the maximum transfer rate, and concurrent I/O.

6 Transition Distributions

This section presents the bases for selecting the four component-related distributions required for this model; time to operational failure, time to restore an operational failure, time to generation of a latent defect, and time to scrub HDDs for latent defects. The simulations in this paper use a three-parameter Weibull probability density function, $f(t)$, with location parameters γ, characteristic life η, and shape parameter β, as shown by

$$f(t) = \left(\frac{\beta}{\eta}\right)\left(\frac{t}{\eta}\right)^{\beta-1} exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right]. \quad (6)$$

6.1 Time to Operational Failure (TTOp)

Recent field data analyses [12], [13], [14], [26], [27] show that HDD failure distributions are anything but constant. Data for specific HDD products often indicate subpopulations with specific characteristics, such as "infant wear-out." Fig. 5 shows data for three different, nonconsecutive HDD vintages from a single manufacturer. The plot shows the probability of failure versus time. A straight line indicates it fits a Weibull time-to-failure distribution. If the slope (β) is 1.0, then the distribution is exponential and has a constant failure rate. As indicated by the values of β in the legend at the bottom left of the chart, only one subpopulation vintage appears to have a constant failure rate, labeled Vin. 1. Vintages 2 and 3 both exhibit increasing failure rates but to different extents, since the β's are different values ($\sim$1.21 and $\sim$1.48).

A Weibull failure distribution with a slightly increasing failure rate is used for these analyses. The characteristic life, η, is 461,386 hours, and the shape parameter, β, is 1.12, both based on measured field data. These parameters are from a field population of over 120,000 HDDs that operated for up to 6,000 hours each. A graphic depiction of this probability density function is shown in Fig. 6.

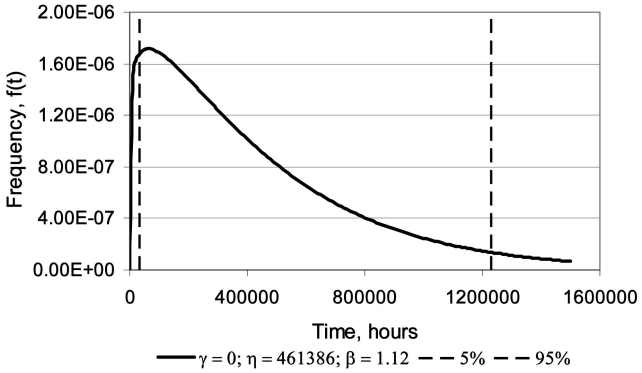

Fig. 6. Probability density function for operational failures, including the times for 5 percent and 95 percent cumulative failures.

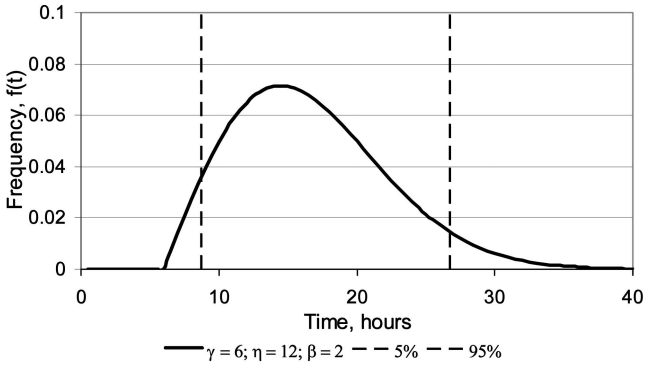

Fig. 7. Probability density function for time to restore an operational failure: 5 percent completion occurs at 8 hours and 95 percent completion occurs around 27 hours.

6.2 Time to Restore an Operational Failure (TTR)

A constant restoration rate implies that the probability of completing the restoration in any time interval is equally as likely as any other interval of equal length. Therefore, it is just as likely to complete restoration in the interval 0 to 48 hours as it is in the interval 1,000 to 1,048 hours. However, this is clearly unrealistic for two reasons. First, the time required for the HDD to reconstruct all the data on the HDD is finite and is a function of the HDD capacity, the data rate of the HDD, the data rate of the data bus, the number of HDDs on the data bus, and the amount of I/O transferred as a foreground process. Reconstruction is performed on a high priority basis but does not stop all other I/O to accelerate completion.

This model recognizes that there is a minimum time before which the probability of being fully restored is zero. Fiber channel HDDs can sustain up to 100 Mbyte/s data transfer rates, although 50 Mbyte/s is more common. The data bus to which the RAID group is attached has a capability of only 2 Gbps. Thus, a RAID group of 14 144-Gbyte fiber channel HDDs on a single data bus will require a minimum of 3 hours to reconstruct the failed HDD with no other I/O. A 500-Gbyte Serial ATA HDD on a 1.5-Gbit data bus will require 10.4 hours to read all other HDDs and reconstruct a replaced HDD.

The added I/O associated with continuing to serve data will lengthen the time to restore an operational failure. Some operating systems place a limit on the amount of I/O that takes place during reconstruction, thereby assuring reconstruction will finish in a prescribed amount of time. This results in a maximum reconstruction time. The minimum time of 6 hours is used for the location parameter. The shape parameter of 2 generates a right-skewed distribution, and the characteristic life is 12 hours. The probability density function for the time-to-restore distribution is shown in Fig. 7.

6.3 Time to Latent Defect (TTLd)

Personal conversations with engineers from four of the world's leading HDD manufacturers support the contention that HDD failure rates are usage dependent, but the exact transfer function of reliability as a function of use (number of reads and writes, lengths of reads and writes, sequential versus random) is not known (or they are not telling anyone). These analyses approximate use by combining read errors per byte read and the average number of bytes read per hour. The result is shown in Table 1 and the following discussion is the justification.

Schwartz et al. [7] claim that the rate of data corruption is five times the rate of HDD operating failures. NetApp completed a study in late 2004 of 282,000 HDDs used in RAID architecture. The read error rate (RER), averaged over three months, was 8×10^{-14} errors per byte read. At the same time, another analysis of 66,800 HDDs showed an RER of approximately 3.2×10^{-13} errors per byte. A more recent analysis of 63,000 HDDs over five months showed a much-improved 8×10^{-15} errors per byte read. In these studies, data corruption was verified by the HDD manufacturer as an HDD problem and not a result of the operating system controlling the RAID group.

Although Gray and van Ingen [28] asserts that it is reasonable to transfer 4.32×10^{12} bytes/day/HDD, the study of 63,000 HDDs read 7.3×10^{17} bytes of data in five months, an approximate read rate of 2.7×10^{11} bytes/day/HDD. The following studies used a high of 1.35×10^{10} bytes/hour and a low of 1.35×10^{9} bytes/hour. Using combinations of the RERs and number of bytes read yields the hourly read failure rates in Table 1. The graphic depiction of the exponential distribution used for the time to latent defect is shown in Fig. 8.

6.4 Time to Scrub (TTScrub)

Latent defects (data corruptions) can occur any time the disks are spinning. However, these defects can be eliminated by "background scrubbing," which is essentially preventive maintenance on data errors. Scrubbing occurs during times of idleness or low I/O activity. During scrubbing, data are

TABLE 1
Range of Average RERs, in Errors per HDD per Byte Read per Hour

Read errors per Byte per HDD		Bytes Read per hr. Low Rate 1.35^{+9}	High Rate 1.35^{+10}
Low	8.00^{-15}	1.08^{-5}	1.08^{-4}
Med	8.00^{-14}	1.08^{-4}	1.08^{-3}
High	3.20^{-13}	4.32^{-4}	4.32^{-3}

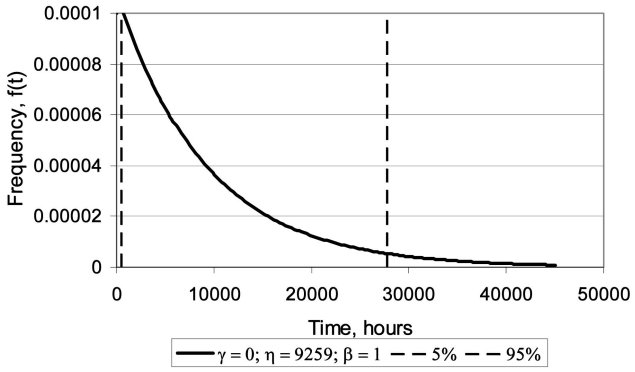

Fig. 8. Probability density function for time to latent defect, showing times for 5 percent and 95 percent cumulative failures.

read and compared to the parity. If they are consistent, no action is taken. If they are inconsistent, the corrupted data are recovered and rewritten to the HDD. If the media is defective, the recovered data are written to new physical sectors on the HDD, and the bad blocks are mapped out.

Scrubbing is a background activity performed on an as-possible basis so it does not affect performance. If not scrubbed, the period of time to accumulate latent defects starts when the HDD first begins operation in the system. The latent defect rate is assumed to be constant with respect to time ($\beta = 1$) and is based on the error generation rate and the hourly data transfer rate.

As with full HDD data reconstruction, the time required to scrub an entire HDD is a random variable that depends on the HDD capacity and the amount of foreground activity. The minimum time to cover the entire HDD is based on capacity and foreground I/O. The operating system may invoke a maximum time to complete scrubbing. In all cases, the shape parameter, β, is 3, which produces a normal-shaped distribution after the delay set by the location parameter, γ. The distribution for the time to scrub is shown in Fig. 9.

7 Results

The results are divided into two sections. Section 7.1 shows the power of the NHPP model, shows sensitivities to latent defects and scrub time distributions, and compares the results to calculations made with the MTTDL method. All models analyzed in this section show time-dependent behavior of DDF over an 87,600-hour (10-year) mission for an $N + 1$ RAID group of eight HDDs (seven data and one parity disk). Section 7.2 compares the model results to actual field data. The RAID group size was increased to match the field RAID group sizes.

TABLE 2
Base Case Input Parameters

Operational Failure Distributions						Latent Defect Distributions					
TTOp			TTR			TTLd			TTScrub		
γ	η	β	γ	η	β	γ	η	β	γ	η	β
0	461386	1.12	6	12	2	0	9259	1	6	168	3

7.1 Parametric Analyses for the NHPP Model

The "base case" data used in these analyses have already been discussed and justified in Section 6. The parameters for the input distributions are summarized in Table 2.

For a Weibull distribution, when the shape parameter $\beta = 1.0$, the distribution degenerates to an exponential distribution in which the parameter η is the MTBF. Therefore, to compare model results to the MTTDL method, the MTBF is 461,386, and the mean time to restore is 12 hours. Using the expression for MTTDL in (2) results in an MTTDL of 37,037 years. Using the common (but erroneous) method to estimate the number of DDFs as described in Section 3 results in 0.27 DDFs per 1,000 RAID groups in 10 years. If this were plotted in Fig. 10, the line for MTTDL would be on the horizontal axis.

Fig. 10 compares the base case (including latent defects and 168-hour scrub) to cases in which the scrub times are 12, 48, and 336 hours. Notice that the plot lines are not linear, showing the effects of the time-dependent failure and restoration rates. The increasing rate of occurrence of failure (ROCOF) for the RAID group is verified by finding the number of DDFs that occur in any fixed time interval (Fig. 11) for scrub times of 168 hours and infinity (no scrubbing).

The increasing ROCOF in Fig. 11 provides evidence that the RAID group does NOT follow a homogeneous Poisson

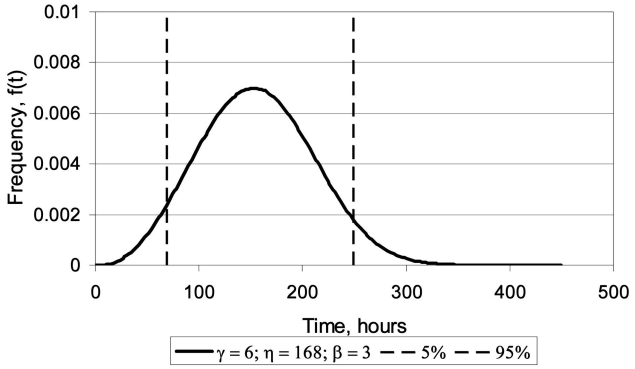

Fig. 9. Time to scrub distribution for the base case: 5 percent of the scrubs will be completed in 70 hours and 95 percent will be completed by 250 hours. None will be completed in less than 6 hours.

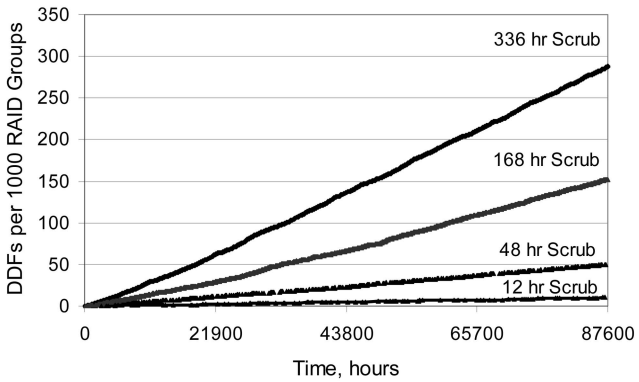

Fig. 10. Effects of scrub time when the latent defect rate is 1.08×10^{-4}/hour.

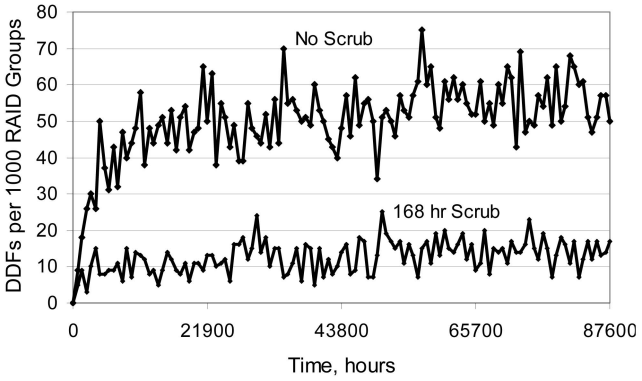

Fig. 11. ROCOFs for two scrub times.

process when using representative field data for the input distributions.

Fig. 12 compares the number of DDFs per 1,000 RAID groups for two different usage rates, both discussed in Section 6.3 and shown in Table 1. The base case assumes the RER is $1.08 \times 10^{-4}/\text{hour}$ of use, or a mean time between read errors of 9,259 hours. Alternatively, the low-usage case estimates an RER of $1.08 \times 10^{-5}/\text{hour}$, a mean time between read errors of 92,950 hours. Even when the latent defect rate is low, the number of DDFs is expected to be 38/1,000 RAID groups in 10 years as compared to the MTTDL estimate of 0.27/1,000 RAID groups in 10 years, a difference of more than 100 times.

This research and new model show a clear difference between the estimated number of DDFs as a function of time based on the MTTDL and the new model. The number of DDFs predicted by the model is, in all cases, greater than the MTTDL when latent defects are included. Without scrubbing, and assuming the distributions in Table 2, this model estimates that in 1,000 RAID groups there will be over 1,200 DDFs in the 10-year mission, contrary to the 0.3 predicted by MTTDL. Table 3 shows the ratio of DDFs expected with the new model to the number estimated using the MTTDL during the first year alone. The highest ratio, $>2{,}500$, is when latent defects are included but there is no scrubbing. Even if scrubbing is completed in 168 hours, the new model predicts over 360 times as many DDFs as the MTTDL method.

TABLE 3
DDF Comparisons

Assumptions	DDFs in 1st year	Ratio
MTTDL	0.03	1
Base Case w/o Scrub	78	2593
336 hr Scrub	21	700
168 hr Scrub	11	360
48 hr Scrub	5	150
12 hr Scrub	1	33

7.2 Comparison to Field Data for RAID Groups

The most meaningful comparison, which validates the model and the associated assumptions, comes from looking at the cumulative number of DDFs in the field for a given RAID size and comparing it to the model results. This comparison is based on a RAID group size of 14 (13 data disks and 1 parity). The HDD operational failure rate is not the same as the base case in Table 2 but is changed to reflect the best estimate for this specific HDD, based on field data. The only variable is the scrub distribution.

Fig. 13 shows the number of DDFs as a function of operating hours in the field (age). There is excellent correlation between the model and the field data for the scrub distribution with $\eta = 48$ hours up until just after 13,140 hours.

This divergence is explained by looking at the quantity that achieved the age of 13,140 hours (Fig. 14). More than 7,000 RAID groups had been in the field for 2,190 hours, but at 13,140 hours of age, the number has decreased to approximately 1,200. This significant reduction in population is the primary reason for the divergence.

Another critical aspect of this analysis is the confirmation that real RAID groups do not follow an HPP. That is, the RAID groups themselves do not have a constant failure rate. This is because the ROCOF is not constant. By taking the derivative of the number of DDFs in time, (dN/dt), it is clear that the number of DDFs in any time interval is not constant but is increasing even at the very lowest ages. The ROCOF (Fig. 15) becomes erratic at 13,870 hours due to the small number of RAID group failures.

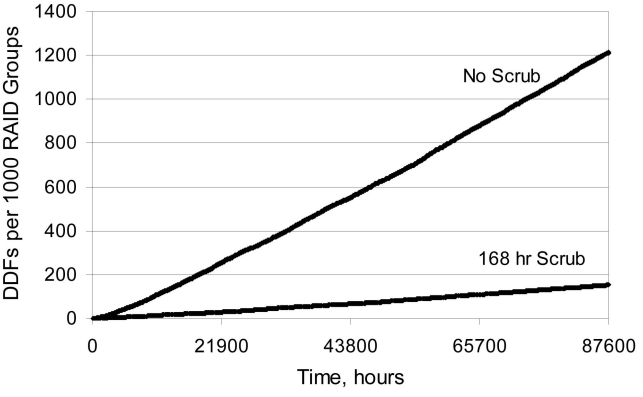

Fig. 12. Effects of latent defect occurrence rates with a 168-hour scrub distribution.

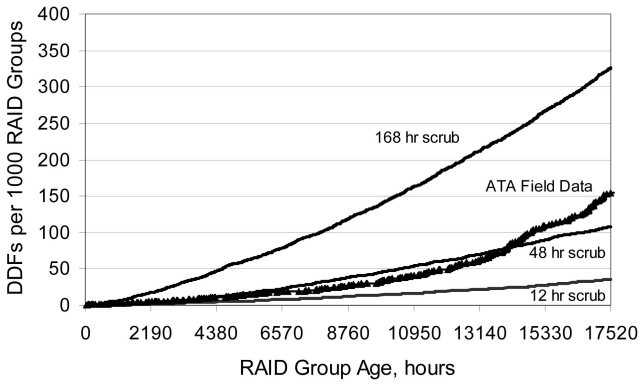

Fig. 13. Number of DDFs as a function of age for ATA field population. RAID group size is 14.

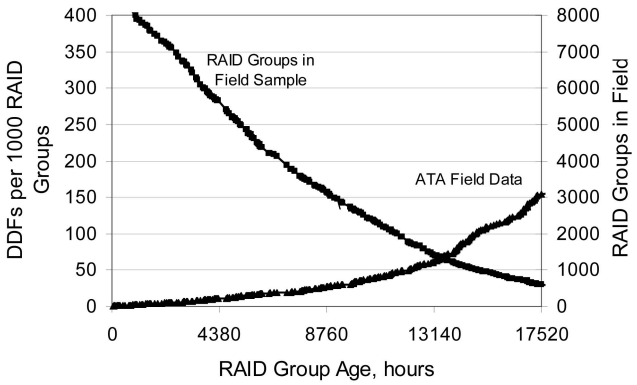

Fig. 14. Quantity and number of DDFs as a function of age for ATA population. RAID group size is 14.

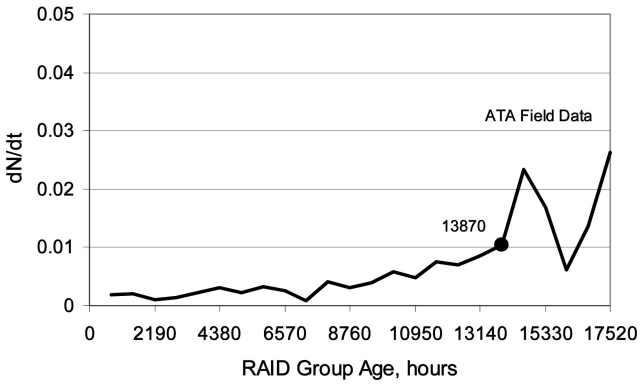

Fig. 15. ROCOF for ATA RAID groups of size 14.

8 Future Enhancements

Two areas of future enhancement and research are discussed here. These relate to improving two distributions and developing specific, simple guidelines for use by architectural developers of hardware and operating systems.

Of the four input distributions, the time to operational failure has the best underpinnings from field data. The time to restore an operational failure has a great deal of intuition embedded and will be fairly easy to collect. However, improvements in this distribution are not expected to have a significant impact on the model results.

Although extensive data were used to determine the RERs, this could be a fruitful area for further research. Simply knowing the average number of read errors as a function of time is not as useful as knowing the distribution of these errors. That is, while this analysis assumed that the errors were equally likely to occur on any surface of any HDD, in reality, specific disk surfaces may experience a large number of errors (for example, due to a scratch). This uncertainty is a prime reason for the sensitivity study on the RER.

Time to scrub, like the time to restore, is fairly well bounded, and improvements are not likely to provide different model results. However, this is one area that could be explored more, because it has the potential to drive decisions made by the developers of the operating system.

This model is fairly complex and requires a great deal of care to exercise correctly. Therefore, a future area of development would be to extract the truly critical distributions or assumptions and to provide guidelines to the software developers so they do not need to run countless Monte Carlo studies.

9 Conclusions

This paper has presented the statistical bases to justify improving the model to assess reliability of RAID groups. The results confirm the significant difference between the MTTDL and the NHPP model methods and show excellent correlation to actual field data for the number of DDFs as a function of time.

The effects of latent defects and scrubbing are explored and shown to be critical in the reliability analyses of RAID systems. Latent defects are inevitable, and scrubbing latent defects is imperative to RAID $N+1$ reliability. As HDD capacity increases, the number of latent defects will also increase and render the MTTDL method less accurate for future products.

Although scrubbing is a viable method to eliminate latent defects, there is a trade-off between serving data and scrubbing. As the demand on the HDD increases, less time will be available to scrub. If scrubbing is given priority, then system response to demands for data will be reduced. Manufacturers of RAID systems will have to address this contention between scrubbing and serving data, and this model allows them to understand the reliability impact of such trade-offs.

Although Monte Carlo simulations are more complex than simple, closed-form calculations such as MTTDL, the improved accuracy justifies the extra effort. The result is a flexible approach that allows use of any distributional form for the inputs and any size of $N+1$ RAID group. The model results show that including time-dependent failure rates and restoration rates along with latent defects yields estimates of DDFs that are as much as 4,000 times greater than the MTTDL-based estimates. Additionally, the ROCOF for a RAID group is not linear in time and depends heavily on the underlying component failure distributions.

Although reliability will be improved when all HDDs are formatted to 4,000-byte blocks (rather than 512-bytes or 520-bytes), this change will be a significant disruption to current system designs. A second alternative to accept latent defects and increase system reliability is to increase redundancy to $N+2$, RAID 6 [29].

References

[1] D.A. Patterson, G.A. Gibson, and R.H. Katz, "A Case for Redundant Arrays of Inexpensive Disks (RAID)," *Proc. ACM SIGMOD '88*, pp. 109-116, June 1988.

[2] G.A. Gibson, "Redundant Disk Arrays: Reliable, Parallel Secondary Storage," PhD dissertation, Dept. Computer Science, UC Berkeley, T7.6 1991 G52 ENGI, Apr. 1991.

[3] J.G. Elerath, "Reliability Model and Assessment of Redundant Arrays of Inexpensive Disks (RAID) Incorporating Latent Defects and Non-Homogeneous Poisson Process Events," PhD dissertation, A. James Clark College of Eng., Mechanical Eng. Dept., Univ. of Maryland, https://drum.umd.edu/dspace/handle/1903/6733, 2007.

[4] H.H. Kari, "Latent Sector Faults and Reliability of Disk Arrays," PhD dissertation, TKO-A33, Helsinki Univ. of Technology, http://www.cs.hut.fi/~hhk/phd/phd.html, 1997.

[5] T.J.E. Schwarz, "Reliability and Performance of Disk Arrays," PhD dissertation, Dept. Computer Science, UC San Diego, 1994.
[6] R. Geist and K. Trivedi, "An Analytic Treatment of the Reliability and Performance of Mirrored Disk Subsystems," *Proc. 23rd Int'l Symp. Fault-Tolerant Computing (FTCS '93),* pp. 442-450, June 1993.
[7] T.J.E. Schwarz, Q. Xin, E.L. Miller, D.D.E. Long, A. Hospodor, and S. Ng, "Disk Scrubbing in Large Archival Storage Systems," *Proc. 12th IEEE/ACM Int'l Symp. Modeling, Analysis, and Simulations of Computer and Telecommunications Systems (MASCOTS),* 2004.
[8] D.A. Patterson, P. Chen, G. Gibson, and R.H. Katz, "Introduction to Redundant Arrays of Inexpensive Disks (RAID)," *Proc. 34th IEEE Computer Soc. Int'l Conf.: Intellectual Leverage (COMPCON '89),* pp. 112-117, Feb. 1989.
[9] P.M. Chen, E.K. Lee, G.A. Gibson, and R.H. Katz, "RAID: High-Performance, Reliable Secondary Storage," *ACM Computing Surveys,* 1994.
[10] W.V. Courtright II, "A Transactional Approach to Redundant Disk Array Implementation," PhD thesis, CMU-CS-97-141, School of Computer Science, Carnegie Mellon Univ., May 1997.
[11] T.J.E. Schwarz and W.A. Burkhard, "Reliability and Performance of RAIDs," *IEEE Trans. Magnetics,* vol. 31, no. 2, pp. 1161-1166, Mar. 1995.
[12] S. Shah and J.G. Elerath, "Reliability Analysis of Disk Drive Failure Mechanisms," *Proc. Ann. Reliability and Maintainability Symp. (RAMS '05),* pp. 226-231, Jan. 2005.
[13] E. Pinheiro, W.D. Weber, and L.A. Barroso, "Failure Trends in Large Disk Drive Population," *Proc. Fifth USENIX Conf. File Storage Technologies (FAST '07),* Feb. 2007.
[14] B. Schroeder and G. Gibson, "Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You," *Proc. Fifth USENIX Conf. File and Storage Technologies (FAST '07),* Feb. 2007.
[15] J.G. Elerath and S. Magie, "Field Reliability from Post-GA Manufacturing Process and Design Changes," *Proc. DISKCON Asia-Pacific,* http://www.idema.org/_smartsite/modules/local/data_file/show_file.php?cmd=download&data_file_id=1441, May 2006.
[16] F. Proschan, "Theoretical Explanation of Observer Decreasing Failure Rate," *Technometrics,* vol. 5, pp. 375-383, 1963.
[17] H. Ascher, "Statistical Methods in Reliability: Discussion," *Technometrics,* vol. 25, no. 4, pp. 320-326, Nov. 1983.
[18] H.E. Ascher, "A Set-of-Numbers is NOT a Data-Set," *IEEE Trans. Reliability,* vol. 48, no. 2, pp. 135-140, June 1999.
[19] W.A. Thompson, "On the Foundations of Reliability," *Technometrics,* vol. 23, no. 1, pp. 1-13, Feb. 1981.
[20] W.A. Thompson, "The Rate of Failure Is the Density, Not the Failure Rate," *The Am. Statistician,* Editorial, vol. 42, no. 4, pp. 288-291, Nov. 1988.
[21] L.H. Crow, "Evaluating the Reliability of Repairable Systems," *Proc. Ann. Reliability and Maintainability Symp. (RAMS '90),* pp. 275-279, Jan. 1990.
[22] W. Nelson, "Graphical Analyses of System Repair Data," *J. Quality Technology,* vol. 20, no. 1, pp. 24-35, Jan. 1988.
[23] V. Prabhakaran, "IRON File Systems," *Proc. 20th ACM Symp. Operating Systems Principles (SOSP '05),* pp. 1-15, Oct. 2005.
[24] C.L.T. Borges, D.M. Falcao, J.C.O. Mello, and A.C.G. Melo, "Composite Reliability Evaluation by Sequential Monte Carlo Simulation on Parallel and Distributed Operating Environments," *IEEE Trans. Power Systems,* vol. 16, no. 2, pp. 203-209, May 2001.
[25] D. Trindade and S. Nathan, "Simple Plots for Monitoring Field Reliability of Repairable Systems," *Proc. Ann. Reliability and Maintainability Symp. (RAMS '05),* pp. 539-544, Jan. 2005.
[26] J.G. Elerath and S. Shah, "Disk Drive Reliability Case Study: Dependence upon Head Fly-Height and Quantity of Heads," *Proc. Ann. Reliability and Maintainability Symp. (RAMS '03),* Jan. 2003.
[27] S. Shah and J.G. Elerath, "Disk Drive Vintage and Its Affect on Reliability," *Proc. Ann. Reliability and Maintainability Symp. (RAMS '04),* Jan. 2004.
[28] J. Gray and C. van Ingen, "Empirical Measurements of Disk Failure Rates and Error Rates," Microsoft Research Technical Report MSR-TR-2005-166, Dec. 2005.
[29] P. Corbett, B. English, A. Goel, T. Grcanac, S. Kleiman, J. Leong, and S. Sankar, "Row Diagonal Parity for Double Disk Failure Correction," *Proc. Third USENIX Conf. File and Storage Technologies (FAST),* 2004.

Jon G. Elerath received the BSME and MS Reliability degrees from the University of Arizona and the PhD degree in mechanical engineering from the University of Maryland. He has been a reliability engineer or reliability engineering manager at the General Electric Co., Tegal, Tandem Computers, Compaq, and IBM. He is currently with NetApp. He has applied reliability techniques to nuclear safety systems of fast breeder reactors, electronics and robotics of plasma-etching equipment, fault-tolerant computers, hard disk drive designs, and RAID data storage systems. He has authored more than 30 technical publications. Currently, his major area of interest is reliability of RAID storage systems. He is active in writing IEEE reliability standards and is a chairman of the Reliability Committee for the International Disk Drive Equipment and Materials Association (IDEMA). He is a member of the IEEE.

Michael Pecht received the MS degree in electrical engineering and the MS and PhD degrees in engineering mechanics from the University of Wisconsin, Madison. He is the founder of the Center for Advanced Life Cycle Engineering (CALCE), University of Maryland, College Park, where he is also a chair professor in mechanical engineering. He has written more than 20 books on electronic products development and on use and supply chain management and more than 400 technical articles. He has been leading a research team in the area of prognostics for the past 10 years and has formed a new Prognostics and Health Management Consortium at the University of Maryland. He has consulted for more than 50 major international electronics companies, providing expertise in strategic planning, design, test, prognostics, IP, and risk assessment of electronic products and systems. He served as a chief editor of the *IEEE Transactions on Reliability* for eight years and on the advisory board of *IEEE Spectrum*. He is a chief editor for *Microelectronics Reliability* and an associate editor for the *IEEE Transactions on Components and Packaging Technology*. He was awarded the highest reliability honor, the IEEE Reliability Society's Lifetime Achievement Award, in 2008. He has previously received the European Micro and Nano-Reliability Award for outstanding contributions to reliability research, a 3M Research Award for electronics packaging, and the IMAPS William D. Ashman Memorial Achievement Award for his contributions in electronics reliability analysis. He is a professional engineer. He is a fellow of the IEEE, the American Society of Mechanical Engineers (ASME), and the International Microelectronics and Packaging Society (IMAPS).