ABSTRACT

Title of Document: RELIABILITY MODEL AND ASSESSMENT OF REDUNDANT ARRAYS OF INEXPENSIVE DISKS (RAID) INCORPORATING LATENT DEFECTS AND NON-HOMOGENEOUS POISSON PROCESS EVENTS.

Jon Garrett Elerath
Doctor of Philosophy, 2007

Directed By: Professor Michael G. Pecht
Department of Mechanical Engineering

Today's most reliable data storage systems are made of redundant arrays of inexpensive disks (RAID). The quantification of RAID system reliability is often based on models that omit critical hard disk drive failure modes, assume all failure and restoration rates are constant (exponential distributions), and assume the RAID group times to failure follow a homogeneous Poisson process (HPP). This paper presents a comprehensive reliability model that accounts for numerous failure causes for today's hard disk drives, allows proper representation of repair and restoration, and does not rely on the assumption of a HPP for the RAID group. The model does not assume hard disk drives have constant transition rates, but allows each hard disk drive "slot" in the RAID group to have its own set of distributions, closed form or user defined. Hard disk drive (HDD) failure distributions derived from field usage are

presented, showing that failure distributions are commonly non-homogeneous, frequently having increasing hazard rates from time zero.

Hard disks drive failure modes and causes are presented and used to develop a model that reflects not only complete failure, but also degraded conditions due to undetected, but corrupted data (latent defects). The model can represent user defined distributions for completion of “background scrubbing" to correct (remove) corrupted data. Sequential Monte Carlo simulation is used to determine the number of double disk failures expected as a function of time. RAID group can be any size up to 25. The results are presented as mean cumulative failure distributions for the RAID group. Results estimate the number of double disk failures can be as much as 5000 times greater than that predicted over 10 years when using the mean time to data loss method or Markov models when the characteristic lives of the input distributions is the same. Model results are compared to actual field data for two HDD families and two different RAID group sizes and show good correlation. Results show the rate of occurrence of failure for the RAID group may be increasing, decreasing or constant depending on the parameters used for the four input distributions.

RELIABILITY MODEL AND ASSESSMENT OF REDUNDANT ARRAYS OF INEXPENSIVE DISKS (RAID) INCORPORATING LATENT DEFECTS AND NON-HOMOGENEOUS POISSON PROCESS EVENTS.

By

Jon Garrett Elerath

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor Michael Pecht, Chair
Professor David Barbe, Dean's Representative
Dr. Diganta Das
Associate Professor Patrick McCluskey
Associate Professor Peter Sandborn

Preface

“All models are wrong, some models are useful.”

- George Box

Dedication

This is dedicated to my father, B. Everett Elerath, and my mother, Betty Jean Elerath. My father, who passed away in December 2002, steadfastly believed he could not help me through every situation I would encounter in life. So, when I would ask how to do something, he often would help only by saying "figure it out". My mother passed away in April, 2004 had a tender heart and cared deeply for her children. My parents both passed while I pursued my Doctor of Philosophy. I miss you Mom and Dad.

.....and Dad, I figured it out.

Acknowledgements

I want to thank my advisor, Dr. Michael Pecht, for encouraging me to pursue my Doctoral degree. His critical eye for detail has only helped me grow professionally. As a distance-learning student residing in California for the duration of my studies, it was often difficult to navigate the maze of written University and Graduate College requirements as well as other unwritten expectations. I want to thank Elyse Beaulieu-Lucy for her help in navigating the objective rules and regulations. I want to thank Dr. Diganta Das for his incessant work, attention to detail, wisdom in vague situations, and advice on subjective requirements. Without his help I'd probably still be filing my study plan.

I want to acknowledge Sandeep Shah and Steven Magie, fellow Reliability Engineers, and Steve Kleiman, Executive Vice-President and Chief Technical Officer of Network Appliance, for their comments and suggestions. I thank Alan Wood, a good friend for many years, for giving up valuable family time to evaluate my first, very broken model. I thank the six managers I have had during this four and one-half year adventure, all of whom have supported me in my quest; Bob Weisickle, Ko Yamamoto, Bill Jacobsen, Dave Barney, Jim Ward and Dave Alexander. They also provided me with data for my analysis of time-between-managers.

My brother Doug earned a Ph.D. in mathematics from L.S.U. One day I asked him, "Doug, what do you do with a Ph.D. in mathematics?" He looked at me, and very seriously said, "Prove theorems." Being a practical sort of person I then asked "What do you do with these theorems?" He again looked at me, waved his hands and

said, "Jon, that is for you engineers to figure out." Here is to the practical side of mathematics. Thanks, Doug.

Lastly and most importantly, I want to thank my wife, Debra, for her tolerance and encouragement in what seemed like a journey without end. I love you.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

According to IDC, a leading research corporation, revenues for disk storage systems grew nearly 9.9% to $6.2 billion for the third quarter of 2006 as compared to the third quarter of 2005. The capacity of these systems was in excess of 783 petabytes, up 50% from the year ago quarter [1]. Much of this storage is deployed in redundant arrays of inexpensive disks, or RAID systems. Storage content of these systems covers everything from "mission critical" on-line transactions and business intelligence, to e-mail, data warehouses, and test and development [2]. Compliance with the Sarbanes-Oxley Act[1] and the desire for disk-to-disk backup have generated a huge potential for the storage industry growth [3]. In all these areas of growth, high reliability is a critical feature as evidenced by the use of RAID.

The most common models for estimating RAID reliability make erroneous assumptions regarding the system rate of occurrence of failures; hard disk drive failure causes, distributions and consequences; realistic restoration distributions; and the tremendous (negative) impact of latent media defects and the beneficial effectiveness of data scrubbing. Although recognized by Kari in 1997 [4], only one recent research effort by Schwarz addresses the concept [5] of latent defects from a system reliability perspective. However, Schwarz continues to use constant failure and restoration rates, treats latent defects as time independent probabilities in the final Markov model, assess mean time between failures (instead of frequencies of failure), develops a simplistic model that does not account for order dependence, and applies

[1] Sarbanes-Oxley Act is a Federal Law to assure better accounting and reporting practices in US companies

his model towards simple mirrored HDDs. As this thesis unfolds, the bases for these criticisms and their solutions will become evident.

1.1 Thesis Problem Statement

Current reliability models of RAID storage systems incorrectly estimate the number of double disk failures per *(N+1)* RAID group due to the following:

a) HDD failures often do not follow a homogeneous Poisson process. Failure rates are often increasing or decreasing, not constant

b) HDD failures do not come from a single population distribution. That is, significant time-to-failure variation exists across populations of HDDs from:

- different manufacturers
- different "families" from the same manufacturer
- different vintages in a single "family" from a single manufacturer

c) RAID group (system) failures do not follow a homogeneous Poisson process, so estimates of the number of DDFs are incorrectly calculated when using renewal theory and assuming a HPP

d) Latent defects in HDD media are not included

e) The RAID system logic modeled does not properly account for conditional order of latent defects and operational failures

This research develops a model that addresses these issues and produces results consistent with the over-arching thesis, with general industry results and with specific Network Appliance field data for double disk failures in RAID systems

1.2 Thesis Organization

The model necessary to support my thesis required research in the areas of HDD designs and failures, RAID architecture, HDD failure distributions and an understanding of renewal theory. The research results are presented in Chapter 2 through Chapter 4. Chapter 2 presents an overview of HDD design. It includes a brief discussion of the basic components in a hard disk drive, shows how data is laid out on the discs and presents some of the high probability failure modes and causes. These are not new, but are the bases for distinctions in consequences of failures discussed later. That is, some failures result in localized data corruption and some result in complete HDD failure.

Chapter 3 describes the RAID system architecture including a high level discussion of the processes for recovering from errors and HDD failures. The concepts presented in Sections 2 and 3 are used extensively in later development of the model.

The Previous Work, Chapter 4, presents a very significant statistical flaw ignored by those assessing RAID reliability. Although known by statisticians and many in the field of reliability, the rate of occurrence of failure of the system has little connection with the failure rates of the components that make up the system. Even if all the components fail according to a Homogeneous Poisson Process (HPP), the system may not. All RAID assessments to date assume the system follows a HPP. This research does not assume any distribution at the system level and shows the consequences of assuming system failures follow a HPP.

Most researchers have little or no true field reliability data for HDDs upon which they base their model assumptions. They cannot see the wide variability that exists across manufacturers, within manufacturers' models, and even within different vintages of product form the same manufacturer. Chapter 5 presents a significant contribution to the analyses of RAID systems. In this Chapter, times to failure distributions are presented based on field data, showing that the constant failure rate assumption used by other researchers is generally inaccurate. Vintage analyses are presented as well. Knowing the true distributions provides a significant basis for the model developed as part of this thesis.

Chapter 6 presents the logic of the new model, another significant contribution from this thesis. In Chapter 7 the Sequential Monte Carlo simulation process is presented with application to the general model concepts in Chapter 6. The logic of the model and the implementation of the logic in code are validated through several studies and comparisons to closed form, manual calculations in Chapter 8.

In Chapter 9 transition distributions are developed, simulation results are presented and the significance of those results is discussed. Chapter 10 provides the conclusions for the research, analysis and justification of the thesis statement. It also summarizes the contributions of this research and identifies future work.

Chapter 2 Hard Disk Drives, Failure Modes and Causes

The hard disk drive is the building block for a RAID storage system. While RAID systems include redundancy, multiple failures can lead to data loss. This section provides the background necessary to understand how the failure causes result in the failure modes and how those modes affect system operation. HDD general design is presented briefly. Because of its importance to the read process, the read element is discussed in more detail. The layout of the data on the discs is shown to clarify some of the failure modes described in section 2.3, the failure modes and causes common to current HDDs. Knowledge of the failure causes is paramount to understanding the basis for the new model presented later. In most previous studies, ref. [6] for example, it was asserted that the only reason data was written corruptly was due to the bit-error rate inherent to the design's recording channel electronics. Once written correctly, it was asserted that a significant cause for corruption after being written correctly was due to "bit-rot" in the media [7]. Other recent research confirms the failure modes and causes [5] that I present. The discussion here augments those studies and identifies additional failure causes that explain and justify the design of the model presented later.

2.1 HDD Description

In 1988, the total capacity of a single 5 ¼" hard disk drive was around 40 MB (megabytes). Today, 3.5" HDDs can be 500 GB (gigabytes) and 750 GB capacity HDDs should be available around April, 2007. In 1988 the HDD discs rotated at 3600-5400 rpm, had areal densities of 25 MB/in.^{2} and had a fly-height (distance

between the head and the disk) of 8-10 μ-in. In 2006, speeds of 10,000 to 15000 rpm are commonplace, areal densities are up to 125,000 MB/in.^{2} [8], and head fly-heights are less than 0.5 μ-in.

All the mechanical components and some of the sensitive electronics are contained within the hard disk drive enclosure. The majority of the electronics are attached to the printed wiring assembly (PWA) affixed to the outside of the drive. The components within the enclosure are assembled in a clean room environment to control contamination, humidity and electrostatic discharges (ESD). The HDD enclosure is sealed in the clean room and a high purity filter allows air pressure inside the enclosure to equilibrate with the local atmospheric pressure. Figure 1 shows a HDD with the cover removed. (The printed wiring assemble is on the opposite side and cannot be seen in this photo.) The main components and assemblies within the HDD enclosure are labeled, including a voice coil motor assembly, actuator arm, read/write heads and the discs (and media). Some designs have a pre-amp attached to a flexible cable inside the enclosure. A more detailed description can be found in reference [9].

Sliders contain the read and write elements, as well as thermal and mechanical shields to protect the sensitive reader from particles. The slider is about the size of a piece of crushed black pepper. Hitachi's "Femto slider", introduced in 2003, is approximately 0.8 x 0.75 x 0.2 mm [10]. Sliders are aerodynamically designed to fly less than 10 nanometers (less than 0.4μ-in.) above the surface of the disc [11]. The sliders are attached to the "suspension" via a gimbal, and the suspension is affixed to the actuator arm. Several actuator arms are stacked on top of each other to form a

head-stack assembly. Usually, there are two heads per disk, although some designs may have one head on a single disk for low capacity. Figure 1 shows a stack of 3 discs and 5 actuator arms. This particular "E-block", to which the suspensions are attached, is designed to accommodate a maximum of 4 discs (8 heads), but the top two suspensions were not installed. This de-populated version has 6 heads and ¾ the capacity of the fully populated version, but uses the same machined aluminum "E-block" to minimize unique parts. The actuator arms are moved across the disks using a voice coil motor assembly.

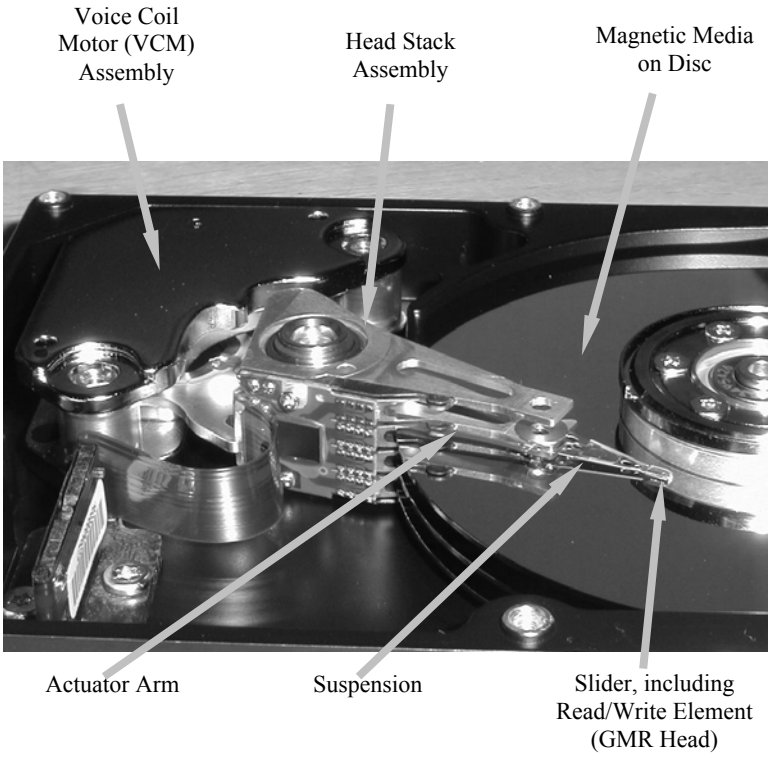


Figure 1 - Hard Disk Drive Components

The giant magneto-resistive (GMR) head that contains the read and write elements has as many as 20 material layers. Additive (electroplating or liftoff

masking) or subtractive (ion milling, wet etch, reactive ion etch, chemical mechanical processing) photolithography is used to create the necessary patterns on the head [12]. The read element alone is composed of multiple layers of materials [13], [14] and [15]. Figure 2 shows the layers common to a giant magneto-resistive head, and shows the function of each of the layers or groups of layers. There are numerous other combinations of materials that can be used to achieve the functions listed. Other common anti-ferromagnetic[2] materials include NiMn, FeMn and IrMn. Each of these has its own set of advantages.

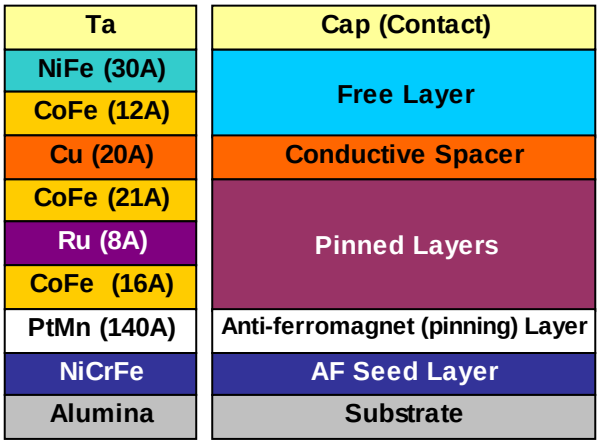


Figure 2 - Reader Element Layers and Thicknesses (Å)

The basic operation of a GMR head can be found in numerous references, one of which is [16]. There are four functional layers in a GMR element that form the foundation of its operation. There is a thick, anti-ferromagnetic layer that serves to fix or "pin" the adjacent magnetic layer. The adjacent layer is magnetic, but its magnetic orientation is "pinned" or fixed so it cannot change. The third layer is a nonmagnetic

[2] The anti-ferromagnetic layer pins the magnetization of an adjacent ferromagnetic layer so that it doesn't reverse in an external magnetic field.

conductive layer (spacer) followed by a free layer whose magnetic orientation changes depending on magnetic fields near it (on the disc itself). The pinned, conductive spacer and the free layers are all relatively thin (5-50Å) as compared to the anti-ferromagnetic layer (140Å). The conductive layer allows electrons to move freely from the pinned layer through the conductor to the free layer. When a magnetic field passes near the free layer, the magnetic orientation in the free layer can change. When the orientations of the free layer and pinned layer line up (parallel), there is less electrical resistance. When the magnetic orientations are not parallel (anti-parallel), higher resistance exists. By passing a small current through the head, the change in resistance produces a change in voltage. After signal conditioning, these binary voltage spikes are interpreted as "0s" or "1s".

2.2 Data Layout

During the manufacturing process, each disc is divided into tracks and sectors as illustrated in Figure 3. There are on the order of 100,000 to 500,000 tracks per radial inch (TPI) that form concentric circles. There are on the order of 100-150 "servo-wedges" the radiate out from the center of the discs creating sectors [17]. The track sectors between servo wedges are available for user data. User data is written in terms of fixed block size, typically 512, 520 or, more recently, 4k Bytes. The arc length at the outer diameter is greater than at the inner diameters, so the physical track length between servo wedges is greater as well. However, since the number of blocks between servo wedges is fixed, the physical length of each block and each sector increases along the radius. More sophisticated layouts recover the added physical space in the outer diameters by (proprietary) changes in block length or block

quantity. The lower bits per inch in the outer tracks helps compensate for the increased angular velocity in the outer tracks.

Longer blocks allow for more error correcting capabilities and greater reliability. The difference between 512 and 520 blocks is 8 added bytes necessary to enhance error correcting capabilities. The recent push for 4k blocks will allow interleaving and even greater error recovery algorithms within the HDD even without RAID.

The servo information is written on a track by track basis, but together the servo data across the tracks at any given angle form a servo wedge. Servo data are written at the factory and contain parametric information to keep the head properly positioned so it seeks, reads and writes to the correct track. While user data, shown as gray, can be re-written as needed, the servo data, shown as brown, cannot. Once servo data is destroyed or corrupted, the HDD can no longer stay on track so it cannot be read or write correctly.

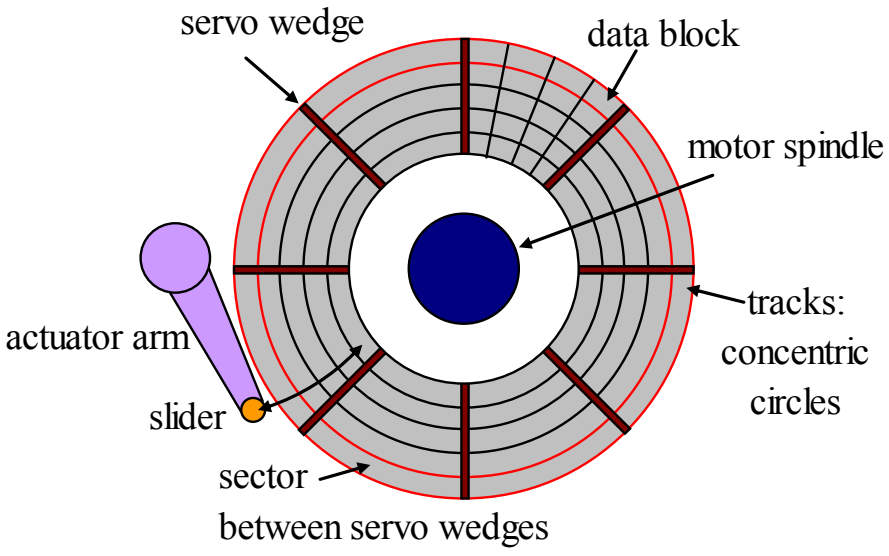


Figure 3 - Track and Data Layout on Discs

2.3 HDD Failure Causes

This section presents the dominant causes of failure for today's HDDs. The consequences or modes resulting from these failure causes are presented in the next section. Figure 4 shows the relationship between the discs, the heads and suspension for reference.

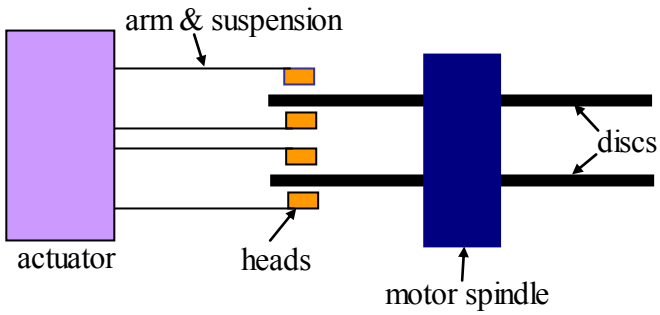


Figure 4 - Head & Disc Positions

2.3.1 Scratches from Hard Particle Contaminants:

In 2006, the read/write heads are flying in the range of 0.3µ-in. to 0.7µ-in. above the media. The heads are so close to the media that particles introduced as a result of the manufacturing process can become wedged between the head and media any time the disk is rotating. Typically, the loose particles are "hard", being composed of stainless steel, SiO_2, carbon, silicon or TiW. Although all disk drive components are rigorously cleaned prior to assembly, contaminants that are only 0.3µ-in. are extremely difficult to remove from every crease, crack or joint. Often these particles will dislodge themselves as the drive is used. The spinning discs create significant air turbulence within the drive enclosure and the motion of the actuator itself (to which the read/write heads are attached) can vibrate enough to shake particle off of the

heads. Although the slider is designed to push particles away with a “plow like” leading edge, particles can become caught between the reader and the media, scratching the media and destroying existing data.

2.3.2 Media damage from “Soft Particles”

While some HDD designs use a ramp to unload the heads from the disc surface when powered off, others allow the read/write heads to rest on the disk in a specific landing zone. If the HDD receives an impact while the heads are in contact with the disc surface, the heads can be damaged, but also, soft particle contamination can result. Particles made up from the coatings on the disc surface are then able to attach to the heads and either interfere with proper reading and writing, or create contact between the head and the media, creating a large smear and corrupting the data. Smears can occur any time the disc-pack is spinning, not just during writing or reading.

The introduction of perpendicular magnetic recording (PMR), instead of longitudinal recording (LMR), requires different material layers in the media. One of these is the (relatively) soft magnetic under-layer [18]. This technology is even more susceptible to scratches and smears from the softer particle contaminants in the HDD. In some instances small particles of the mold-flash from a plastic encapsulated IC within the HDD enclosure caused contamination.

2.3.3 Thermal Effects on Read Heads

Disc surfaces are coated and polished to control flatness measured by optical interferometers. In spite of these efforts, the polishing and burnishing process may

still leave areas that are raised as compared to the rest of the media surface after coating. As the heads pass over these raised areas, one of several things may occur. The head can hit the media, creating soft-particle contamination. As the head touches the media high localized heating occurs, called a thermal asperity (T/A). The localized high heat from the T/A can erase data that is already written. There is a time and usage dependence for this frequently occurring event. The effects of the head hitting the media can be immediate or cause gradual degradation in the magnetic properties of the head if the "bump" is repeatedly impacted. It makes no difference whether the heads are reading, writing or simply passing over the spot on the way to another track on the disc, the impact causes degradation. If data has already been written to that spot then it is lost and a latent defect (failure) exists. If the next operation at that spot is a write without a following verification, data will be written and, in effect, be corrupted. Only in the case of write verification will the defect be identified and the data reallocated to a new sector on the drive.

The amount of twist permitted in the suspension will also affect the fly-height of the heads and therefore affect the frequency of T/As. A suspension that is not robust against torsion will twist each time the actuator changes direction. Frequent changes in direction create more twist and more head/disc contacts. This phenomenon is highly design dependent and there can be significant arm design differences even within the same drive manufacturer.

Temperature has the greatest effect on the reliability of the read elements. Prakash [19] points out that the anti-ferromagnetic layer, PtMn in this case, loses its "pinning ability" as temperature increases. At the temperature called the blocking

temperature, T_b, "exchange anisotropy is lost completely, although exchange coupling at the interface is still present [19]." Since the general operating temperature of heads is approximately 120 ± 30 °C, and the T_b for PtMn is around 325 °C, any thermal source that raises the temperature can cause head failure. Thermal asperities, in which the heads hit the surface of the disk while spinning, are one common source of short, but very high temperatures.

While heat does affect the mechanical and electrical properties of the materials, it has en even more pronounced effect on the magnetic properties. Tsu [20] studied the effects of Joule temperature rise, high current densities and magnetic field resulting from applied bias. He conducted a constant-current life test with current density of 1.8 to 11.5×10^7 A/cm^2 in a convection oven at 25 to 80 °C, for 1000 hours. Furthermore, he ran tests at positive and negative bias to see the effects of current density. Tsu determined the activation energy for failure by amplitude loss was 1.05 +/- 0.16 eV and that of resistance loss was 1.27 +/- 0.2eV. The results are seemingly unaffected by the bias direction. Reliability is less affected by current density for practical reliability than it is by overall sensor temperature. He therefore concludes that electro-migration and diffusion are secondary failure mechanisms, and failure of the pinning layer from heat is the dominant mechanism.

2.3.4 Degraded/Unstable Heads

Read/write heads are as intricate in design as the most sophisticated integrated circuit today. The heads rely on magnetic fields from materials that are only a few atomic layers thick. Due to their design, they are subject to "inherent instability". That is, under normal usage, design voltages and design currents, the read heads can

become unstable. When this happens, they are no longer able to read data. This common failure mechanism creates an operational failure in which the read head cannot read.

Electro-static discharge (ESD) has become one of the most important issues in manufacturing GMR heads [21]. Wallash claims GMR heads are "the world's most static sensitive device in mass manufacturing today" [22], and can become "wounded" during the head manufacturing or HDD assembly process if ESD is not carefully controlled. Many ESD induced failures are immediate and readily apparent. These are usually the ones in which the temperatures from the short duration, high voltage ESD event causes localized heating, melts one or more layers and changes the resistance. As little as 200 mA for 1.5 ns will melt a typical MR sensor [23]. When a metal contact is made to a charged MR head, the stored energy is transferred in about 1 ns [23]. This current can result in temperatures greater than 1400 °C. If not completely failed, the high voltage may cause pitting and resistance changes.

However, the more insidious ESD problem is that of magnetic damage resulting from a combination of Joule heating and internal magnetic field switching during the ESD event [21]. When the sensor's temperature exceeds the critical blocking temperature of the pinning layer (PtMn), the strength of the pinning layer is reduced so the pinned layer will become unstable at lower temperatures. The degree of instability or relationship between temperature and time to failure for a damaged head depend on the energy from the ESD event, the GMR head design and the material layers. The time required to achieve the peak temperature is on the order of 20-30 ns. This is much smaller than the duration of the ESD pulse which is around 150 ns.

Results of a study by Yang [21] indicate the voltage required to induce magnetic damage is about half of that to cause physical breakdown (pitting or melting). This is consistent with results reported by Wallash [23], whose results are shown in Figure 5.

In GMR heads, a single ESD transient (1 ns) with peak current of only 25mA (1nJ) can cause severe magnetic changes. ESD Studies have shown that a GMR head can be partially damaged by ESD, pass the component (quasi-static test) and assembly performance tests and yet rapidly degrade upon use [23], [24]. The mechanism by which the magnetic properties change is explained by the blocking temperature. Below the device specific blocking temperature, T_b, the anti-ferromagnetic layer maintains it anisotropic alignment. Above the blocking temperature, some grains are no longer aligned and the strength of the field is reduced, allowing the pinned layer to be affected (unstable) when magnetic fields are present.

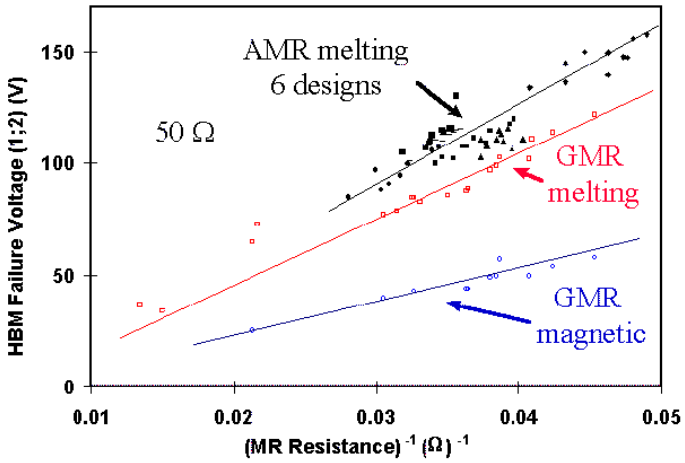


Figure 5 - Voltages Required for Melting and Magnetic Damage [23]

Another aspect of heads damaged by ESD is that other events that create high temperatures, such as thermal asperities, are more likely to cause complete failure of the damaged heads by bringing the sensor temperature closer to the blocking temperature. For the PtMn GMR sensor, magnetic failure occurs at $34V_{HBM}$ and the resistive failure point is 43 V_{HBM}. At 150 °C, the magnetic failure occurs at $24V_{HBM}$ and the resistive failure point is 32 V_{HBM} [25].

2.3.5 Contamination from Lubrication

The motors for new HDD designs are often fluid-dynamic bearings. The oil in the bearings is contained using barrier films, close tolerances and the inherent oil viscosity. Since the bearings are inside the HDD enclosure, any leakage or residual oil can become volatile and condense on surfaces such as the discs. The heads pick up the oil and are no longer able to read or write. This is a time dependent mechanism that is affected by spinning and usage. Properly written data may be unreadable if this mechanism occurs. The actuator bearing also has lubricant in it. Any leakage or residual will have the same effect as the motor bearing oils. Residual oils from the base casting machining processes also can get on the heads and interfere with reading.

During manufacturing, the discs are coated with a special lubricant to mitigate damage if the head should contact the media surface. However, depending on the head design, lubrication composition and actuator activity, the heads can pick up lubricant from the disc surface. When a sufficient amount of lube is built up, the head will take on different aerodynamic characteristics, often flying much higher than it is designed to fly. The next time it writes, the head may be too far from the surface to

induce adequate magnetic field to set the bits correctly. These are often referred to as a “high fly writes” which are magnetically too weak to be read.

Fly-heights need to be low to increase reliability of the data read/write process and to increase the areal density of the discs. It is not likely that the fly-heights will ever increase; only that they will decrease and heads will continue to get closer to the discs. Contaminant reduction efforts will continue at the drive manufacturer, but this dominant failure mechanism is not likely to be fully controlled even in the next generation of disk drive.

2.3.6 Other Causes

Motors: Most motor bearings for HDDs spinning at 10k - 15k rpm use fluid dynamic bearings (FDBs) to reduce non-repeatable run-out and vibration. In a FDB the motor spindle is separated from the hub by a small amount of fluid. When stopped, the spindle touches the hub. The fluid is “held in” only by the mechanical properties of the fluid and the bearing design. It is possible that the bearing can be under filled from the outset in the factory or that the fluid leaks during use. Although less probable than the mechanisms listed above, fluid loss does occur. Bearing failure can occur anytime the discs are spinning and does not depend on the amount of data read or written.

PWBAs: The printed wiring board assembly can also fail. Most frequent causes include DRAM failures and mechanical damage. Resistors and capacitors have been known to be knocked off the board. These cannot always be found during the test process but may show up some later time during use. Both of these failures can render the entire drive as intermittently failing or completely failed.

2.4 HDD Failure Modes (Consequences)

The consequences of the HDD failures are a critical aspect of the new model developed since each mode has a specific result. Based on the consequences and the system operation (discussed later), there are two basically different failure modes: cannot find the data and data missing. The read error failure modes discussed below are shown graphically as a fault tree in Figure 6.

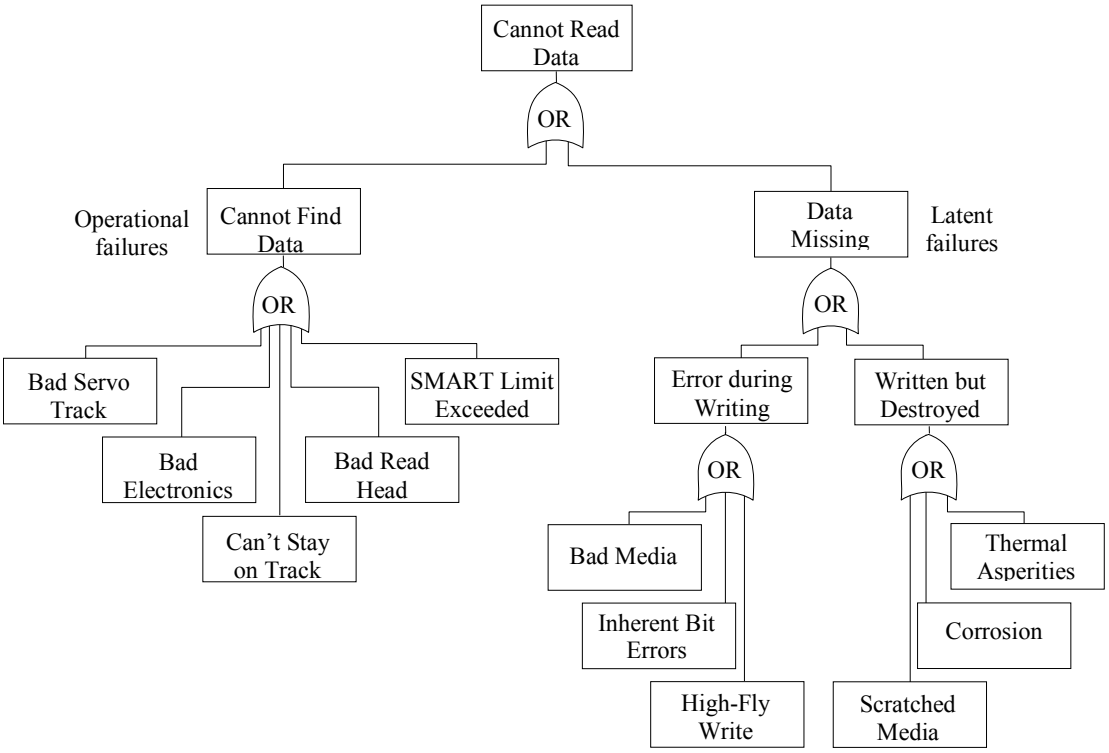


Figure 6 - Fault Tree for Read Errors

2.4.1 Cannot Find Data

Servo data is written periodically on every data track of each disc surface in what are called servo "wedges", as shown in Figure 3b. The servo wedges contain data

used solely to control the positioning of the read/write heads and are required for the heads to stay on track, whether executing a read, write or seek command. Servo-track information is written during the manufacturing process and cannot be reconstructed using RAID or rewritten in the fields. Media defects in the servo-wedges can cause the HDD to lose track of the heads' locations or where to move the head for the next read or write. These faulty servo tracks will result in the inability to access data, even though the data is written and uncorrupted. Servo-wedges can be damaged by scratches or thermal asperities.

Tracks on a HDD are never perfectly circular. The present head position is continuously measured and compared to where it should be and a position error signal (PES) is used to properly reposition the head over the track. This repeatable run-out (RRO) is all part of normal HDD head positioning control. Non-repeatable run-out (NRRO) cannot be corrected by the HDD firmware since it is non-repeatable. NRRO, caused by mechanical tolerances from the motor bearings, 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. This mode can be induced by excessive wear and is exacerbated by high rotational speeds. It affects both ball and fluid-dynamic bearings, but to different amounts.

All recent HDDs collect and analyze functional and performance data to try and predict impending failure using self-monitoring analysis reporting technology (SMART). For example, the number of sector reallocations is monitored. If an excessive number occur in a specific time interval (values are proprietary), the HDD is deemed unreliable and is failed out. In general, reallocations are expected and

many spare sectors are available on each HDD as long as the SMART threshold is not exceeded.

Most head failures are due to changes in the magnetic properties, not electrical. Electro-static discharge (ESD), high temperatures and physical impact all affect magnetic properties. As with any highly integrated circuit, ESD can leave the read heads in a degraded mode, undetectable by testing. Subsequent moderate to low levels of heat will then be sufficient to fail the read heads (magnetically). The read element is physically hidden and difficult to damage, but heat can be conducted from the shields to the read element, affecting magnetic properties of the reader element, especially if already weakened by ESD.

The electronics on a HDD are complex. Failed DRAM and cracked chip-capacitors have been known to cause HDD failure.

2.4.2 Data Missing

Data can be missing either because it was not written well initially or because it was erased or corrupted after being written well. All errors resulting from "data missing" are latent because the corrupted data is resident without the knowledge of the user or the HDD software knowing

a) 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 is rarely checked immediately after writing, BER can also produce

corrupted data during writes. While BER does account for some fraction of defective data written to the HDD for both read and write combined, a greater source of write-errors is magnetic recording media coating the discs. If the media is already scratched, contains voids or bumps, or has a hydrocarbon contaminant (machine oil) on its surface, write errors will result. The magnetic media in a HDD consists of various metallic layers with controlled grain sizes. These are applied through sputtering. The surfaces are polished prior to the final application of sputtered carbon overcoat. The sputtering process leaves bumps on the disc surfaces which are removed with the polishing, but scratches can occur during polishing.

Writing on scratched, smears or pitted media can also result in corrupted data. 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 or during field use.

A common cause for poorly written data is "high-fly writes". Magnetic field strength decreases rapidly as a function of distance between the head and the magnetic media. The heads are aerodynamically designed to have a negative pressure and maintain the small, fixed distance above the disc surface at all times. However, if the aerodynamics are perturbed the head can fly too high resulting in weakly (magnetically) written data that cannot be read. All discs have a very thin film of lubricant on them as protection from head-disc contact, but lubrication build-up on the head is one way to affect the flying height.

b) Data Written but Destroyed

All previous RAID system reliability models presume that once written, the data will remain undestroyed except by degradation of the magnetic properties of the media ("bit-rot"). Media can degrade, but is inconsequential and failure for other reasons is much more probable. Data can become corrupted any time the discs are spinning, even when data is not being written to or read from the disc. Three common causes for erasure include thermal asperities, corrosion and scratches/smears.

Thermal asperities (T/As) are instances of high heat for a short duration caused by head-disc contact. This is usually the result of heads hitting small "bumps" created by particles embedded in the media surface during the manufacturing process, even after burnishing and polishing the surfaces. 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 so they are not trapped between the head and disc surface, but particles do get caught there. Hard particles used in the manufacture of a HDD, such as Al_2O_3, TiW and C will cause surface scratches and data erasure. 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, also can cause data erasure and may be accelerated by T/A generated heat.

Chapter 3 RAID System Architecture

RAID is more than a simple *N+1* or *N+2* hardware redundancy. RAID allows data to become corrupted from partially failed HDDs as described by the failure modes in Section 2.4. This Chapter presents the general concept of RAID, the Berkeley taxonomy of RAID architectures, and a summary of how the HDD firmware and RAID operating system can correct corrupted data within an HDD and across HDDs in the RAID group.

3.1 General RAID Concepts

Redundant Arrays of Inexpensive Disks (RAID) are created to enhance data input/output (I/O) performance and reliability by grouping together multiple inexpensive disks rather than using one large disk. Multiple disks change the I/O from a serial data transfer process to a parallel transfer process. The I/O transfer rate is enhanced for small RAID group sizes, but can decrease as the group grows larger because of the increased time required to manage data locations and data flow. Systems are often composed of more than one RAID group and it is not uncommon to have one computer manage 2 to 20 RAID groups.

Usually, RAID employs parity checking for data distributed across multiple disks and uses it to reconstruct data that has been corrupted or "lost" due to damaged or "defect-laden" disc media. Using parity across the data disks creates HDD redundancy because any amount of corrupted data on a single HDD can be reconstructed as long as all the data on all the other HDDs is uncorrupted. When a specific location on a HDD cannot be read, the corrupted data is reconstructed using

parity and stored at a new physical location on the HDD. If the data had been on a single HDD, corrupted data could not have been be recovered. Most frequently, small amounts of data are reconstructed and reallocated to new physical locations.

3.2 RAID Definitions

There are generally considered 7 levels of RAID, as summarized in Table 1, each having different properties in terms of latency, performance, reliability and a number of other attributes. Combinations of these levels result in RAID-10 and RAID-50. Paterson et al. [26] and Shooman [27] provide detailed discussions of RAID groupings and data storage patterns. RAID-0 is without redundancy, but may have data striped across multiple HDDs for performance reasons, and RAID-6 uses parity across multiple groups of HDDs to allow data recovery if two HDDs fail simultaneously [28]. RAID 6 effectively becomes an "N+2" logic.

The most common implementation of RAID uses an "N+1" architecture in which N logical HDDs store data and 1 logical HDD holds parity. Note that parity bits may be restricted to a physically separate HDD, as in RAID-4, but are more likely intermingled with data. However, since the reliability logic is the same, it is easier to think of having N physical data HDDs and 1 physical parity HDD. Theoretically, there can be any number of data disks, but for performance groups of 4-16 are commonly used. The most reliable configuration achieves the redundancy with the fewest number of disks, meaning $N + 1 = 2$. This is also the most expensive in that 2 disks are used to store the data of 1 disk, a 100% overhead in capacity. Increasing the number of disks beyond 2 reduces both cost per gigabyte and reliability.

Table 1 - RAID Designations

Name	Attributes
RAID-0 Also known as just-a-bunch-of-disks, or JBOD	No redundancy. Data may be striped across more than one disk for performance reasons. Reliability is lower than a single disk.
RAID-1 Mirrored disks	Two physical disks that store identical copies of the data. Highly reliable, high performance, high cost.
RAID-2 Hamming Error Code Correcting (ECC) with bit-level interleaving	Single error correction capability with double error detection. ECC is striped across disks. Not used often because of reduced performance in determining ECC as compared to parity.
RAID-3 Block based parity-bit codes	Data-block based parity-bits are stored on a separate disk.
RAID-4 Sector based parity-bit codes	Sector based parity-bits stored on a separate disk. Data is striped across the other disks in the group.
RAID-5 Sector based parity-bit codes	Sector based parity-bits are striped across multiple disks along with the data.
RAID-6 Dual parity	Sector based parity-bits are derived from two different RAID groups. Slows performance but greatly improves reliability.

Note that, statistically, there is also a limit for k-out-of-n redundancy called the "cross-over" point. Beyond this level of redundancy, the reliability of a single disk is greater than the reliability of the k-out-of-m system. For an *N+1* system (k = N; m = N + 1) with reliability, *R*, of a single disk drive, that point is defined as:

$$R = \sum_{x=k}^{m} \binom{m}{x} R^x (1-R)^{(m-x)} = NR^N(1-R) + R^{N+1} \qquad \textbf{eq. 1}$$

Since the reliability changes in time, this may be difficult to evaluate and render a crossover point that changes in time.

3.3 Data Loss and Error Correction

Error correcting codes (ECC) *on the disk* and parity *across the disks* is a common method to assure accurate data recording. ECC, often based on Reed-Solomon codes [27], uses Boolean operations to encode blocks of data, saving the resultant as well as the data. If bits are lost, the ECC is decoded to determine the value of the lost bit and data integrity is preserved. The strength of ECC is enhanced by interleaving multiple blocks of data so that if a large physical area of a disk (many bits) is not readable, the corrupted area does not affect all the data in a single block, but spreads the errors out over multiple blocks. The ECC must then correct multiple blocks, which is easier than recovering the same number of lost bits from a single block.

Losing access to an HDD occurs when the supporting hardware or software prevents data from being written to or read from the disk. Examples include loss of electric power, failure of the processor or failure of the host-bus adapter. Once the support function is fully restored, the HDDs are accessible and, assuming graceful shutdown, the data on the disks remains uncorrupted and readable. Data loss occurs when the stored data has been corrupted to the point that on-board ECC and parity across disks cannot reconstruct the missing bits or when a catastrophic disk drive failure occurs. When one disk is in the reconstruct mode, a read error on any other HDD will result in lost data. Certain errors do not cause failure when all *N+1* disks in the RAID group are fully functioning, but will cause failure if one disk has failed.

A functional representation of RAID-4 with 3 data disks and one parity disk is shown in Figure 7. Parity of data stream "A" is computed and stored on disk 4 while the data is striped across disks 1 to 3. Figure 8 depicts the process of reconstructing

data onto disk 3. The data striped across disks 1 and 2 and the parity for "A" on disk 4 are used to reconstruct the data that was stored on disk 3. Many RAID systems have "hot plug" capability on the disks so that in the event of a complete disk failure, the failed disk can be removed, a new one installed and the data restored while continuing to process other data streams. Thus, very high data availability is achieved. On-line spares reduce the replacement time further.

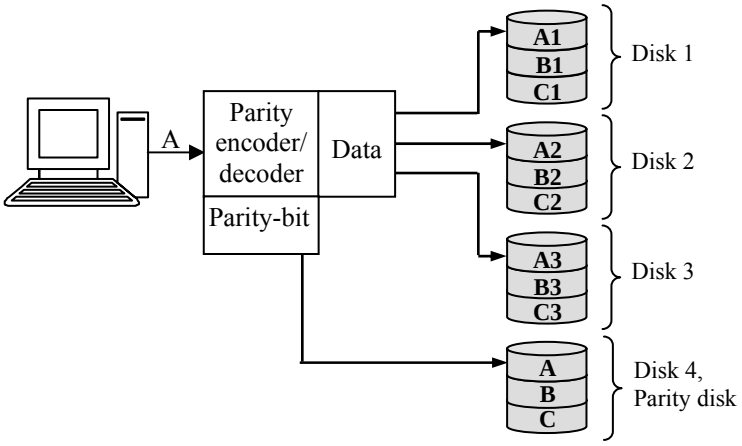


Figure 7 - Write Process for RAID-4

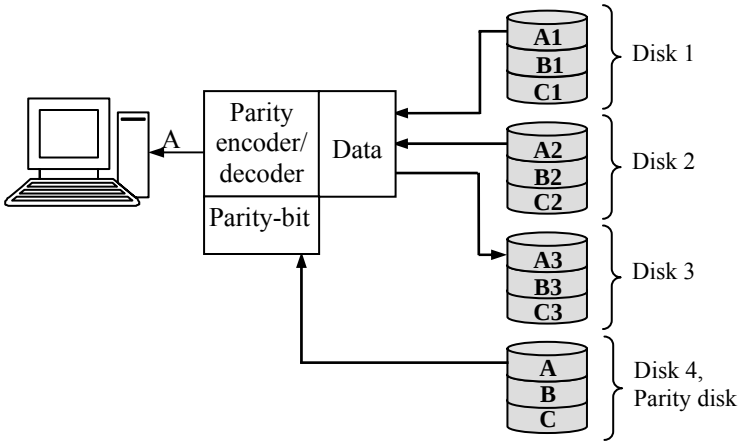


Figure 8 - Read and Correct Process for RAID-4

Chapter 4 Previous Work

Currently, RAID reliability modeling is dominated by two methods. The mean time to data loss (MTTDL) is a closed form equation that attempts to estimate the mean time between double disk failures (DDFs). The Markov model is more complex, usually requires a computer to solve and produces a probability of double disk failure in a specified time. Both of these approaches and previously proposed enhancements are in Section 4.1. Discussion of the assumptions and errors inherent with these two methods are presented in Section 4.2. The new method proposed in this thesis appears in Chapter 6.

4.1 MTTDL and Markov Models

Most researchers ([3], [5], and [29] through [46]) use the same concepts and assumptions in assessing the reliability of RAID systems and develop a MTTDL. MTTDL attempts to estimate the average time between simultaneous failures of two hard disk drives in an *N+1* RAID group, resulting in the inability to deliver data upon request. This method is predicated on the assumption that not only do all the HDDs in the RAID group follow a homogeneous Poisson process, but that the RAID system (composed of the *N+1* HDDs) also follows a homogeneous Poisson process. Therefore, all HDDs in the RAID group are assumed to have the same constant failure rates, λ, and constant repair (restoration) rates, μ. The failure rate is assumed to be the reciprocal of the mean time to failure as follows:

$$\lambda_{disk} = \frac{1}{MTTF_{disk}} \qquad \textbf{eq. 2}$$

The restoration rate is assumed to be the reciprocal of the mean time to restore.

$$\mu_{disk} = \frac{1}{MTTR_{disk}} \quad \textbf{eq. 3}$$

For these assumptions, a RAID group composed of *N + 1* disks has a MTTDL calculated as follows:

$$MTTDL = \frac{(2N+1)\lambda + \mu}{N(N+1)\lambda^2} \quad \textbf{eq. 4}$$

Since the repair rate is usually much larger than the failure rate, the term $(2N+1)\lambda$ in eq. 4 can be ignored and eq. 2 and eq. 3 can be substituted into eq. 4. The resultant expression for MTTDL is shown in eq. 5.

$$MTTDL_{Indep} = \frac{\mu}{N(N+1)\lambda^2} = \frac{MTTF_{disk}^2}{N(N+1)MTTR_{disk}} \quad \textbf{eq. 5}$$

At the system level, the assumption of a homogeneous Poisson process means the "system failure rate", $\lambda_{syst}(t)$, is constant and can be estimated by the reciprocal of the MTTDL. Then, the number of DDFs can be calculated by the product of $\lambda_{syst}(t)$ and the time at risk just as in a HPP, as shown in eq. 6.

$$E[N(t)] = \lambda_{syst}(t)t = \lambda_{syst}t \quad \textbf{eq. 6}$$

Based on eq. 6, the estimated number of failures for an MTTDL of 36,162 years (MTBF = 461,386 hrs; MTTR=12 hrs; N=7), 1000 RAID groups and 10 years of operation is shown in eq. 7.

$$N(t) = \frac{\frac{10\ \text{yrs}}{\text{RAID Group}} \text{x}\ 1000\ \text{RAID Groups}}{\frac{36{,}162\ \text{yrs}}{\text{Failure}}} = 0.28 \qquad \textbf{eq. 7}$$

Markov models are an alternative approach to estimating RAID reliability. Generally, the major drawback of Markov models is their high level of complexity. A basic Markov model for a k-out-of-n system is not difficult to create or evaluate if all the HDDs failure rates are constant and the same value, all repair rates are constant and the same value, and the repair strategy is simple. In Chapter 5, however, it will be shown that HDDs rarely have constant failure rates. The model logic itself, discussed in Chapter 6, will illustrate why the repair and restoration rates are not constant. Inclusion of delay times, the location parameter in a Weibull distribution, adds significant model complexity. In such a delay time, the probability of restoration is zero. For HDDs there is a minimum amount of time required to reconstruct the data on the HDD based on the number of HDDs on the bus, the capacity of the HDDs, the maximum sustained data rate and the amount of foreground simultaneous I/O.

A Markov model calculates the *probability* of one or more failures during the mission time, but the expected number of failures is not calculated. Estimates of the number of failures in time are usually based on *nF(t)*, where n is the total number of

units at risk at the beginning of the mission. Thus, at any point in time, t_i, the number of failures is computed.

Numerous researchers have evaluated the MTTDL method and a few have identified problems, yet failed to develop a new model to correct the issues. Malhotra [36] recognized that MTTDL is the metric commonly used to assess reliability, but acknowledged that mean values are not as interesting to system users as the probability of failing in a specific time frame. He states "...transient solution is of more interest than steady state solution." Malhotra introduces several new metrics, including $L(t)$, the probability that no data loss has occurred until time t. While Malhotra's thesis offers insights for complex modeling of fault tolerant systems, he resorts to constant failure rates and evaluates the RAID system using Markov models.

Kari [3] shows the importance of latent fault detection and its effect on the reliability and data availability of disk arrays. He proposed analytic methods, measurements on existing systems, and reliability simulations as possibilities to assess the MTTDL and data availability. He has a brief table of failure modes and causes, but his list omitted several important causes of disk drive failure which should be modeled. While Kari provides a different view of Markov analyses with improvements, his basic assumptions are the same as all others and he arrives at the same point, constant failure rates and MTTDL.

Kari, as with most other researchers listed earlier, incorrectly assumes the deterioration of the magnetic media is independent of the usage. This means that operations of reading from and writing to the disk will not cause the media to deteriorate. He assumes the media will deteriorate by itself, through what is often

termed "bit-rot", a degradation in the magnetic properties in the media causing areas previously unread and unwritten to become faulty. He assumes that the heads do not touch the media during use and overlooks many of the causes discussed in previous sections.

Courtright [32] acknowledges that disk drive failure rates may not be constant and cites the International Disk Equipment Materials Association (IDEMA) standard [33] for specifying disk drive reliability based on non-constant failure rates. However, he then assumes constant failure rates throughout his paper and asserts that MTTDL is an insightful metric.

Geist and Trivedi [34] use Markov models and constant transition rates, but add an interesting insight. They assume that the failure rate for the second failure is higher than that for the first failure, although still constant. While this is an interesting approach, there is little justification for it from a failure cause perspective. Their discussion includes only non-redundant disks as compared to mirrored disks. In spite of these research efforts, their model again assumes constant failure rates.

More recently Schwarz et al. [5] recognized the problems with undiscovered (latent) data corruptions and developed a model that includes scrubbing for an archival storage system. As other researchers, however, they assume constant failure rates and constant repair rates, a HPP at the system level, and calculate a mean time to data loss as the reciprocal of the "failure rate". While providing interesting approaches to modeling latent defects and alternative scrubbing strategies for a massive array of independent disks (MAID), the paper contains the same statistical assumptions and prevalent in previous papers. Several other errors include

multiplying a failure rate by time and asserting the result is a failure rate. The approximate equivalence between probability and frequency, shown in [5] is used without checking its validity (see eq. 8). The system modeled is a "two-way mirroring system with *N* disk drives"; a far simpler model than the RAID system considered for this new model.

$$F(t) = 1 - e^{-\lambda t} \approx \lambda t \qquad \textbf{eq. 8}$$

4.2 Assumptions and Errors in Past RAID Models

In nearly all past RAID reliability models, the emphasis has been placed on either the mean time to data loss (MTTDL) or the probability of failure in some time period. However, the expected number of failures as a function of time is more insightful to everyone from the engineers designing RAID systems, to the executives trying to understand unhappy customers, to the service personnel who are asked whether a particular number of failures is higher than expected. For both the MTTDL and the probability of failure, (erroneous) assumptions are made in order to calculate the expected number of failures. Neither of these methods really solves the problem of determining the cumulative number of double disk failures as a function of time. While all attempt to do so, this section will illustrate the statistical errors inherent in their efforts.

The greatest fallacy that afflicts all the analyses to date is the assumption that the *system* follows a homogeneous Poisson process (HPP). This significance of this fallacy is followed closely by the assumption that individual HDDs have constant failure rates and that repair/restorations have constant rates and unbounded durations.

It is important to spend a fair amount of time understanding these assumptions and their consequences. These three assumptions introduce error in the estimated number of system failures based on MTTDL and Markov models.

The goal of this thesis is not to derive new mathematics or statistics, but to use those already derived and apply them to develop a new, novel RAID reliability model. Therefore, this section provides conclusions from authorities Ascher in [47] - [52], Thompson in [53], Nelson in [54], and Crow in [55]. In the following discussion, I draw heavily on their expertise, statements and examples to support these concepts.

4.2.1 Hazard Rate versus Rate of Occurrence of Failure

To set the stage for discussions in sections 4.2.2 and 4.2.3, I must first elaborate on the distinction between "failure rate" and "rate of occurrence of failure" (ROCOF). For a set of observed times to failure having a probability density function, f(t), the cumulative distribution function, F(t), is the time based integral.

$$F(t) = \int_0^\infty f(t)dt \qquad \textbf{eq. 9}$$

If the cumulative distribution function (CDF) is known, the probability density function (pdf) can be determined by differentiation.

$$f(t) = \frac{d}{dt}F(t) \qquad \textbf{eq. 10}$$

The hazard rate (instantaneous failure rate) is a function of the pdf and CDF.

$$h(t) = \frac{f(t)}{1 - F(t)} \qquad \textbf{eq. 11}$$

When *h(t)* is independent of time, it is often referred to as the failure rate.

The failures of a *system* are described by a stochastic point process [52]. The time based number of failures is *N(t)*. The mean cumulative number of failures for the process is *E[N(t)]*. We use the following notation for the cumulative expected value:

$$V(t) \equiv E[N(t)] \qquad \textbf{eq. 12}$$

The ROCOF for the process, *v(t)*, is the derivative of *V(t)*.

$$v(t) \equiv \frac{dV(t)}{dt} \qquad \textbf{eq. 13}$$

The hazard (failure) rate and ROCOF represent two very different concepts. The hazard rate, $h(t_i)$, at time t_i, is a property of a time to failure distribution, related through both the pdf and CDF of the distribution. The ROCOF is a property of a sequence of times to failure [48]. However, under certain specific conditions, they are numerically the same [47]. This leads to confusion between the two and misuse of hazard rate instead of ROCOF. This research draws on the distinction and develops a model method consistent with cumulative number of failures, *N(t)*, and the ROCOF, *v(t)*. This research focuses on the ROCOF for determining the number of DDFs.

If a large number of similar components (HDDs, perhaps) are tested to failure, the CDF and pdf can be determined. The pdf and CDF for these components are based on the components themselves. There is no logical organization among them, such as simple redundancy, m-out-of-n, or conditional order of occurrence. If these failures follow a homogeneous Poisson process, the times from the start of the test until failure will be independent and identically distributed (iid), follow an exponential distribution and will have a constant hazard rate, λ. The MTBF will be the reciprocal of the hazard rate.

Now consider a system composed of a single component. Each time the component fails, a new component sampled from the original pdf is inserted and the system is once again operative. It should be apparent that each sample is iid. The system is completely restored to an as-good-as-new condition because the new component fails according to the exact same distribution. Since each component is drawn from the same distribution with mean $1/\lambda$, the mean time to failure for the system will also be $1/\lambda$. This is an example of a HPP process. Notice that the mean for the components is the same as the mean for the system.

Consider the same system as described above, but instead of a single component, there are multiple components in series, all drawn from the same distribution. This system will exhibit the same characteristics as the single component system. The observed *m* sequential times to failure (inter-arrival times) for the system are essentially the same as *m* observations from an exponential distribution, because, in each case the *m* observations are statistically independent observations from an exponential distribution [52] and the system is restored to as-good-as new. There is

no time memory to the system and probability of failing in any interval, Δt, is constant. This is a homogeneous Poisson process. This is the ONLY set of conditions in which the ROCOF will be numerically equivalent to the hazard rate of the components and order of occurrence is not important.

When analyzing the data, the times between failures are ordered smallest to largest, and a distribution is fit to the data. The HPP exists only when the inter-arrival times are iid. Consider a set of data used by Ascher [47] as an example of a repairable system. He lists five inter-arrival times for system failure:

14, 34, 42, 72, and 244 hours.

If ordered as above, an exponential distribution will fit the data. Suppose the sequential order of the data is different. He considers three possibilities as follows:

---14------34-----42----------72----------244

----------244----------72--------42-----34---14

------34-----14----------244----------72---------42

In the first case, the time between failures is increasing. In the second, the time between failures is decreasing. In the third, it does not appear to be either increasing or decreasing. So, while order is not important for determining a pdf, the mean, and failure rate for the components, order is critical for accurately understanding the ROCOF of the system. Before a HPP is assumed, a trend test should be conducted.

4.2.2 Non-Homogeneous Processes

A process is said to be a HPP when the probability of experiencing a specific number of failures, N, in a specific time interval, $(t, t + x)$, does not depend on time, t. It is proportional to t through a constant, λ, where $0 < \lambda < \infty$. Thompson expresses it as follows [53]:

$$P\left[N(t,t+x) = k\right] = p(k;\lambda x);\ k = 0,1,2,... \qquad \textbf{eq. 14}$$

Put another way, a process is HPP if and only if the times between failures are independent and with common exponential distribution. Components with constant failure rates may, depending on system logic, result in a ROCOF that follows a HPP. However, components with time dependent hazard rates (e. g., Weibull) will *not* result in a HPP and will *not* have a constant ROCOF. Weibull distributed times to failure are not iid, but inter-arrival times must be iid for the ROCOF to be a HPP.

Extensive field reliability data, which will be presented in great detail in Chapter 5, shows that HDD hazard rates are rarely constant. Furthermore, the HDDs in a RAID system often come from different vintages from the same supplier as well as different suppliers. Even if all HDDs were all exponentially distributed but with different hazard rates, the system would exhibit a decreasing failure rate [56] and, therefore, not follow the rules for HPP. It has been found that HDD failure distributions are often mixtures of distributions, resulting from some HDD sub-populations having failure mechanisms that other sub-populations do not have, even though the sub-populations are the same model from the same manufacturer, but

different vintage. An example of this is the failures caused by air borne contamination. When a HDD is closed up in the clean-room, it either does or does not contain large quantities of contaminants. Those that do often fail earlier than those that do not have contaminants, rendering an early wear-out mechanism for a subpopulation of HDDs.

In addition to HDD times to failure following a NHPP, the system failure logic (architecture) must be examined. A RAID system always has some level of redundancy, *N+1* in the cases we are considering. Even if all the components in a redundant system have constant failure rates, the system does not. Consider three simple configurations in which success is defined by the following:

- one-out-of-one
- one-out-of-two, and
- two-out-of-three.

Table 2 shows the hazard rate for three configurations. When all the constituent elements of the system have constant failure rates, only the serial system (reliability-wise) has a constant failure rate. In the 1-out-of-2 and 2-out-of-3 systems, the hazard rates are time dependent; therefore, the systems do not have constant failure rates and do not conform to the requirements for a HPP. However, the mean time to failure is not time dependent, and the failure rate is not the reciprocal of the MTBF as is clear from a review of Table 2. RAID systems under consideration in this thesis require n-out-of-n+1 to operate, where *n* is between 1 and 24. These do not conform to HPP requirements.

Table 2 - Hazard Rates for Three Different Levels of Redundancy

# of units required and total	R(t)	E(t)	h(t)
1 out of 1	$e^{-\lambda t}$	λ	$\frac{1}{\lambda}$
1 out of 2	$2e^{-\lambda t} - e^{-2\lambda t}$	$\frac{3}{2\lambda}$	$2\lambda\frac{1-e^{-\lambda t}}{2-e^{-\lambda t}}$
2 out of 3	$3e^{-2\lambda t} - 2e^{-3\lambda t}$	$\frac{5}{6\lambda}$	$6\lambda\frac{1-e^{-\lambda t}}{3-2e^{-\lambda t}}$

4.2.3 Errors Resulting from the HPP Assumption

To date, all RAID analyses provide estimates of MTTDL or the probability of failure (Markov models), and the concepts of failure rate and ROCOF are interchanged. Sequence is ignored and component concepts are used on the system. Under a HPP, the expected rate of failure, $v(t)$, is constant and numerically equal to the component failure rate, λ, and the two are interchanged. From renewal theory, the expected number of failures for a single system is λt and for n replications, nλt.

Under the non-Homogeneous Poisson Process (NHPP), the cumulative number of failures follows a Poisson distribution, but the mean, $E[N(t)]$, is not proportional to time [52], so eq. 15 must be employed.

$$V(T) = E[N(T)] = \int_0^T v(t)dt \qquad \textbf{eq. 15}$$

Ascher [47] summarized it best stating, there is little connection between the properties of component hazard rates and the properties of the process that produces a sequence of failures. He also states that an increasing (decreasing) component failure

rate has no connection with the tendency for times between successive failures (at the system level) to become smaller (larger) [50]. Furthermore, times between successive system failures can become increasingly larger even though each component hazard rate is increasing [48]. In another publication [52] he goes on to explain that for a NHPP, the inter-arrival times do not come from a single distribution so it is meaningless to even try to fit a distribution to them or calculate mean and variance.

Since estimates of *V(t)* using the MTTDL and Markov models are based on the invalid assumption of HPP for the system, the results are questionable, at best. The component failure rates are not constant, the ROCOF of the system is not constant and the system ROCOF is not the reciprocal of the MTBF (MTTDL). Markov models are no better than MTTDL at resolving problems associated with assuming HPP for the system. Thompson [57] provides an important observation relative to using probability to assess *E[N(t)]*. Again using a life distribution with pdf of *f(t)*, the probability of failure before some time, *t*, is described by the CDF (see eq. 9). He then notes that the number of components that fail as a function of time, *N(t)*, is a random variable that is binomially distributed as in eq. 16.

$$\Pr\left[N(t)=k\right]=\binom{n}{k}\left[F(t)^{k}\right]\left[1-F(t)\right]^{n-k} \text{ for } k=0,\ 1,\ldots.\ n \qquad \textbf{eq. 16}$$

For a binomial distribution, the expected value, *E[N(t)] = nF(t)*. However, replacing *E[N(t)]* with *nF(t)* and dividing by *n* leads to an interesting observation. Differentiating *V(t)* to get the ROCOF results in the expression shown in eq. 17.

$$\frac{d}{dt}\frac{E[N(t)]}{n} = \frac{d}{dt}F(t) = f(t) \qquad \textbf{eq. 17}$$

This is density function, not a hazard rate! So the ROCOF cannot be determined by using *nF(t)*. This is the process used when Markov models are employed.

Chapter 5 Realities of Field Data

Numerous data sets for HDDs used in RAID systems in the field were analyzed to better understand the potential impact of real data on the model. The results confirm that the assumption of constant HDD failure rates in RAID systems *must* be revised. Some of the results have been published [58], [59], [60], [61], [62], and [63], and new data analyses are included to provide new insights. This section provides examples of analyses to illustrate the following points:

- Failure rates are *usually not* constant in time
- Failure distributions depend on manufacturers
- Failure distributions change as a function of manufacturing vintage
- Failure distributions are affected by RAID architecture and recovery techniques
- Failure distributions are often mixtures of multiple distributions which are difficult to segregate into the constituent distributions

When one compares the following results to the manufacturers' advertised "MTBFs", you will find great differences between the two. Knowing that HDD failure distributions usually do not have constant failure rates invalidates the MTTDL and Markov models used to date and the failure rate times time calculation used to determine numbers of failures using these two methods. The data presented below provide the basis for better DDF estimates using the model presented in Chapter 6 and Monte Carlo simulations, Chapter 7.

In general, failure rates are dynamic throughout the HDD life as well as the production life. Much of the data presented to illustrate this is for significant subpopulations over a significant time. These numbers do not represent an entire product line and should not be interpreted as representative of all the HDDs in a particular system, although any specific system may be composed of HDDs from a specific manufacturing vintage. Two systems of the same make and model sitting side by side in a user environment may consist of a highly reliable "angelic" system and a very unreliable "evil twin" because of the difference in composition of HDDs. The significance of this data is to provide justification that failure rates are generally not constant and that "evil twin" systems can be a reality.

5.1 The Bathtub Myth

In most text books, it is stated that component failures follow a "bathtub" curve. First, the "infant mortality" portion starts at time zero and has a decreasing failure rate. When the hardware prone to early life failures diminishes significantly, the "useful life" begins, wherein the failure rate fairly constant. At some point "wear-out" begins and the failure rate begins to increase. The bathtub curve is actually not a single distribution, but is composed of three different distributions. The three theoretic segments are shown in Figure 9. As will be seen in Section 5.2 through 5.5, HDD field data does not match the theory. Significant findings are as follows:

- Decreasing HDD failure rates in the short term
- Decreasing failure rates after many months/years of increasing failure rates
- Early life wear-out (increasing failure rates in early life)

- Extrapolation beyond the existing data is an art based on understanding failure mechanisms

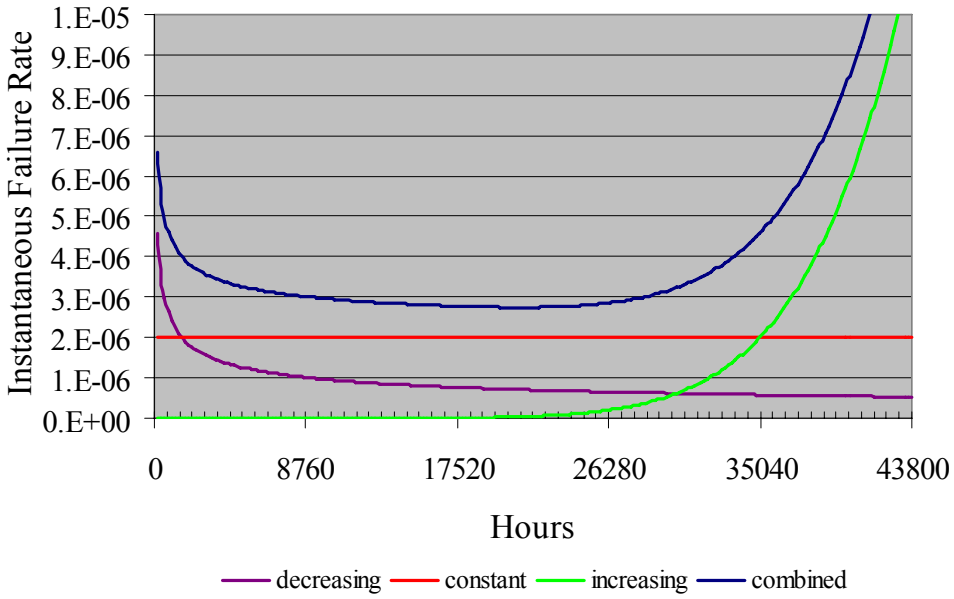


Figure 9 - The Theoretic Bathtub Curve
This shows the theoretic failure rate as a function of time

5.2 The Importance of Vintage Analyses

Vintage analyses are especially effective in discerning the effects of design or manufacturing process changes over time. Significant design changes in HDDs are infrequently introduced after mass production begins because customers must re-qualify the products, a very expensive and time consuming endeavor. However, small manufacturing process changes occur on a regular basis to enhance manufacturing yields. Yield improvements are readily apparent on a weekly basis, during which a supplier may produce 100,000 HDDs. Reliability is presumed to improve along with

yields, but only after significant design changes are hundreds of HDDs run through weeks of reliability testing to verify improvements or even the lack of degradation.

The definition of a vintage is rather arbitrary but can be viewed as a change that affects the entire subpopulation of HDDs as a function of build date. If significant design or manufacturing process changes are implemented, HDDs produced before and after the change are reasonable vintage groups. For the data analysis shown in Figure 10, there were no significant changes as confirmed by the manufacturer. In these analyses, it was found that a population of 10,000 to 15,000 HDDs provided a reasonable basis for a vintage population.

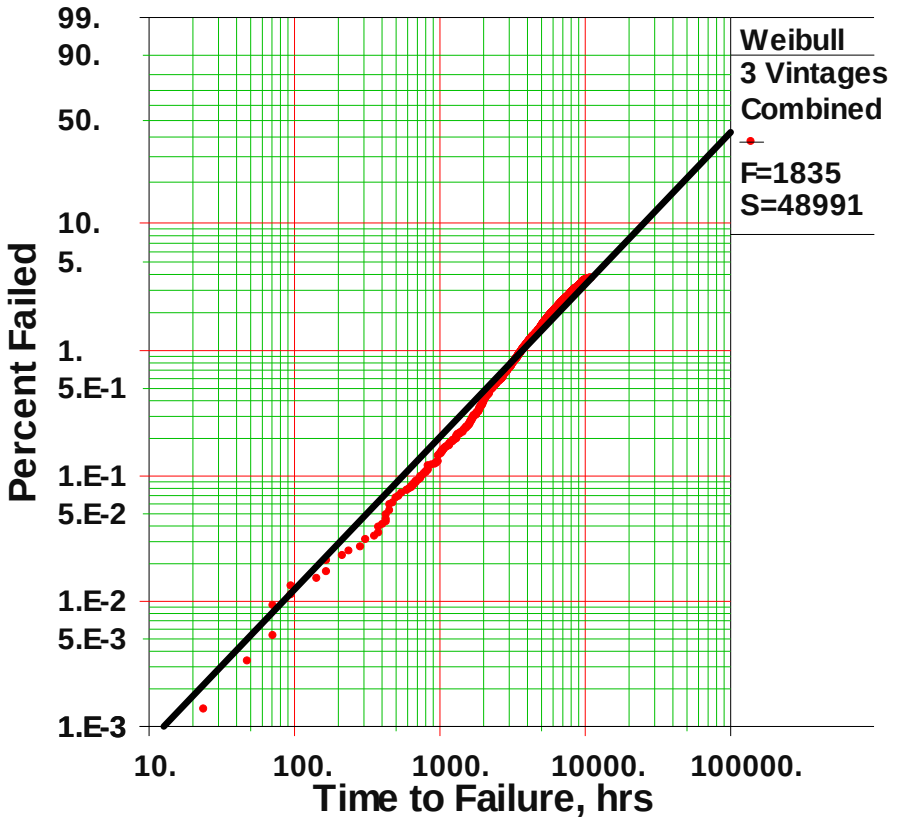


Figure 10 - Weibull Plot of Combined Vintages

Since the monthly usage exceeded 10,000 HDDs, monthly shipments sufficed as the definition of a vintage. The plot contains three months of production combined. The data are plotted in red as individual failure times and the straight black line is the "best fit" using the Maximum Likelihood Method (MLE). MLE is preferred over rank-regression because of the high number of right-censored HDDs. The data appears to have a slight "s-shape" to it. The single Weibull distribution with the calculated slope of 1.2 is not an especially good fit. This data was then divided into three constituent vintages, each one month in duration and the three vintages are plotted in Figure 11.

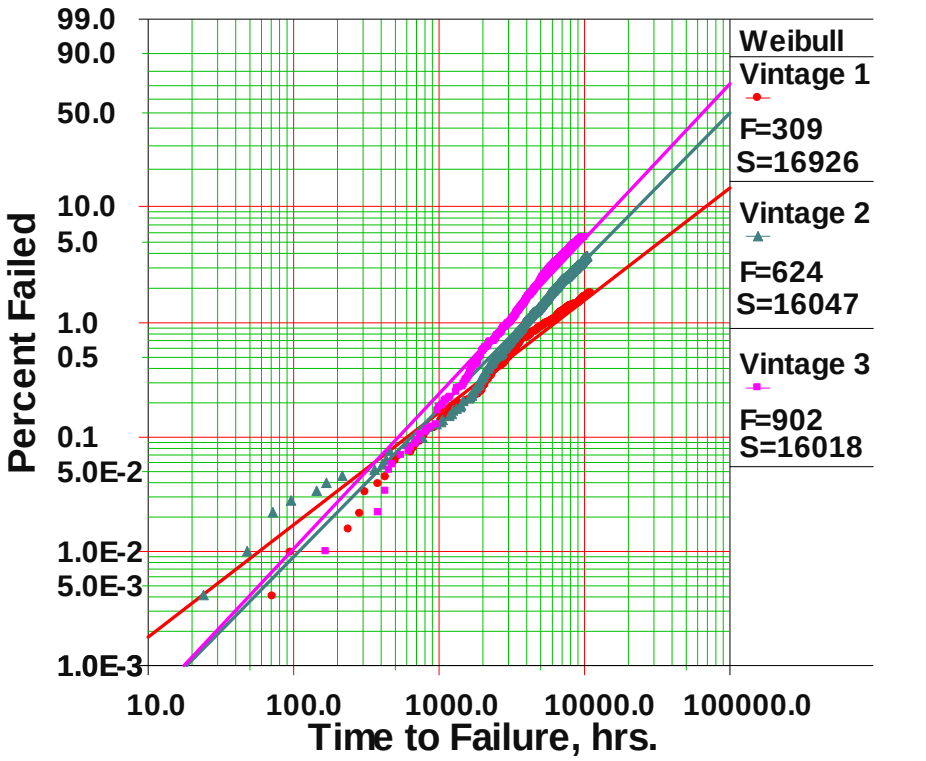


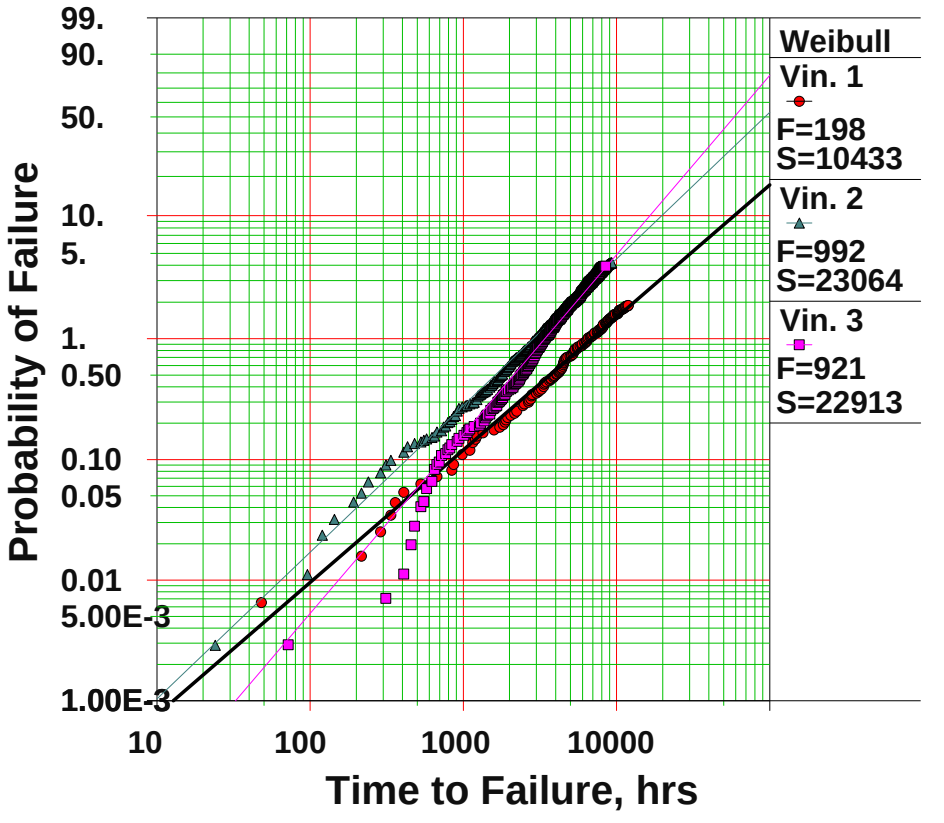
Figure 11 - Weibull Plot of Three Constituent Vintages

In Figure 11, note that the three months all have significantly different slopes, β, and characteristic lives, η. Even when 90% confidence limits were placed on the data, there is a statistical difference between the three lines. On the right side of the plots, "S" indicates the number of HDDs censored (succeeded) in the population and "F" indicates the number failed. The parameters of the three distributions are shown below the plot. One line is nearly exponential, for which $\beta = 1.0$. The other two are increasing failure rates with $\beta = 1.3$ and $\beta = 1.4$. The percent of HDDs that will fail in 1 year (8760 hours) is 1.4, 2.9 and 4.1 respectively for these parameters. While vintage 1 has a fairly constant failure rate, both vintages 2 and 3 have increasing failure rates or wear-out within the first 30 days of usage.

Figure 12 shows another three vintages. Again, two of the slopes (β's) indicate increasing failure rates while the third is nearly 1.0, indicating a fairly constant failure rate.

Observation #1: HDD failure distributions can vary significantly from month to month even for the same HDD from the same manufacturer.

Observation #2: "Infant" wear-out is observer in two of the vintages. Wear-out is not just a late-life phenomenon as described in the bathtub curve.



β1=1.0987, η1=4.5444E+5
β2=1.2162, η2=1.2566E+5
β3=1.4873, η3=7.5012E+4

Figure 12 - Second Group of Vintages

5.3 Competing Risks versus Distribution Mixtures in Data

When all units in a population can fail from any of several causes, the causes are said to present competing risks. The reliability for the unit is the product of the reliabilities due to each of the causes. For *n* different failure mechanisms, *m*, the reliability is the product:

$$R_S = R_{m1} \, x \, R_{m2} \, x \, R_{m3} \, x \, ... \, R_{mn} \qquad \textbf{eq. 18}$$

The HDD population in Figure 13 shows the effects of competing risks. All the HDDs were subject to the same failure mechanisms. Early in their life, mode "A" dominated and later mode "B" dominated. The failure analyses shows that nearly all HDDs will suffer from this long-term wear-out phenomenon related to lube buildup on the heads causing high-fly writes. This population does not exhibit an initially decreasing failure rate as in the bath-tub curve, but it does exhibit useful life and (long-term) wear-out. This population did not contain distinctive vintages in subpopulations. Extrapolation of the percent to fail will follow an extension of the steep right end of the curve.

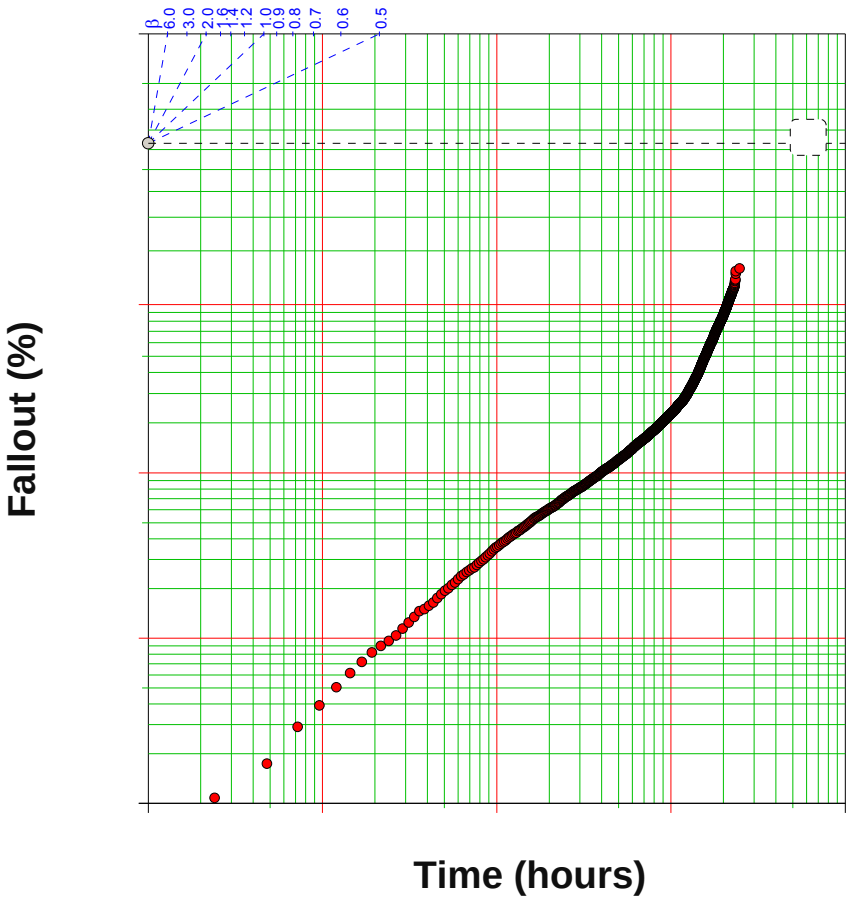


Figure 13 - Plot Showing Competing Risks (No Vintage Effects)

5.4 Mixtures of Distributions

Figure 14 shows a plot of a single HDD product from a single supplier. When the population is taken as a whole, denoted as "composite" in the figure, it shows a distinctly increasing failure rate that tends to diminish until, in the later times, it has a decreasing failure rate[3]. When segregated into five vintages, also plotted in Figure 14, all vintages exhibit similar behavior. This indicates that all vintages contain at least two distinctly different subpopulations regardless of the vintage.

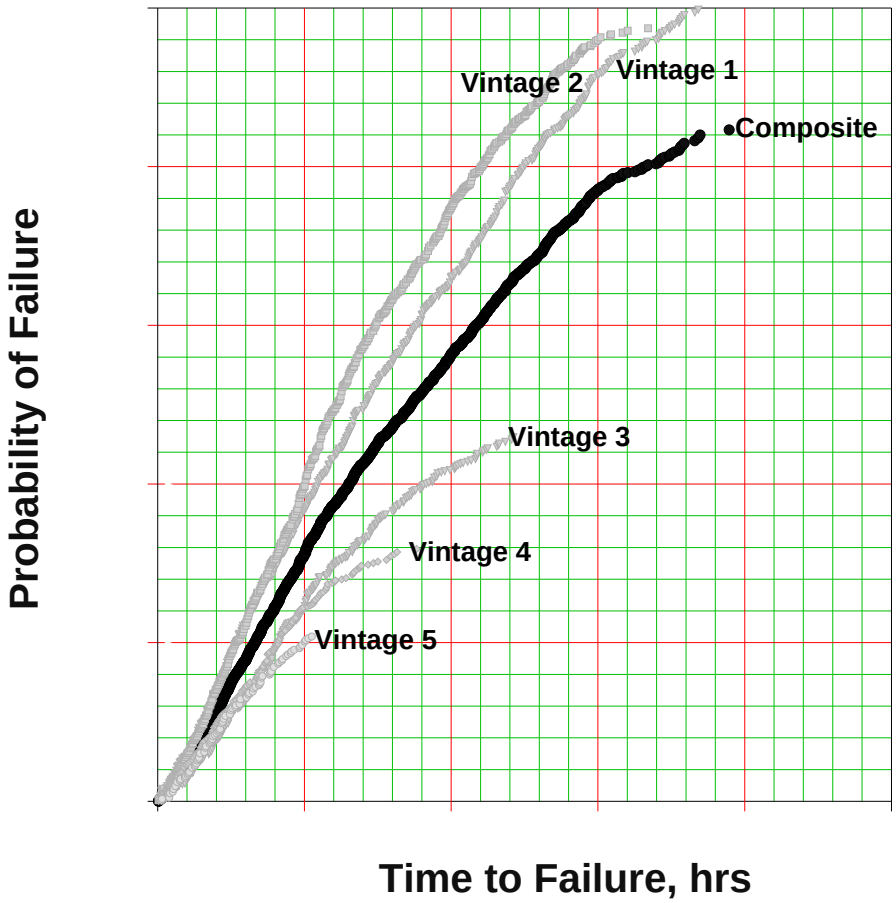


Figure 14 - Data Plot Showing "Distribution Mixtures"

[3] When plotted on linear paper, a slope greater than 1.0 indicates an increasing failure rate. If equivalent to 1.0, it has a constant failure rate and if less than 1.0, has a decreasing failure rate.

Failure analyses by the HDD manufacturer showed these characteristics resulted from the manufacturing process in which some HDDs contained high levels of contamination when the cover was put on and some did not. Those with higher levels of contamination followed an "infant wear-out" distribution and those with lower levels of contamination had either a constant or decreasing failure rate[4]. Thus, some fraction of all HDDs on any given day had excessive levels of contamination. These distinct distributions in the population render what is termed "mixed distributions".

5.5 Highly Complex Distributions

In general, I tried to fit data first to Weibull distributions because of its flexibility. Under some circumstances it is difficult to discern the difference between a Weibull and ln-Normal distribution, but from a Weibull it is easy to determine whether the failure rate is increasing, constant or decreasing. Figure 15 shows some very complex distributions. Clearly, only one is even close to a Weibull distribution. The other two contain mixtures of distributions, competing risks and vintage changes. Population HDD #3 initially has a fairly constant failure rate. It then exhibits a decreasing failure rate followed by a strongly increasing failure rate. This product experienced a design change that substantially improved its head reliability (sub-population with much higher reliability) but then all the HDD heads are subject to lube build-up and high-fly writes.

[4] Proschan [56] demonstrated that a population consisting of two subpopulations, both with constant failure rates, will exhibit what appeared to be a decreasing failure rate. As the units with the higher failure rate fail, the remaining population has a higher proportion of units with the lower failure rate. Unless specifically analyzed to discern the truth, it is not possible to tell whether the HDDs with low contamination levels were truly decreasing or whether the remaining population appeared to have a decreasing failure rate as a direct result of Proschan's observation.

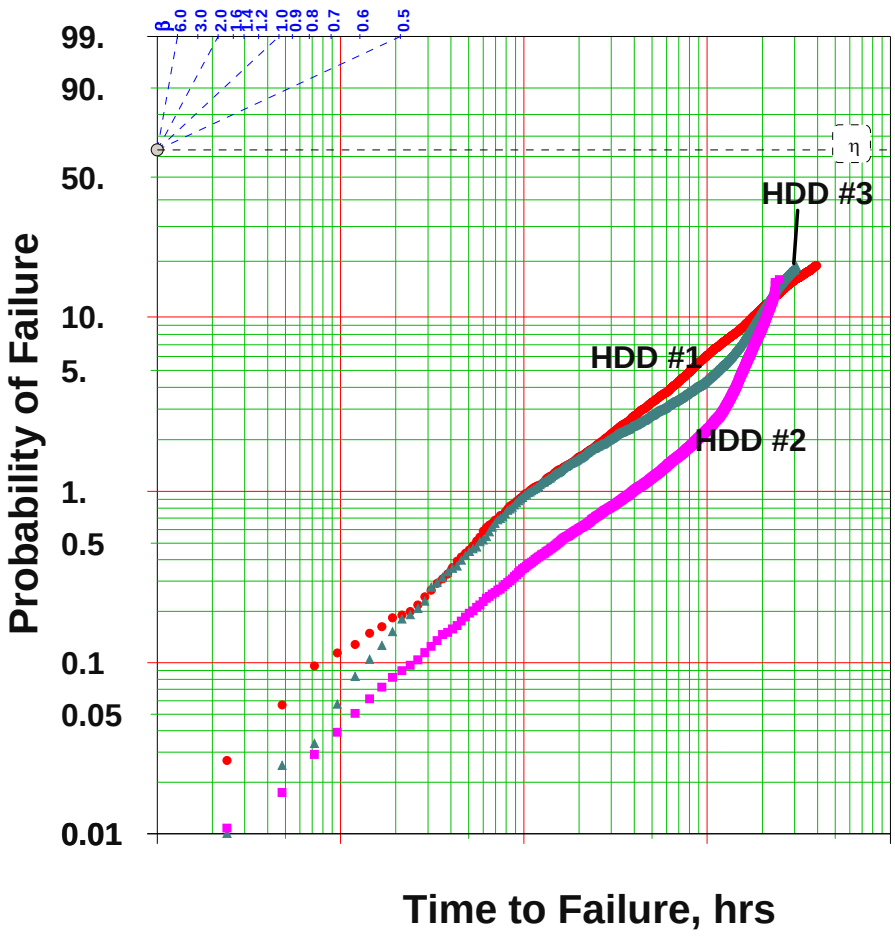


Figure 15 - Three Separate Populations: Two are Complex

5.6 Data Summary

The data presented are a significant contribution to the knowledge of HDD failure distributions. Based on my literature search, I could find no other HDD data analyses that clearly illustrate the distributions and the effects of vintage, mixtures, contamination, and wear-out. These data clearly show that the assumption of a constant failure rate over the life of the product is generally invalid. Most subpopulations analyzed do not have a constant failure rate. In fact, throughout the analysis of nearly one-half million HDDs from four manufacturers, consisting of 10

different products (excluding capacity differences), constant failure rates are rarely seen, even in monthly vintages.

The significance of this observation is essential to this thesis. Lack of constant failure rates at the HDD level negates the ubiquitous assumption that the system will follow a homogeneous Poisson process. This assumption appears, either implicitly or explicitly, in all other RAID system analyses. The rest of this thesis is devoted to understanding the significance of this finding by developing a model that includes non-constant failure rates, non-constant repair/restoration rates and accounts for a variable scrubbing of latent defects.

Observation #3: Vintage differences can result from either competing failure mechanisms or "mixtures" of mechanisms.

Observation #4: Complex distributions can result from mixtures of causes that cannot be segregated. The HDD failure rate does not follow a closed form distribution (e.g., Weibull) and predictions of future behavior may not be possible.

Chapter 6 New Model Logic

The bases for needing a new RAID reliability model are presented in Chapter 4 and Chapter 5. At the heart of the errors are the non-inclusion of latent defects and the assumption that the HDD failures and repairs, and the RAID group (as a system) failures and restorations, all follow a homogeneous Poisson process. This Chapter presents the new model that accounts allows the HDD failures to follow a non-HPP, incorporates the logic resulting from the failure modes and mechanisms identified earlier, and incorporates operating policies such as "scrubbing" durations. This Chapter begins with some definitions (6.1), describes a common operational profile for the RAID group (6.2), the model logic (6.3), the model in detail (6.4) and finally the transition distributions (6.5).

6.1 Definitions

There are no “standard” definitions for many of the terms used in describing HDD failures. However a consistent set of terms is paramount in creating a model. This paper uses the following definitions.

Correctable Data Errors: Corrupted data that can be corrected by using the HDD’s onboard Error Correction Code (ECC) algorithms. “On-the-fly” corrections can be performed for small errors (few bytes) in less than a revolution. Somewhat larger data corruptions, 20-80 bytes, require more calculation time to correct and so may require another revolution. These errors are corrected by the drive and do not

require RAID for data recovery. These are typically, SCSI[5] "01" codes, such as 01-17 and 01-18. Correctable errors are not part of the RAID reliability calculation.

Uncorrectable Data Errors: Corrupted data that cannot be corrected using the HDD's onboard ECC algorithms, off-track (offset) reads, parametric changes to the heads, and numerous other methods to recover data. Typically these are SCSI "03" errors, such as 03-03, 03-11, 03-0C and 03-13.

Recoverable Data: Uncorrectable Data Errors that are recoverable at the system level using the remaining HDDs in the RAID group, including parity data. Data recovered at the RAID level is rewritten to the same drive that had the error, but to a physically different location. The offending location and some amount of "track padding" are logically mapped out, never to be used again.

Latent Defects: Data with undetected defects (data corruption) in the media. Since data corruption is not discovered until the data is read, it may be corrupt and either undiscovered or discovered but uncorrected. Both of these constitute a latent defect. Latent defects contain uncorrectable, recoverable or unrecoverable data and may be due to poor writing, bad media or post-writing corruption.

Operational Faults: (Hardware) Faults (errors) in which data or servo tracks cannot be read even if the data is not corrupted (SCSI code 04). Typically, this requires a hardware failure, such as a failed head or bad motor bearing, lubrication buildup on a head or electronic circuitry failure. These errors can occur during a *read* or *write*. Staying on-track requires the following to be good: media servo tracks,

[5] SCSI is the small computer standard interface

read/write heads, motor bearings, and numerous electronic components. Even during a "write" the servo track must be read, so inability to read servo tracks will result in an operational failure. Operational errors do not include defective media or corrupted data.

Media Defect: A part of a disc surface that has a physical condition such as scratch, bump (from contamination buried in the media), or a pit that either prevents data from being written correctly or results in erasure or corruption of data after being written. For this research, all media defects have corrupted data or will not allow data to be written without becoming corrupted.

Reallocation: If corrupted data can be recovered using RAID, the bad sectors on the affected disk are mapped out and the data is saved in a new location on the disk.

Reconstruction: If a HDD is no longer able to read or write data (operational error or SMART threshold exceeded), all the data on the affected HDD is reconstructed from parity and the rest of the HDDs in the RAID group and written on a spare HDD.

6.2 System Operation Profiles

Most HDD manufacturers agree that system operation can greatly affect HDD reliability. However, none has developed a verifiable relationship between the various kinds of use and the effects on reliability. That is, HDD reliability is affected by the length (number of successive blocks) of writes and whether they are sequentially or randomly written. This research assumes that order of operation (sequential versus random), the length of writes (number of blocks), the number of seeks and the seek

length all affect reliability, but their effects are averaged out over the all customer use profiles. Therefore, these effects are inherently contained in the field data collected at the HDD level.

A. All System Operating Profiles:

Systems are fully utilized 24 hours per day, 365 days per year and are rarely stopped. This is common practice in the business critical applications in the United States. In other countries, the systems are sometimes turned off during off-peak hours and over some fraction of the weekends to conserve electricity.

Data written to HDDs are not checked for integrity after writing. This is done to enhance performance. Write errors are not identified until the next time the data is read. Recoverable errors can be corrected using all the remaining disks in the RAID group. Unrecoverable errors cannot be corrected even using the other disks in the RAID group, either because it is an "operational defect" or because the required data is not available on the remaining RAID disks (one HDD has already failed or the specific data necessary to correct has also been corrupted). ALL data on ALL other HDDs are required to reconstruction of a full HDD in a RAID group.

Total number of operational disks in a RAID group is $N+1$. The capacity equivalent of one HDD is required for parity. So, whether RAID-4 or RAID-5, all data is stored on N HDDs, the minimum number required to serve data per specification. All HDDs experience the same workload. No HDD (even in RAID-4) is used significantly more than any other, so all HDDs in a RAID group have the same failure distributions, although the model is capable of changing the failure distribution of one HDD in the RAID group for RAID-4 if desired.

Adjacent track interference is not considered in this analysis. Data may be wholly contained on a single HDD or striped across HDDs. Parity is calculated based on physical HDD blocks, so the location of the file and the amount of striping is not important. Location of data is striped solely for performance reasons.

B. Spares:

Availability of spares affects the “at risk” time when one HDD in the RAID group has failed. Having a spare HDD in the system and operating means that the delay time to begin reconstruction can be treated as “zero”. However, "hot-standby" spares are spinning and the heads are moving so they can induce media defects, damage heads and experience all types of operational failures. Operational failures are readily detected and replaced, but they are not part of the cut-sets for the model.

Latent defects in the media of a spare HDD do not cause HDD failure. Data is written to a location with defective media just as any other data-write to a HDD. Spares with latent media defects do not fail a reconstruction during a “write” unless the defect is in an area used to simply stay on track, such as a servo block. Operational failure of a spare will prevent reconstruction but otherwise not affect any data on the other HDDs in the RAID group.

C. Attributes during reconstruction:

During reconstruction, 100% of all idle time is spent on reconstructing data on the “new” HDD. The RAID group is still processing I/O commands and serving data as requested. When I/O and reconstruction occur simultaneously, the reconstruction uses 40% of the HDD bandwidth, 40% of the I/O bus bandwidth, or 40% of the cpu cycles, whichever is the most limiting.

New (additional) data may be stored completely on one of the original RAID group HDDs or striped across other HDDs, including the one under reconstruction. During reconstruction, the entire HDD is reconstructed, even the blocks that have no useful data. Since the HDD may be "zeroed" before it leaves the integrator's factory, parity should always be valid.

During reallocation or reconstruction, corrupted data may be written to the new sectors or HDD. If a bad block is read from one of the HDDs being used for reconstruction the reconstructed HDD contains corrupted data but is not immediately failed. The bad data block is tagged as questionable. If the data on that block is rewritten (updated) before it is read, then the data is no longer "bad", no data is lost and the HDD is not considered as failed. However, if the bad sector on the newly reconstructed HDD is read before being re-written, then the block is considered bad, the data is corrupted and the HDD fails. This model assumes that locations with latent media defects will be read before being rewritten.

D. Latent Defect Scrubbing:

Latent data corruptions that have occurred during operation (reading, writing or spinning - since spinning can cause T/A's and corrupt data) can be discovered by scrubbing. Scrubbing is a process in which data is read and compared to the parity. If they are consistent, no action is taken. If they are inconsistent, the corrupted data is recovered and rewritten to the same physical location on the HDD or, if the media is defective, the recovered data is written to new physical sectors on the HDD and the bad blocks and adjacent blocks are mapped out and not reused.

Scrubbing is a background activity performed on an as-possible basis so it does not affect system performance. The time required to scrub an entire HDD is a random variable that depends on the HDD capacity and the amount of foreground activity. During times of heavy I/O the scrubbing rate may not keep up with the bit error rate (BER) or added media defects (from T/As, etc). So, time to completely scrub may be very lengthy. It is possible that the BER is so high that new errors are being added to the data at a faster rate than the scrubbing is removing them.

Scrubbing can prevent latent defects from being discovered during a full HDD reconstruction, but does not reduce the number of latent media errors that occur throughout the life of the HDD. By recovering corrupted data before a reconstruction, latent media defect related double-disk failures are reduced. Since scrubbing requires reading and writing data, scrubbing acts as time to failure accelerator for HDD components with usage dependent time to failure distributions. For example, if scrubbing causes the slider to move too slowly across the disc surface, disc lubrication can build up on the head causing it to fly too high above the media. If the head is too high when performing a write, the magnetic field strength in the written media can be too low to be read.

From a modeling perspective, scrubbing reduces the opportunities for latent defects to accumulate, thus reducing the probability of failure during reconstruction. If not scrubbed, the period of time to accumulate latent defects starts when the HDD first begins operation in the system and depends on BER and amount of data transferred and operational failure mechanisms that corrupt data such as thermal asperities.

6.3 Model Logic

In RAID-4 and RAID-5 configurations, a single additional HDD within the RAID group is employed for redundancy. As part of the write process, an "exclusive OR" calculation generates parity bits that are also written to the RAID group. Error correcting codes (ECC) *on the HDD*, and parity *across the HDDs* is a common method to assure accurate data transfer and recording. ECC uses Boolean operations to encode blocks of data, interleaving the data and the ECC bits. On each read command, user data and ECC are read. If inconsistent, the data is corrected on the fly (less than one revolution), data integrity preserved and performance is not degraded. ECC strength is enhanced by interleaving multiple blocks of data so that if a large physical area of a disk (many bits) is not readable, it does not affect all the data in a single block, but spreads the errors out over multiple blocks. 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.

Losing access to a HDD occurs when the supporting hardware or software prevents data from being written to or read from the disk. Examples include loss of electric power, failure of the processor or failure of the host-bus adapter. Once the support function is fully restored, the HDDs are accessible and, assuming graceful shutdown upon loss of the support function, the data on the disks remains uncorrupted and readable. This model does not consider lost access, but focuses on data loss.

There are four distributions to consider based on observed HDD failure modes and mechanisms and common system operating rules; 1) time to operational failure, 2) time to latent defect, 3) time to operational repair and 4) time to latent defect

recovery (through scrubbing). System failure occurs when two HDDs fail simultaneously. An operational failure (Op) is one in which no data anywhere on the HDD can be read, even though the data may have no defect whatsoever. Removal and replacement of the HDD is the only resolution to an operational failure. Latent defect (Ld) refers to unknown or undetected data corruption. Latent defects can exist because either the data was not written well initially or the data was destroyed after successfully being written to the disc.

Poorly written data can be a result of the inherent bit-error rate (BER), but this has been determined to be a second order effect. More likely, data is poorly written because of high-fly writes or other head related magnetic, electrical or mechanical problems. Data is often destroyed after being successfully and properly written. This can be the result of contaminants buried in the coatings on the media surface, loose contamination within the HDD enclosure, or thermal erasures from brief or repeated head-disc contacts.

Remedies for latent defects occur only upon reading the corrupted data and rely on reading the data on all other HDDs in the RAID group and the associated parity bits to reconstruct the lost data. If small amounts of data are lost, the reconstructed data is physically written to another good section of the HDD and the faulty section is mapped out to prevent reuse. When a HDD is removed due to an operational failure, this same process is used to reconstruct all the data lost on the failed HDD. MTTDL calculations include only a single failure rate for all HDD failures and a single repair rate for all restoration (reconstruction) processes.

A significant modeling difference results from the order of occurrence for latent and operational failures. Latent defects go undiscovered until the corrupted data is read. Only at that time is the data corrected through use of the parity bits. Data that is not read remains in a corrupted state indefinitely. If an operational failure occurs after the existence of a latent defect on any other HDD, the data cannot be reconstructed on the replacement HDD because data required is corrupted or missing. Thus, a latent defect followed by an operational failure results in a double disk failure (DDF).

While the HDD experiencing reconstruction may suffer write errors, these will be corrected during the next read or remain as latent defects. Their creation during the 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 do not occur during reconstructions. Multiple HDDs with latent defects do not constitute DDF unless they happen to coexist within the same block of data across more than one HDD, also an extremely low probability event which is not included in the model.

In recent years, the problem of latent defects has been recognized by some system integrators and been significantly reduced by data “scrubbing” [5]. During scrubbing, data on the HDD is read and checked against its parity bits even though the data is not being requested by the user. The corrupt data is corrected, bad spots on the media are mapped out and the data is saved to good locations on the HDD. Since this is done as a background activity and does not impede performance, it is rather slow. The scrub time may be as short as the maximum HDD and data-bus data rates

permit. But, depending on the foreground I/O demand and the HDD capacity, it also may take weeks or months to complete the scrub.

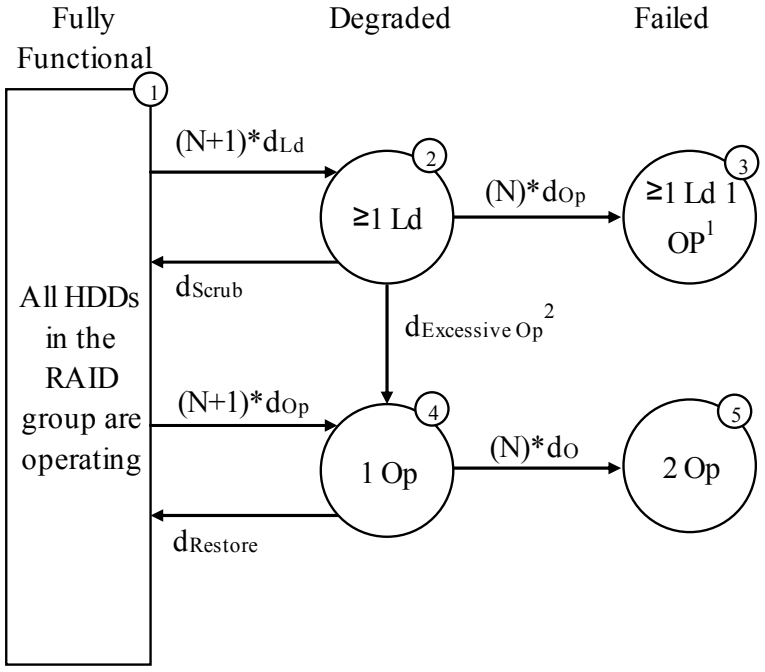
In summary, the two conditions that result in a DDF are two simultaneous operational failures and an operational failure that occurs after a latent defect has been introduced and before it is corrected. Simultaneous latent defects do not constitute failure. This conditionality is not considered in MTTDL or the Markov models reviewed.

6.4 Model Description

The model logic is partially depicted in Figure 16. While this logic appears to represent a Markov model, it certainly is not evaluated as a Markov model. The notation of distributions and transitions from state to state is generic. Evaluation process and the Monte Carlo simulation will be described later. In State 1, the data and parity HDDs are good, there are no latent defects and a spare HDD is available. State 2 represents the condition in which 1 or more HDDs have developed latent defects. The transition from State 1 is a function of the *N+1* HDDs developing a latent defect according to the failure distribution, d_{Ld}. 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 3 occurs in accord with the operational failure distribution d_{Op}.

Transition from State 2 to State 4 occurs when media defects become so prevalent on the HDD which had the latent media/data defect, that the HDD causes a time-out for the system while trying to reallocate large numbers of defective blocks, or because the HDDs SMART threshold, such as “excessive block reallocations”, was

exceeded. In this transition massive media problems render the HDD as inoperative, just like any other operational failure. 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}.



Note 1: Op failure must be a different HDD than the one with a Ld.
Note 2: This transition does not have an explicit rate. It is included in the measured rate of "Op" from field data.

Figure 16 - State Diagram for N+1 RAID Group

State 4, the second possible transition from State 1, represents one operational failure. The transition to this state from State 1 is based on the full complement of *N+1* HDDs and in accord with the distribution for operational failures, d_{Op}. There are two possible transitions out of State 4. A second (simultaneous) operational failure

results in transition to State 5, also a DDF state. The operational failure is 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 distribution, $d_{Restore}$, includes the delay time to physically incorporate the spare HDD and has a minimum time to reconstruct based on HDD the capacity and the maximum transfer rate and concurrent foreground data I/O activity.

6.5 Transition Distributions

Four component related distributions required for this model: time to operational failure (TTOp), time to restore an operational failure (TTR), time to generation of a latent defect (TTLd), and time to scrub (TTScrub) HDDs for latent defects. All Weibull distributions have the following parameters:

Location parameter = γ (gamma)

Shape parameter = β (beta)

Characteristic Life = η (eta).

A. Time to Operational Failure (TTOp)

Chapter 5 illustrated the myriad of possibilities for TTOp distributions. This analysis uses a single TTOp distribution to illustrate the difference between a representative failure distribution in the proposed model and constant failure rates of the MTTDL method. A Weibull failure distribution with a slightly increasing failure rate is used. The characteristic life, η, is 461,386 hours. The shape parameter, β, is 1.12. These parameters are taken from a field population of more than 120,000 HDDs

that operated for up to 6,000 hours each. The location parameter is 0. The distribution is shown in Figure 17.

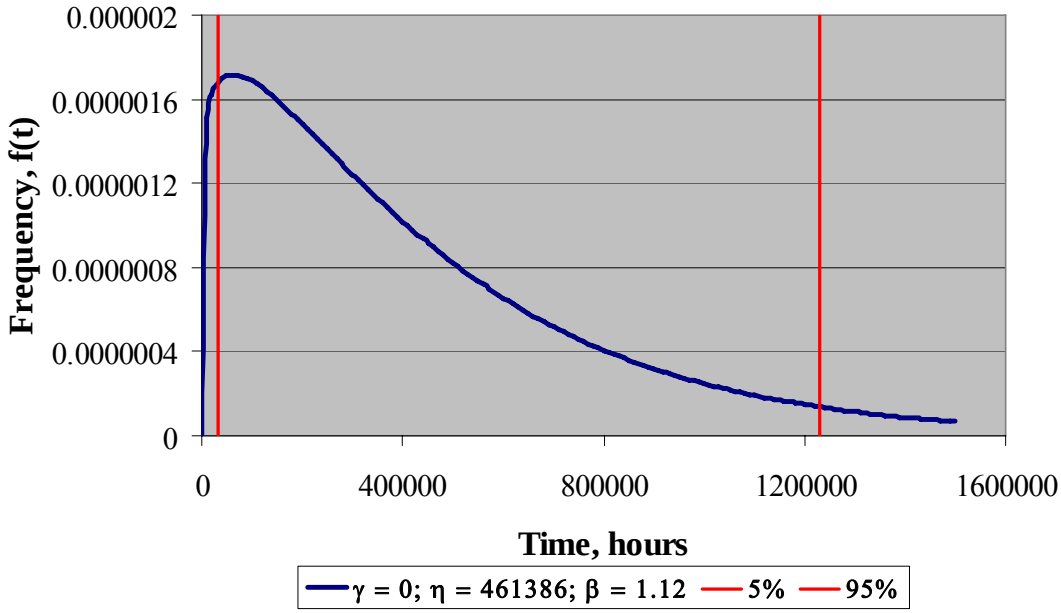


Figure 17 - TTOp failure distribution

Weibull distribution with location parameter = 0, shape = 1.12 and characteristic life = 461,386 hours

B. Time to restore (TTR)

A constant restoration rate implies 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. But this is clearly unrealistic for two reasons. First, there is a finite amount of time required for the HDD to reconstruct all the data on the HDD. It 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. Fibre channel HDDs can sustain up to 100MB/second data transfer rates, although 50MB/sec is more common. The data-bus to which the RAID group is attached has only a 2 giga-bits per second capability. Thus, in a RAID group of 14, a 144GB HDD on a fibre channel interface will require a minimum of 3 hours with no other I/O to reconstruct the failed HDD. A 500GB, Serial ATA HDD on a 1.5Gb data-bus will require 10.4 hours to read all other HDDs and reconstruct a HDD that has been replaced.

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 complete in a prescribed amount of time.

Three parameter Weibull distributions are used for the time to repair/restore distribution. The minimum time of 6 hours is used for the location parameter, γ. The shape parameter, β, of 2 generates a right-skewed distribution, and the characteristic life, η, is 12 hours. Figure 18 shows the shape of the probability density function for this distribution.

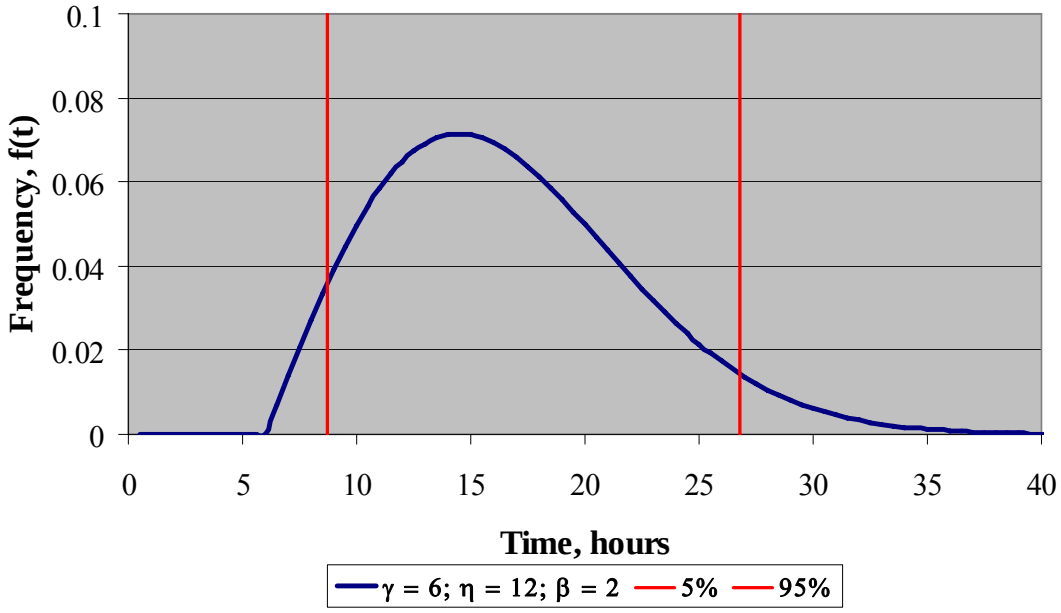


Figure 18 - TTR Distribution

Location parameter is 6 hours, shape parameter is 2.0, and characteristic life is 12 hours. No repairs are completed within 6 hours and 95% of all restorations are completed in 26.77 hours.

C. Time to latent defect (TTLd)

Personal conversations with engineers and design teams from 4 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 aren't 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 3 and the following discussion is the justification.

Gray [64] concludes that the "bit error rate", (BER) is fairly inconsequential in terms of creating corrupted data. Schwartz [5] claims the rate of data corruption is five times the rate of HDD operating failures. Analyses of corrupted data identified

by specific SCSI error codes and subsequent detailed failure analyses shows that the rate of data corruption for all causes is significant and must be included in the reliability model.

Network Appliance completed a study in late 2004 on 282,000 HDDs used in RAID architecture. The read error rate (RER), over three months, was $8x10^{-14}$ errors per Byte read. At the same time, another analysis of 66,800 HDDs showed a RER of approximately $3.2x10^{-13}$ errors per Byte. A more recent analysis of 63,000 HDDs over 5 months showed a much improved $8x10^{-15}$ errors per Byte read. In these studies, data corruption is verified by the HDD manufacturer as a HDD problem and not a result of the operating system controlling the RAID group.

While Gray [64] asserts that it is reasonable to transfer $4.32x10^{12}$ Bytes/day/HDD, the study of 63,000 HDDs read $7.3x10^{17}$ Bytes of data in 5 months, an approximate read rate of $2.7x10^{11}$ Bytes/day/HDD. The following studies use a high of $1.35x10^{10}$ Bytes/hour and a low of $1.35x10^{9}$ Bytes/hour. Using combinations of the RERs and amount of Bytes read yields the hourly read failure rates Table 3.

Since usage varies so greatly by customer and by application, it is assumed that the latent defects are created at a constant rate. Therefore, an exponential distribution is assumed for this distribution. Two different values are used; 9259 hours is the mean for high bytes-read rate, low read-errors per byte; 92950 hours is the mean for low bytes-read, low read-errors per byte. While some latent defects are created by wear-out mechanisms, data is not available to discern wear-out from those that occur randomly at a constant rate. A distribution is shown in Figure 19.

Table 3 - Range of Average Read Error Rates

Read Errors per Byte per HDD		Bytes Read per Hour: Low Rate 1.35×10^{9}	Bytes Read per Hour: High Rate 1.35×10^{10}	
Low	8.0×10^{-15}	1.08×10^{-5}	1.08×10^{-4}	Err/hr
Med	8.0×10^{-14}	1.08×10^{-4}	1.08×10^{-3}	Err/hr
High	3.2×10^{-13}	4.32×10^{-4}	4.32×10^{-3}	Err/hr

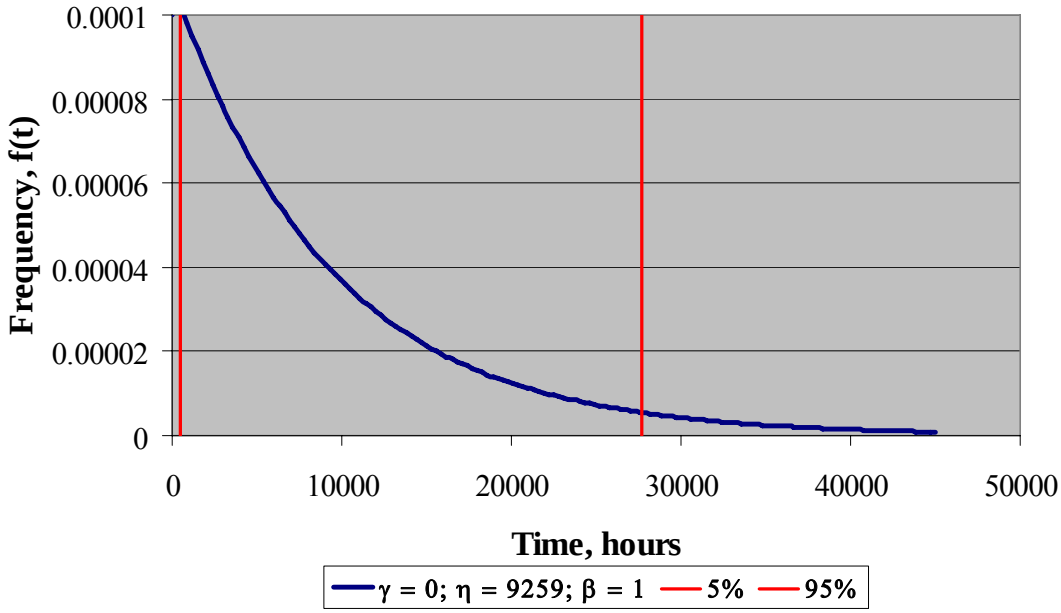


Figure 19 - Exponential distribution for time to occurrence for latent defects
This figure shows the distribution for high bytes-read per hour, low read-errors per byte, a mean of 9259 hours.

D. Time to scrub (TTScrub)

Latent defects (data corruptions) can occur during almost any HDD activity: reading, writing or simply spinning. If not corrected they will result in lost data when an operational failure occurs. 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 is read and compared to the parity. If they are consistent, no action is taken. If they are inconsistent, the corrupted data is recovered and rewritten to the HDD. If the media is defective, the recovered data is written to new physical sectors on the HDD and the bad blocks are mapped out.

Since scrubbing is a background activity performed on an as-possible basis, it does not affect system performance. The time required to scrub an entire HDD is a random variable that depends on the HDD capacity and the amount of foreground activity. 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 and is based on the error generation rate and the hourly data transfer rate.

Since scrubbing requires reading and writing data, scrubbing can act as a time-to-failure accelerator for HDD components with usage dependent time to failure mechanisms. The optimal scrub pattern, rate and time of scrubbing is HDD specific and must be determined in conjunction with the HDD manufacturer to assure that HDD operational failure rates are not increased.

Scrubbing, as with full HDD data reconstruction, has a minimum time to cover the entire HDD. The time to complete the scrub is a random variable that depends on HDD capacity and I/O activity. The operating system may invoke a maximum time to complete scrubbing. The simulations in this paper use a 3-parameter Weibull distribution for time to scrub. 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 $\gamma = 36$, $\eta = 168$ hours, and $\beta = 3$ is shown in Figure 20. The distribution for an alternate, quicker scrub is shown in Figure 21, with $\gamma = 0$, $\eta = 12$ hours and $\beta = 3$.

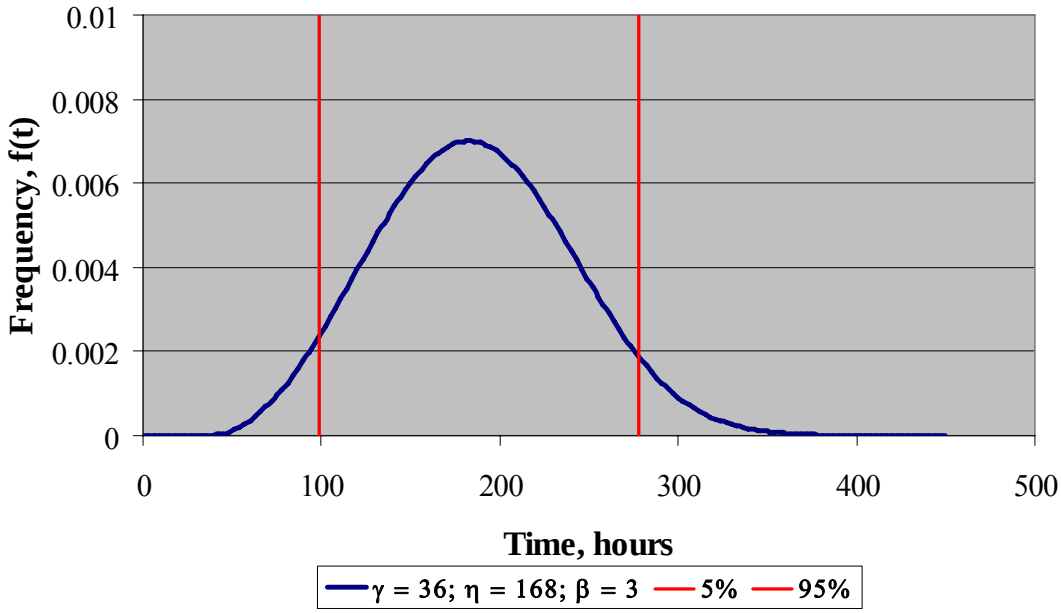


Figure 20 - Time to scrub density function

After a defect has been introduced, the time to scrub and remove the defect is assumed to follow a "normal" looking distribution. The probability of a complete disk scrub in less than 36 hours is 0. 95% of all scrubs will be completed within 278.2 hours.

For a sensitivity study, the shape of the TTOp distribution was varied while the characteristic life kept constant at 461,386 hours. The probability density and cumulative distribution functions are shown in Figure 22 and Figure 23 respectively.

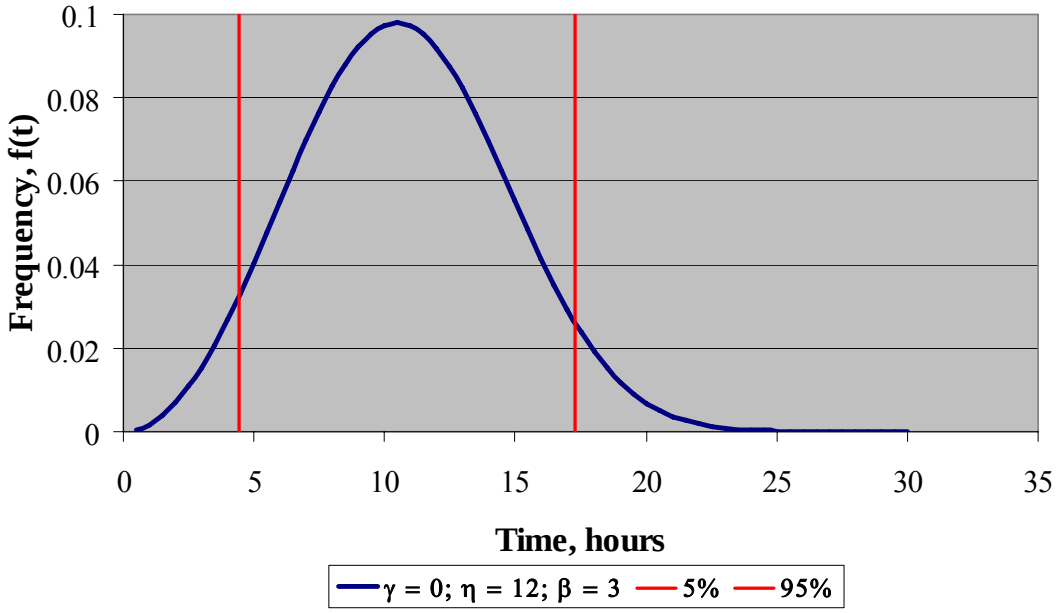


Figure 21 - Short scrub distribution

This distribution has no delay between the occurrence of the defect and the possible time of completing the scrub. With a characteristic life of 12 hours and a shape of 3, 5% of the scrubs will be completed in less than 4.46 hours and 95% will be completed by 17.3 hours.

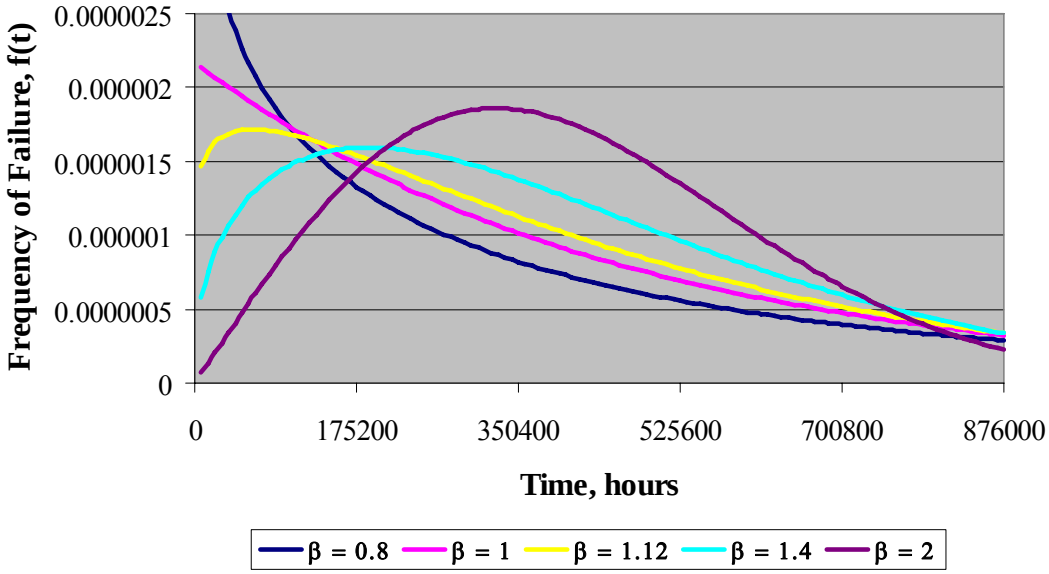


Figure 22 - Probability density function for TTOp sensitivity study

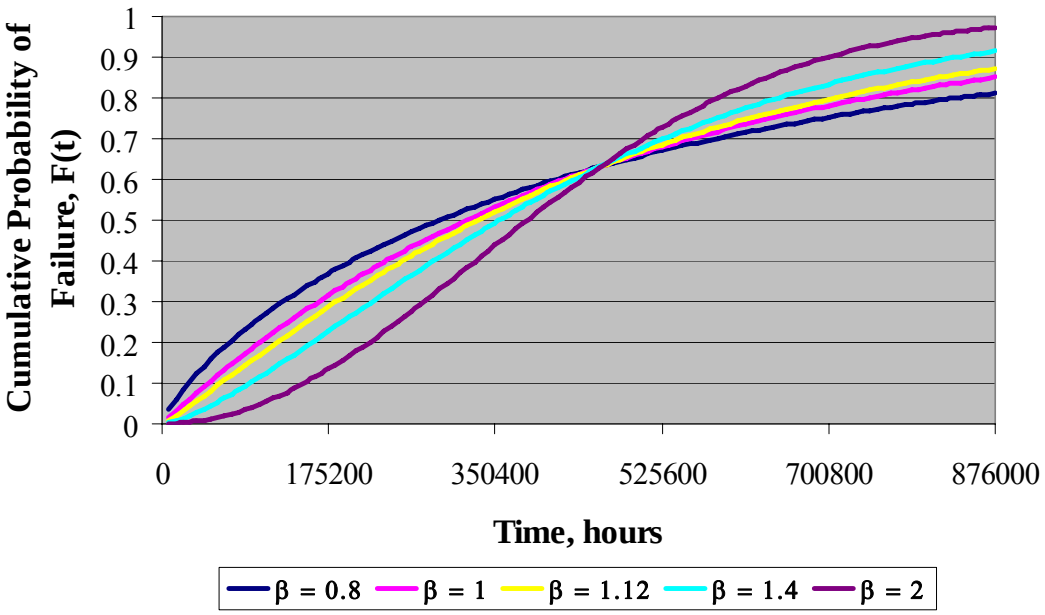


Figure 23 - Cumulative distribution function for study of Weibull shape parameter

The cumulative probabilities of failure cross over each other at the characteristic life for all values of beta. The shape of the distribution prior to the characteristic life is critical to the analysis of the RAID system. For example, at time 175200 hours, the cumulative probability of failure for shape parameter $\beta = 0.8$ is0.37, whereas for $\beta = 2$, the CDF is 0.13; a three-fold difference.

Chapter 7 Sequential Monte Carlo Simulations

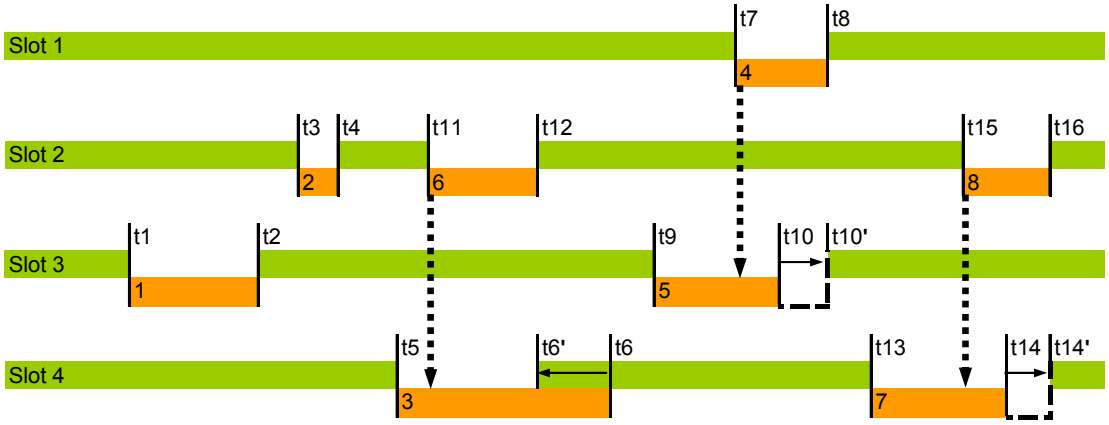
The most useful reliability model for a RAID system must be able to properly include the following:

- Failure rates that are time dependent (non-constant)
- Failure rates that are manufacturing vintage dependent
- Repair models that have an offset (non-zero location parameter) and are time dependent (not constant)
- Permanent errors (data corruptions) that occur during non-write operations
- Latent defects (data corruptions) that persist
- Latent defect elimination (through data scrubbing)

Additionally, it must be easy to change the RAID group size and set a mission time. The output must provide the maximum amount of information that can be easily manipulated to render failure quantities, frequencies and probabilities for the RAID group. This model focuses on developing an understanding of the time dependent nature of double disk failures (DDFs). Since this problem does not conform to renewal theory [53], the simple multiplication of failure rate and time does not render expected number of failures in time. The times to failure generated by the model are used to plot the mean cumulative failure distribution [65]. A sequential Monte Carlo process was selected because of its adaptability to account for the conditions described above.

In a sequential Monte Carlo simulation, the time dependent, or chronological, behavior of the system is simulated [66]. For each HDD in the RAID group, a (time) sample of each transition time in Figure 16 is taken from the associated underlying transition distribution (time to failure or time to restore). The operating and failure times are accumulated until a user specified mission time is exceeded, 10 years for this research. During that time, the sequence of HDD failures, repairs, latent defects, scrubs and DDFs are tracked. Each sequence of sampling required to reach 10 years is a single simulation and represents one possible system operating chronology. Depending on the input parameters, as many as 50,000 simulations are needed to develop the cumulative failure function, described in [65]. Since each simulation is equivalent to a single RAID group operating for the 10 year mission, the complete set of simulations is equivalent to watching 50,000 systems for 10 years and monitoring the DDFs.

Figure 24 shows the sequential process used in this simulation. For simplicity, only four HDD slots are shown. The graph looks like a digital timing diagram with the "high signal" representing the operating (non-defective) condition and the "low signal" representing the failed (defective) condition. Throughout this process, each HDD "slot" in the RAID group carries its own times to failure (both TTOp and TTLd) and times to restore (TTR and TTScrub) distributions. Upon a DDF involving a HDD with a latent defect, the TTR for the failure is the same as the concomitant operational failure time. The Excel Visual Basic program and the flow diagram for the program are in the Appendix.



Comparison	DDF?	Next sample processes
TTF TTR		
Old t1 t2		
New t3 t4		
Is t1<t3<t2?	no	Sample new TTF & TTR for slot 3 (t9 & t10)
Old t3 t4		
New t5 t6		
Is t3<t5<t4?	no	Sample new TTF & TTR for slot 2 (t11 & t12)
Old t5 t6		
New t11 t12		
Is t5<t11<t6?	yes	Shift restart time (t6) to coincide with restoration of slot 2 (t12) Sample new TTF & TTR for slot 4 (t13 & t14)
Old t11 t12		
New t9 t10		
Is t11<t9<t12?	no	Sample new TTF & TTR for slot 2 (t15 & t16)
Old t9 t10		
New t7 t8		
Is t9<t7<t10?	yes	Shift restart time (t10) to coincide with restoration of slot 1 (t8) Sample new TTF & TTR for slot 3 (t17 & t18) not shown
Old t13 t14		
New t15 t16		
Is t13<t15<t14?	yes	Shift restart time (t14) to coincide with restoration of slot 2 (t16) Sample new TTF & TTR for slot 4 (t19 & t20) not shown

Figure 24 - "Timing" diagram for sampling TTFs and TTRs

Initially, a TTF and TTR are sampled for each HDD slot, denoted t1 to t8. Then, pair-wise comparisons are made as indicated below the "timing" diagram. If the time to failure for the second-youngest event, t3, occurs after the youngest event, t1, and before the repair of the youngest failure, t2, then a failure would occur. After the comparison, new times are sampled and the comparisons shift to the next older time to failure, t5, and its repair time, t6. In the 3rd, 5th and 6th comparisons, DDFs occurred. TTF for a HDD with a latent defect is the same as the TTR for the Operational failure (see events #3, 5 and 7). The process continues until there is only one slot with a cumulative TTF less than the mission time.

The Monte Carlo simulation process begins with sampling a TTOp and a TTLd for every HDD. The TTOp and TTLd arrays are kept separate and sorted separately in order to implement conditional failure logic (order of events is critical). For both of the two HDDs with the shortest times to failure (or defect), a time to restore (or time to scrub) is sampled. If both are operational failures, a DDF exists when the second shortest TTOp occurs during the interval in which the other HDD was failed. If any sampled time to failure or cumulative run time exceeds the mission time, that slot is excluded from the calculation (sorting) process. This is true whether it occurs on the first sample or after many samples. This process is reiterated until all the cumulative operating times, TTOp plus TTR, for all HDD slots have exceeded the mission time. Once a DDF has occurred, a subsequent one cannot occur until the first is restored.

If both the shortest and second shortest event times are for latent defects, the time is not marked as a system failure time because two latent defects will not fail the system, and the sampling process continues. If one event is an operational failure and one is a latent defect, a DDF exists when the operational failure occurs after the latent defect has occurred and before the scrub process corrects the corrupted data from the latent defect. This means that the second shortest time to failure must be the operational failure and the shortest must be the latent defect.

Corrupted data introduced as a result of the write process during a reconstruction does not cause a DDF. So, a system failure does not occur when the shortest event time is the operational failure and the second shortest is a latent defect. If the defect is introduced onto the new disk to which data is being reconstructed, it will be corrected next time it is read. The only time that a latent defect will cause a DDF during

reconstruction is when it is a non-write related cause (thermal asperity) occurring in an unread sector. This is an extremely low probability event and so is omitted.

If no DDF is detected, then the TTR (or TTScrub) that has already been sampled and used in the preceding comparison is added to the earliest time to failure. A new TTOp (or TTLd) is sampled, added to the previous sum, and the complete set of times to failure for all HDDs in the RAID group are once again resorted and reduced if the cumulative time exceeds the mission time. The process is repeated until there is only one HDD slot left.

It could be argued that one-half of the latent defects would occur in locations that had already been read (not a DDF) and one-half would occur in locations that had not yet been read (DDF). However, the data collected for developing TTLd distributions cannot discern between the relative contributions from the many possible causes.

Chapter 8 Model Validation

The results were generated using Microsoft Excel™ with the @RISK™ add-in from Palisade. Visual Basic for Excel (VBE) was also used for storing intermediate data, sorting routines, making comparisons and printing outputs to text files for post processing and charting. Using the notation of Ascher [48], these analyses measure the rate of occurrence of failure (ROCOF) for the system (not the same as the failure rate) and present it as a mean cumulative failure function (MCF). The MCF measures the cumulative number of failures in a period of time for a repairable system and can be greater then unity [65].

The theory behind a Markov process is exactly the same as the process used for this simplified Monte Carlo based model. The Monte Carlo simulation can be viewed as a discrete time to failure implementation of the Markov process, which requires many iterations, sometimes as many as 100,000 to render the probabilities that are produced using commonly available computer codes to evaluate Markov models. Since the Monte Carlo model used for validation uses only constant failure and repair rates, the ROCOF and the Monte Carlo results will be numerically the same even though the concepts are different. The need, therefore, is not to validate the theoretic equivalence of the two, but the implementation of the Monte Carlo simulation in VBE code to make sure it has no errors.

One nice feature about VBE is the ability to step through the code as it executes. Values for all variables (inputs, intermediate calculations and outputs) are visible. This way, each calculation, decision, loop and result can be verified visually, one line of code at a time. The code created is rather simple by most software developers'

standards. Of the 14 subroutines, six are used only once in a simulation, for input, initial filling of arrays and for output. Of the other 8 subroutines, four are repeatedly used to gather new samples from distributions and two are sorting (operational sorts are separated form latent defect sorts). There are only six decision statements in the main part of the program; "If" or "else" conditions. Thus, validation, while tedious, was complete and detailed.

The correctness of the simulation was assessed through two comparative studies. The first study duplicates the conditions of a simple Markov model in which the distributions for all operational failures and operational repairs were assumed to follow an exponential distribution. The latent defects were removed from the model by setting an input parameter. The combinations involving latent defects are eliminated from the outset. The only combinations remaining are multiple operational failures, as in the Markov model.

This particular set of conditions omits the effects of latent defects, yet still uses the VBE code-stream (loops, "if statements", counters, comparisons) in the same manner as in cases in which latent defects are included. These assumptions and conditions produce a model that can be directly compared to both the MTTDL equation and a simple Markov model.

The Markov model for a RAID group of 14 is shown in Figure 25. All transition rates are assumed to be constant, with the failure rate, λ, of $2x10^{-6}$ (MTBF = 500,000 hours), and repair rate, μ, of 0.0208 (MTTR = 48 hours). A calculation using the equation 2 yields MTTDL = 28,617,216 hours (3,266 years). With the assumptions used, the Monte Carlo simulation produces a HPP system response so the ROCOF of

the system can be compared numerically to the reciprocal of the MTTDL. This hourly "rate" is compared to the results of the Markov model in Table 4. In Figure 26 the "rate" from the MTTDL calculation is plotted with the results of 10,000 simulations.

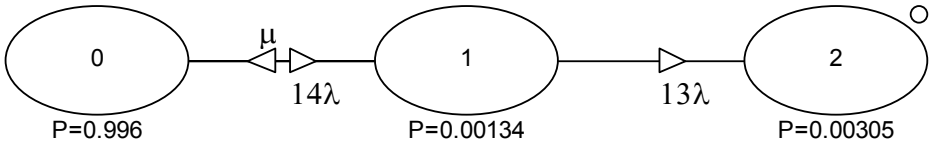


Figure 25 - First Markov model to validate Monte Carlo model and VBE code
In state 0, any of the 14 HDDs can fail so the transition rate from state 0 to state 1 is 14λ. Similarly, in state 1, any of 13 HDDs can fail, so the transition rate has a multiplier of 13.

For additional code confirmation, a second model comparison was performed in which the latent defects were included and treated as a second path for failure with constant failure rates. The Markov model for this case is shown in Figure 27 for a RAID group with 14 HDDs. The path for operational and latent defects is kept separate, with the operational failure leading to State 1 from State 0 and latent failures to State 3 from State 0. The latent defect failure rate is the same as the operational rate, $2x10^{-6}$ per hour and the repair rates are the same for both operational and latent defects, 0.0208 per hour. The intent of this is to verify the subroutines and coding relating to the latent defects. This is not the model used in the RAID analyses. Again, the simulation (Figure 28, Run #2, 2a and 2b) matches the linearized Markov result well.

Table 4 - MTTDL Failure Frequencies Compared to Markov Model Probabilities

Time, hrs.	MTTDL	MKV
0	0	0
730	2.551E-05	2.380E-05
1460	5.102E-05	4.920E-05
2190	7.653E-05	7.470E-05
2920	1.020E-04	9.660E-05
3650	1.275E-04	1.260E-04
4380	1.531E-04	1.510E-04
5110	1.786E-04	1.760E-04
5840	2.041E-04	2.020E-04
6570	2.296E-04	2.270E-04
7300	2.551E-04	2.530E-04
8030	2.806E-04	2.780E-04
8760	3.061E-04	3.040E-04
13140	4.592E-04	4.560E-04
17520	6.122E-04	6.090E-04
21900	7.653E-04	7.610E-04
26280	9.183E-04	9.140E-04
30660	1.071E-03	1.070E-03
35040	1.224E-03	1.220E-03
43800	1.531E-03	1.520E-03
52560	1.837E-03	1.830E-03
61320	2.143E-03	2.130E-03
70080	2.449E-03	2.440E-03
78840	2.755E-03	2.740E-03
87600	3.061E-03	3.050E-03

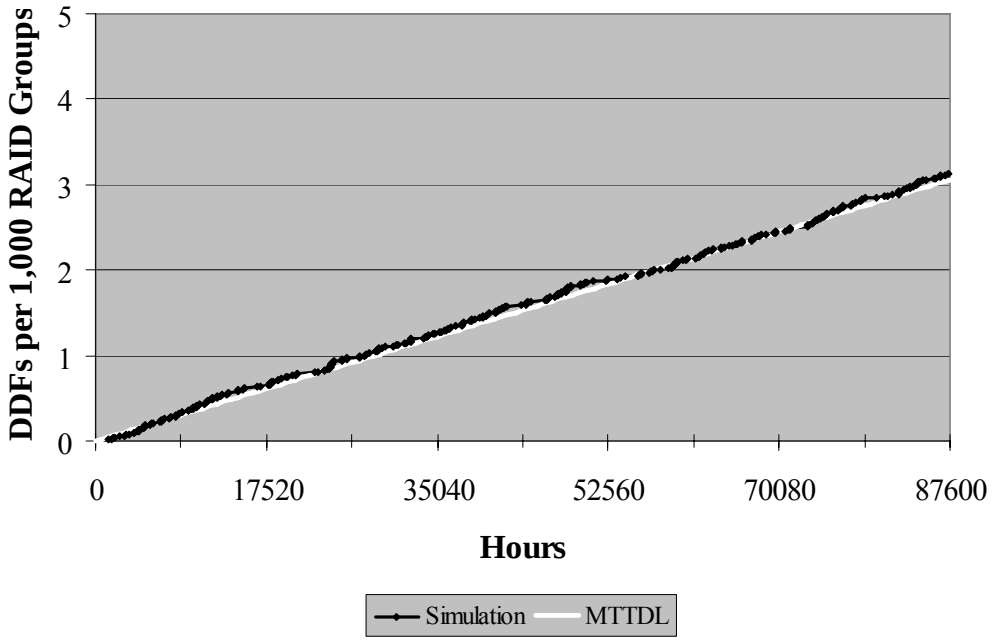


Figure 26 - Results Comparison for Markov Model and Monte Carlo Simulation

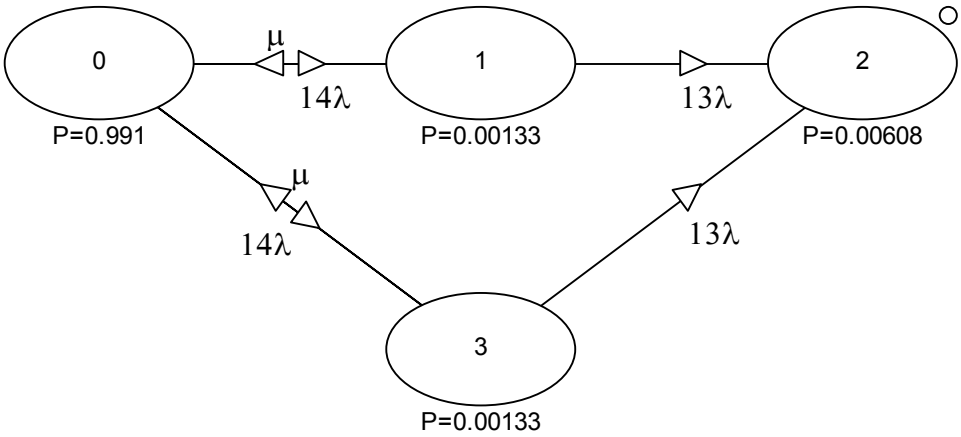


Figure 27 - Second Markov Model to Validate VBE Code

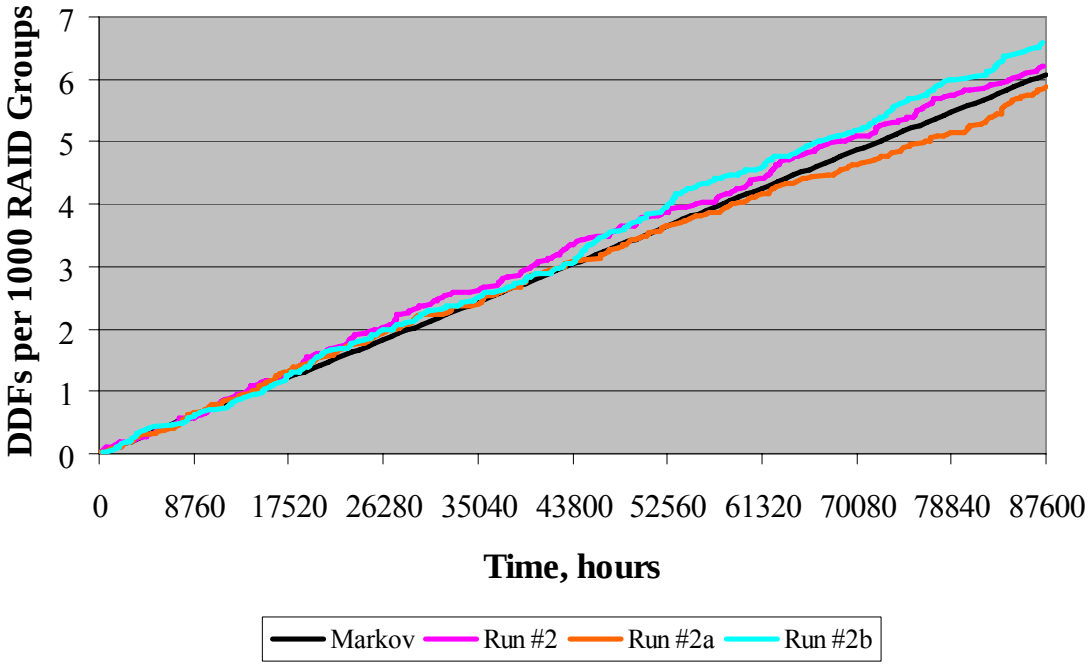


Figure 28 - Results from Second Markov and Monte Carlo Comparison

Chapter 9 Results

So far, I have presented justifications for the underlying causes for my thesis:

a) HDD failures often do not follow a homogeneous Poisson process. Failure rates are often increasing or decreasing, not constant

b) HDD failures do not come from a single population distribution. That is, significant time-to-failure variation exists across populations of HDDs from:

- different manufacturers
- different "families" from the same manufacturer
- different vintages in a single "family" from a single manufacturer

Additionally, I have provided the statistical bases behind the assertion that the MTTDL and Markov models are incorrect for this type of analysis and I have provided the electro-mechanical bases to justify the existence and creation of latent defects (undiscovered data corruptions).

In this section, I complete the defense of my thesis by showing:

c) RAID group (system) failures may not follow a homogeneous Poisson process, so estimates of the number of DDFs are incorrectly calculated when using renewal theory and assuming a HPP

d) Latent defects in HDD media are either ignored or are modeled too simplistically, omitting numerous DDFs

e) The RAID system logic modeled does not properly account for conditional order of latent defects and operational failures

To understand what is driving system behavior, a series of analyses is developed that begins with simple assumptions, similar to the HPP assumption associated with both MTTDL and Markov models, and gradually builds in complexity. HDD failure rates and repair rates are initially assumed to be constant and latent defects non-existent. Time dependent failure and restoration rates are introduced and latent defects are added in, first without scrubbing then with scrubbing. The effects of RAID group size is explored along with effects of changing only the TTOp failure rate from decreasing to constant, to increasing while including latent defects and scrubbing. Lastly, I present actual field data form Network Appliance systems in the field that supports the model results.

Since the model does not assume that the system ROCOF follows a HPP, it provides powerful insights as to the true system behavior in time. The most significant observations to come out of these results are as follows:

Observation 5: RAID group (system) failures do not follow a homogeneous Poisson process.

Observation 6: The new model predicts a much higher number of DDFs than estimated using MTTDL or Markov models.

Observation 7: Latent defects combined with operational failures dominate the DDFs.

Observation 8: The time to scrub is critical to the number of DDFs calculated.

Observation 9: When all the HDD times to failure are drawn from the same distribution, the Weibull shape parameter can create significant differences in the resulting number of DDFs.

Observation 10: The field observed read error rate is critical to estimating the number of DDFs

Observations 7 through 10 are important because the MTTDL and Markov models have never included these concepts and it is not clear that they are capable of including them.

9.1 Comparisons

A series of studies was performed assuming a RAID group of 8 (7 data +1 parity) and a mission duration of 87,600 hours (10 years). The first four cases investigate only the effects of the distribution shape for the TTOp and TTR. Latent defects are not included. Case #1 assumes TTOp and TTR distributions have constant failure rates. Case #2 assumes a time dependent rate for the TTOp, Case #3 assumes a time dependent rate for TTR, and Case #4 assumes time dependent rates for both TTOp and TTR. Effects of latent defects are explored in Cases #5 and #6. Scrubbing is added in Case #7 through #10. Parameters for Cases 1-10 are shown in Table 5.

Table 5 - Summary of Input Parameters

		TTOp			TTR			TTLd			TTScrub		
Code	Case	γ	η	β	γ	η	β	γ	η	β	γ	η	β
c-c-na-na	1	0	461386	1.00	0	12	1.0		n/a			n/a	
ft-c-na-na	2	0	461386	1.12	0	12	1.0		n/a			n/a	
c-mt-na-na	3	0	461386	1.00	6	12	2.0		n/a			n/a	
ft-mt-na-na	4	0	461386	1.12	6	12	2.0		n/a			n/a	
c-c-c-na	5	0	461386	1.00	0	12	1.0	0	9259	1.0		n/a	
ft-mt-c-na	6	0	461386	1.12	6	12	2.0	0	9259	1.0		n/a	
	7	0	461386	1.12	6	12	2.0	0	9259	1.0	6	336	3.0
	8	0	461386	1.12	6	12	2.0	0	9259	1.0	6	168	3.0
	9	0	461386	1.12	6	12	2.0	0	9259	1.0	0	12	3.0
	10	0	461386	1.12	6	12	2.0	0	92590	1.0	6	336	3.0

9.2 No Latent Defects and No Scrubbing

In Chapter 8, Figure 26 compares the number of DDFs estimated from MTTDL to the number estimated by the new model with RAID group size of 14 with a mean time to repair of 48 hours, and shows the estimates are comparable. A similar comparison is shown in Figure 29 for RAID group size of 8 with MTBF = 461,386 and MTTR = 12 hours (see Table 5 for complete set of parameters). The results of the first four comparisons and the MTTDL are listed in Table 6. The plots in Figure 29 use the following notation:

c = constant rate, failure or repair

ft = failure rate is time dependent

mt = restoration rate is time dependent

na = not applicable because the distribution is not used in the analysis

The positions in the string (w-x-y-z) also have significance as follows:

Position "w": Failure rate

Position "x": Restoration rate

Position "y": Latent defect generation rate

Position "z": Scrub rate

In each of the four cases shown in Figure 29, 500,000 simulations were required to generate the results. The plot indicates there is no significant difference between the MTTDL base calculation, Case #1 (c-c-na-na), which assumes the TTOp and TTR are constant failure rates, and Case #4, which includes time dependent failure rates and repair rates. Case #3 stands out as being significantly higher than the base case

(Case #1), indicating the selected restoration distribution tends to increase the number of DDFs. Case #2, shows the time dependent (increasing) failure rate has decreased the number of DDFs expected in 10 years with respect to the base case. Since latent defects are not considered, TTLd and TTScrub are "not applicable".

Table 6 - Results for the First Four Cases and the MTTDL Calculation

Configuration		DDFs/1000 RAID Groups/10 years
Case # 1	c-c-na-na	0.26
Case #2	ft-c-na-na	0.164
Case #3	c-mt-na-na	0.374
Case #4	ft-mt-na-na	0.268
MTTDL	MTTDL	0.27

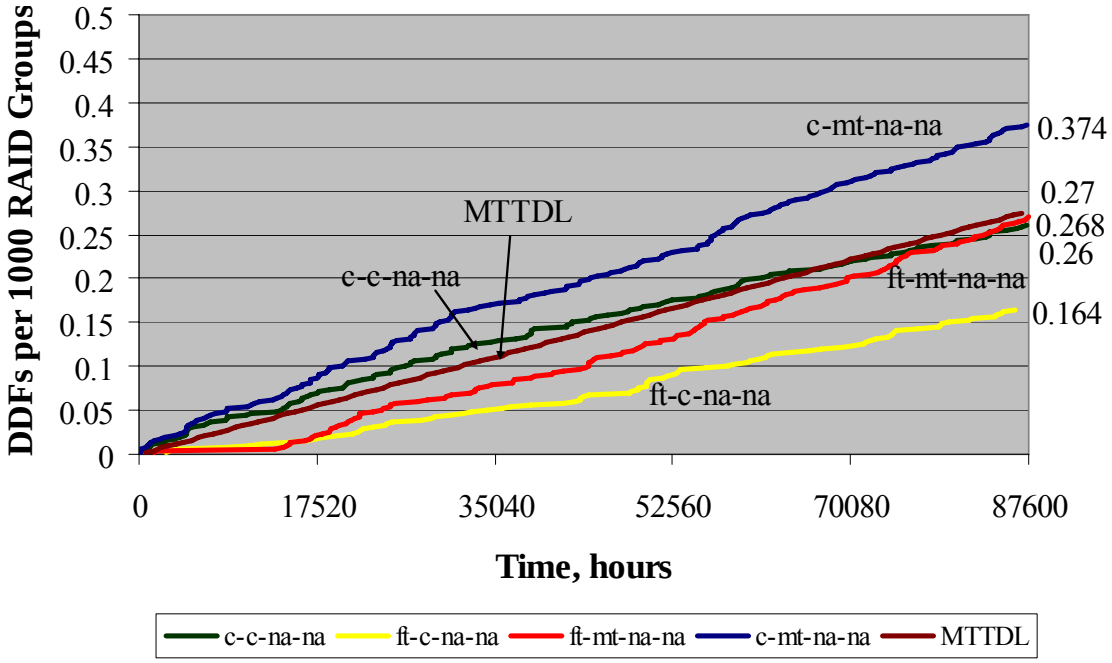


Figure 29 - Basic Comparisons

These five comparisons show the impact of constant failure rates and repair rates to those that are a function of time.

The results of Cases #2 and #3 may seem counter intuitive and deserve some discussion. In comparing Case #1 and Case #2 the characteristic life is the same for both (η = 461,386) but Case #2, with an increasing hazard rate (β = 1.12), slightly shifts the mass of the probability density function to the longer times, decreasing the probability of failure for each HDD for the time period shown in the plot.[6] Changing the restoration distribution from a constant failure rate (Case #1) to the Weibull with parameters show in Table 5, Case #3 (c-mt-na-na), results in an increase in the expected number of DDFs. Using the Weibull parameters for the restoration distribution shifts the mass of the probability function to the longer times, resulting in a longer time that the system is at risk for another failure. The probability density functions for the Exponential and Weibull restoration distributions are shown in Figure 30.

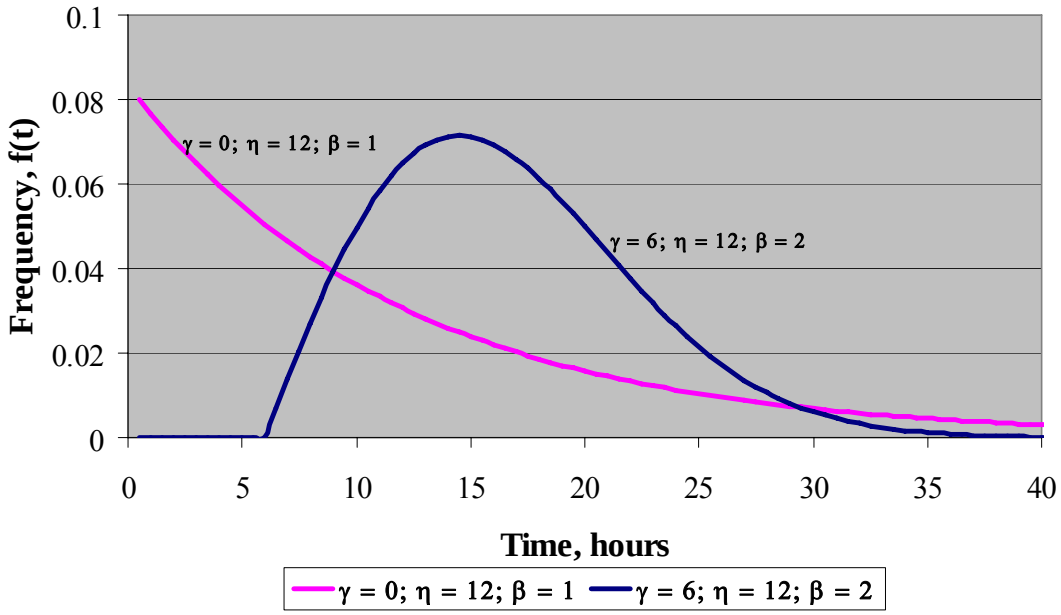


Figure 30 - Shift in pdf: Exponential and 3-parameter Weibull

[6] By changing only the shape parameter, the mass of the p.d.f. shifts. F(Exp:λ, t) = Exp(461386, 87600) = 0.173. F(Weibull:η, β, t) = Weibull(461386, 1.12, 87600) = 0.144.

Figure 31 shows the probability of completing the restoration for the Exponential and the Weibull, as well as the difference between the two. At time t = 15 hours, the probability of restoration based on the Exponential is 0.71 whereas the probability for the Weibull is only 0.43. The Exponential has a greater probability of completing the restoration than the Weibull until 23 hours have elapsed, at which time they are equal (Figure 31).

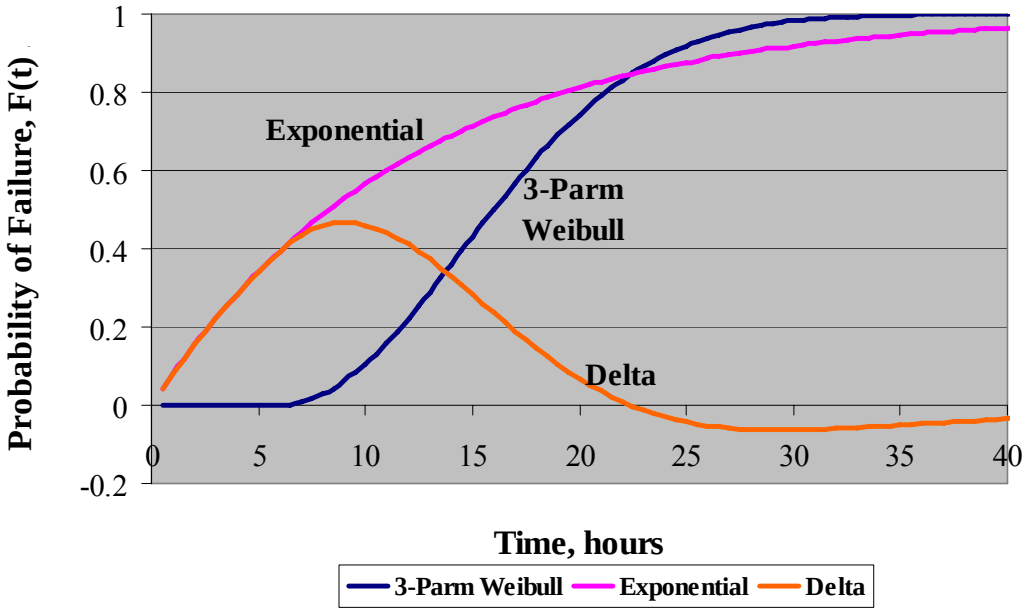


Figure 31 - Cumulative Probability of Restoration as a Function of Time

The probability of completing the restoration is greater for the Exponential than the selected 3-parameter Weibull until 22.5 hours. This generally increases the system ROCOF under the Exponential assumption when all else remains constant. The Delta line is the difference between the Exponential and Weibull probabilities of completion.

When combining the time dependent failure rate with the time dependent restoration (Case #4, line ft-mt-na-na), the expected number of DDFs is greater than for Case #2 and less than Case #3 and, coincidentally, ends up being about the same as Case #1 and the number of DDFs based on the MTTDL calculation.

The results in Figure 29 show only small changes. From a practical viewpoint, one can question whether the model approximations and unavailability of accurate data might render the number of DDFs indistinguishable in a real use of this model. The importance of the results in Figure 29 is the confirmation that the model is sensitive to changes in the failure and restoration distribution assumptions even for small numbers of DDFs. Also notice in Figure 29 that the simulation predicts the number of DDFs to be fairly proportional to time, so the ROCOF is fairly constant.

9.3 Latent Defects without Scrubbing

In Figure 32, latent defects are added to the model. The latent defect distribution is assumed to have a constant rate, so all three distributions used to create the DDF plot in Figure 32 have constant failure rates (Case # 5 in Table 5). The expected number of DDFs increases to over 1430 in 10 years. Again, for these input distributions, the number of DDFs appears fairly proportional in time.

Next, we explore the impact of non-constant failure rates for the operational failures only (TTOp) when the model includes latent defects, but still no scrubbing. Figure 33 shows the effects of using the failure and restoration distributions listed in Table 5 for Case #5 (constant rates) and Case #6 (all constant rates except TTOp).

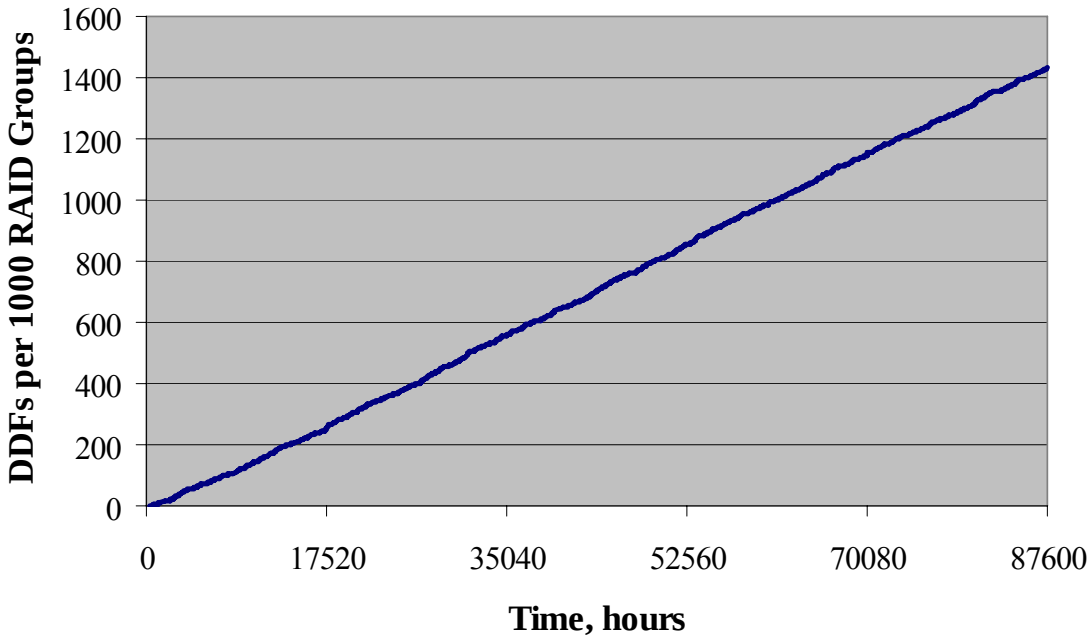


Figure 32 - Case #5: Number of DDFs when Latent Defects are added to Case #1
This continues to assume constant failure and restorations rates.

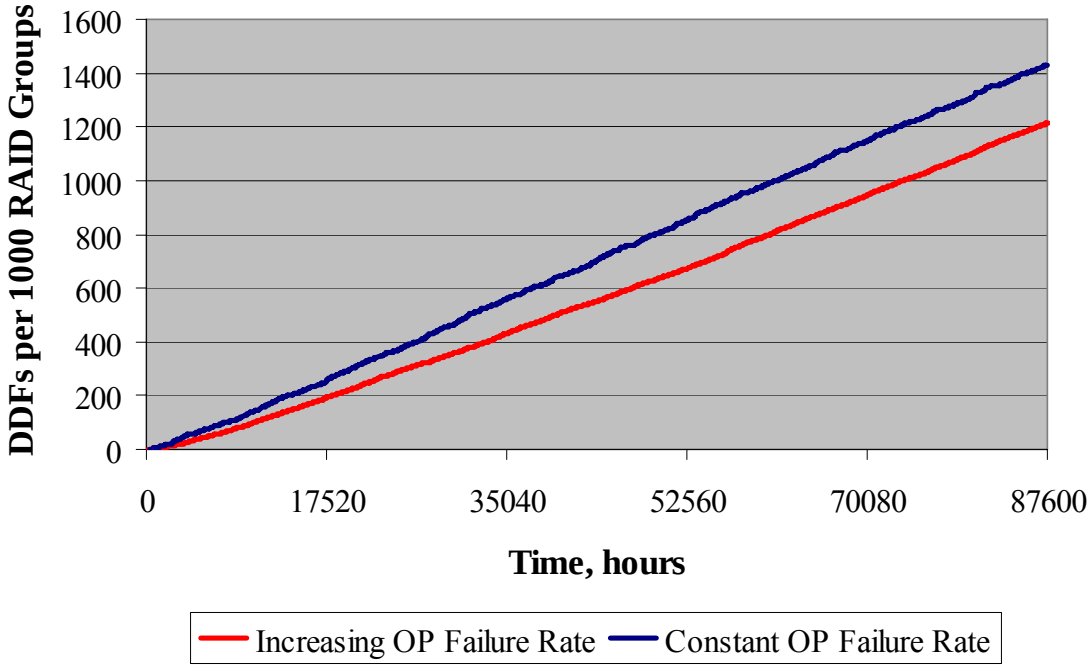


Figure 33 - Case #5 vs. Case #6: Effects of Input Distributions when Latent Defects are not Scrubbed

This is the first example in which the effects of the non-constant failure rates can be easily observed. Figure 34 shows the ROCOF as a function of time. An interval of 1,460 hours was selected. The number of DDFs in each of the intervals was used to calculate dN/dt. Note that Case #5, which has all constant failure rates as inputs, exhibits a fairly constant ROCOF for the system. Case #6, however, indicates an increasing failure rate, indicating a NHPP prevails.

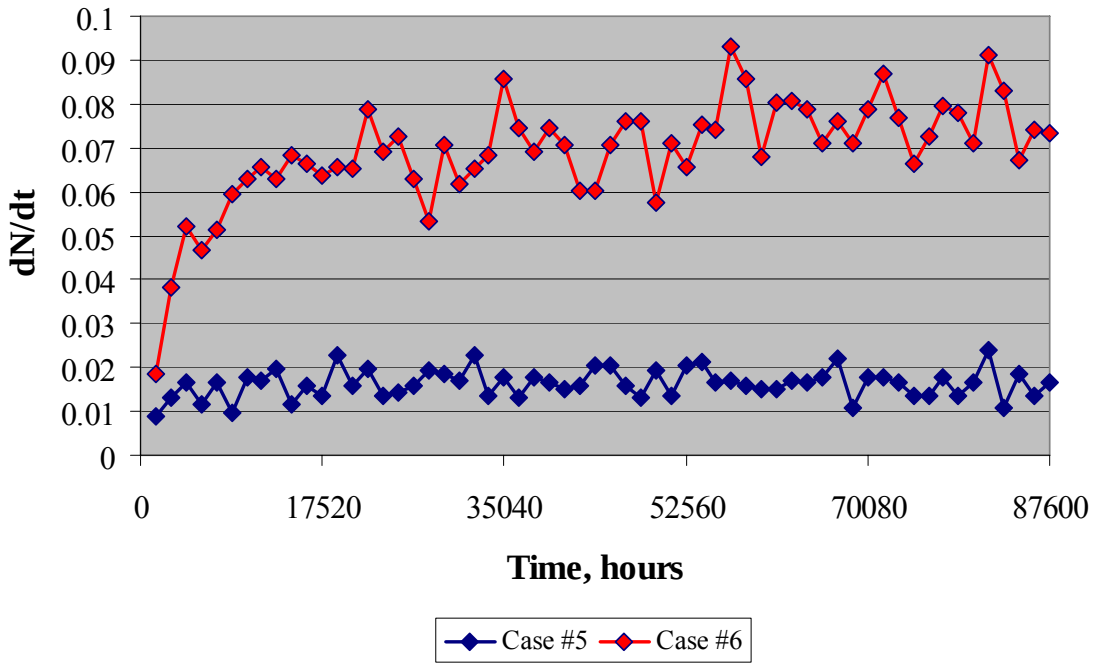


Figure 34 - Rate of Occurrence of Failure (ROCOF) for Cases #5 and #6

9.4 Latent Defects with Scrubbing

Some RAID manufacturers, such as Network Appliance, have added scrubbing as an operational feature. Figure 35 compares Case #6, which has latent defects without scrubbing, to Case #7, which has the same distributional parameters as Case #6, but includes a complete HDD scrub for latent defects every 336 hours with a

minimum of 6 hours after the defect is introduced. This scrub reduces the number of DDFs from over 1,200 to just over 280 in 10 years.

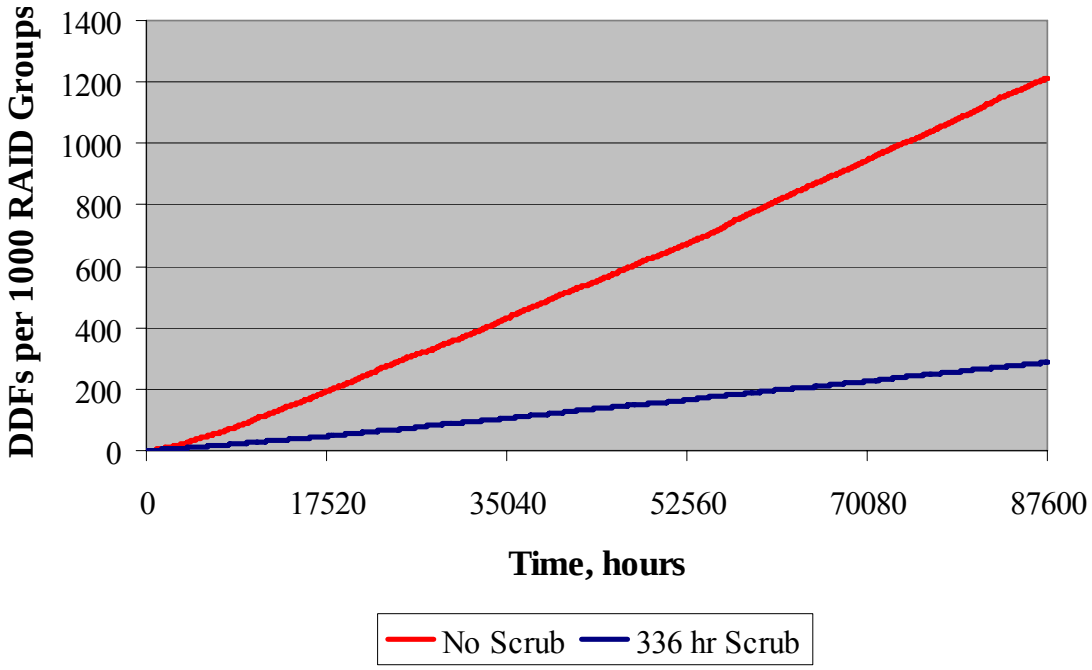


Figure 35 - Case #6 vs. Case #7: Benefits of 336 hour Scrub

Figure 36 shows the effects of three different scrub schemes. The 336 hr and 168 hr time to scrub curves assume scrubbing cannot be completed in less than 6 hours. The "12 hour scrub" plot assumes there is no minimum time (location parameter $\gamma = 0$), the distribution for time to complete the scrub has a 12 hour characteristic life ($\eta = 12$ hours) and, as in both of the others, the shape is 3.0 ($\beta = 3.0$).

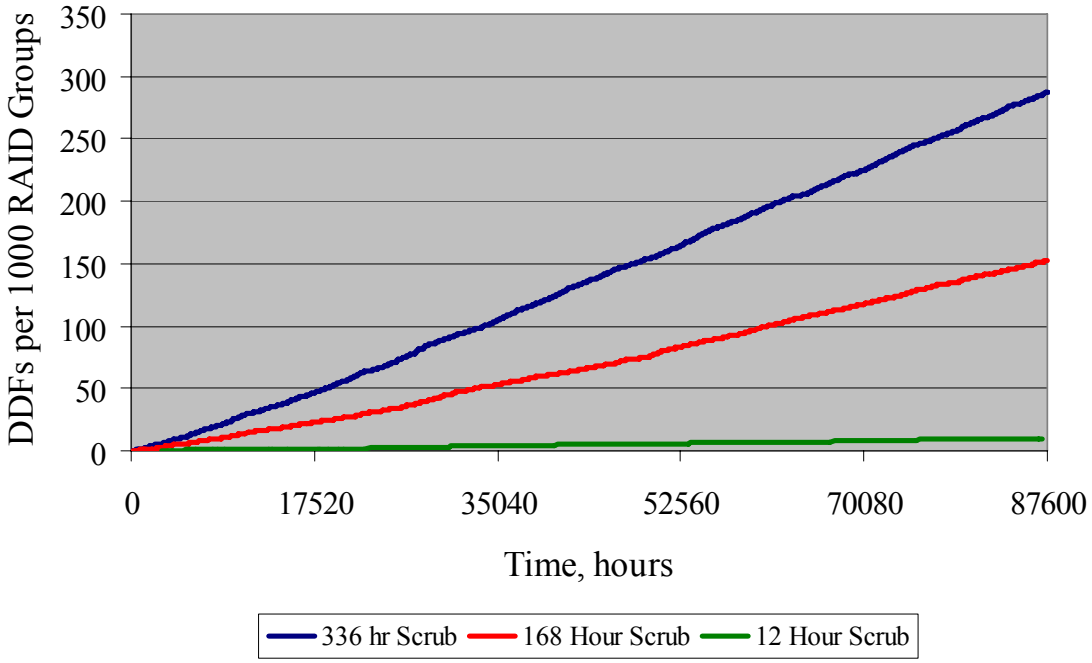


Figure 36 - Case #7; Case #8 & Case #9: Effects of Scrub Times

The next comparison explores the effects of the time to latent defect (TTLd). In all the cases so far, the characteristic life for the TTLd has been 9,259 hours, per the calculation shown in Table 3. However, if operational procedures or HDD manufacturers can reduce the rate of occurrence by a factor of 10, the number of DDFs also decreases. Figure 37 shows plots comparing Case #7 to Case #10. The only difference between these two cases is an order of magnitude improvement in the TTLd for Case #10 (92,590 hours) over Case #7 (9,259 hours). Although somewhat difficult to see in these plots, a dN/dt plot (Figure 38) shows the TTLd with the shorter characteristic life (9,259) results in an increasing ROCOF (NHPP). For the longer characteristic life (92,590) the ROCOF appears very constant.

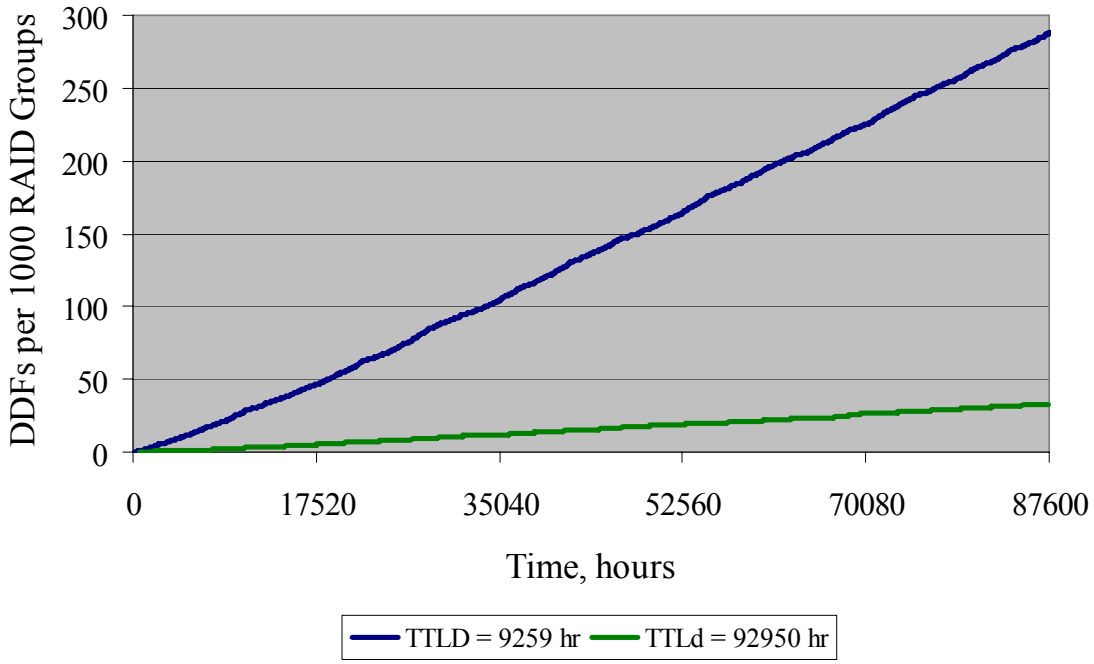


Figure 37 - Effects of Time to Latent Defect Distribution

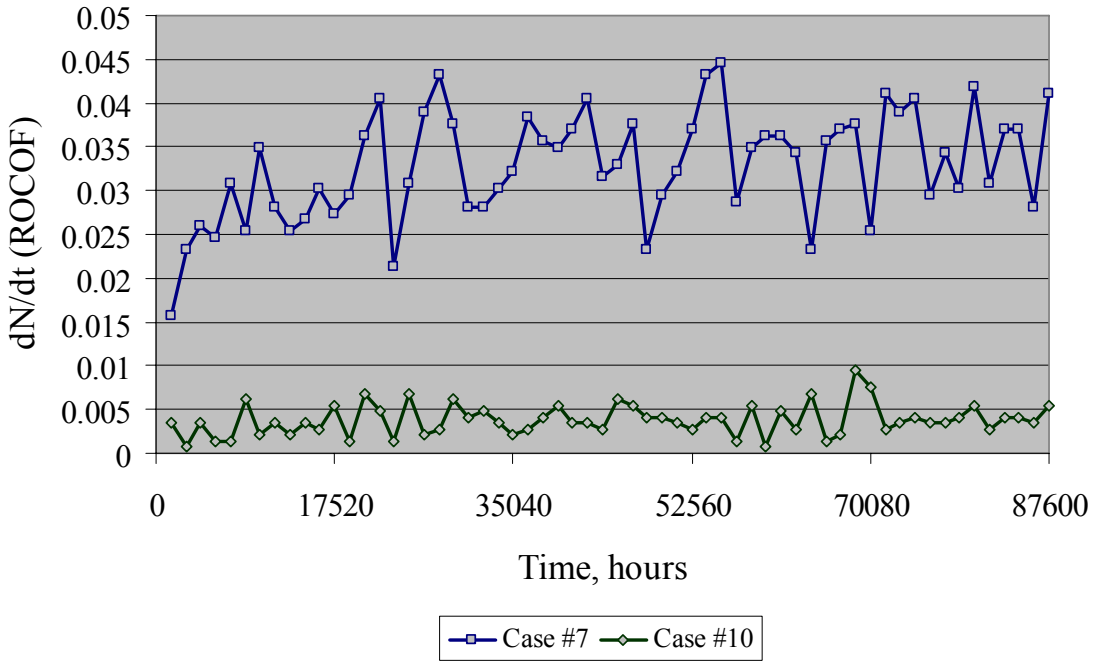


Figure 38 - ROCOF for Case #7 and #10

Case #7 shows the ROCOF more than doubles over the 10 years simulated, starting at 0.15 and ending around 0.35 (on average)

9.5 RAID Group Size Effects

Basic reliability theory indicates that as the quantity of HDDs in a RAID group increases, the number of DDFs should also increase for any fixed set of distribution parameters. Figure 39 through Figure 42 show the reduction in DDFs as a function of RAID group size for four sets of parameters shown in Table 7. The quantity of expected DDFs is, to some degree, dependent on the number of combinations of HDDs in a RAID group when taken 2 at a time, so one should expect that the number of DDFs is greater in a RAID group of 14 than in a smaller group. That is, the larger the RAID group, there are more combinations that can result in 2 simultaneous failures.

Table 7 - Parameters for RAID Group Size Studies

	Operational						Latent					
	TTOp			TTR			TTLd			TTScrub		
Case	γ	η	β	γ	η	β	γ	η	β	γ	η	β
11	0	461386	1.12	6	12	2.0	0	9259	1.0	6	336	3.0
12	0	461386	1.12	6	12	2.0	0	9259	1.0	3	12	3.0
13	0	461386	1.12	6	12	2.0	0	92590	1.0	6	336	3.0
14	0	461386	1.12	6	12	2.0	0	92590	1.0	3	12	3.0

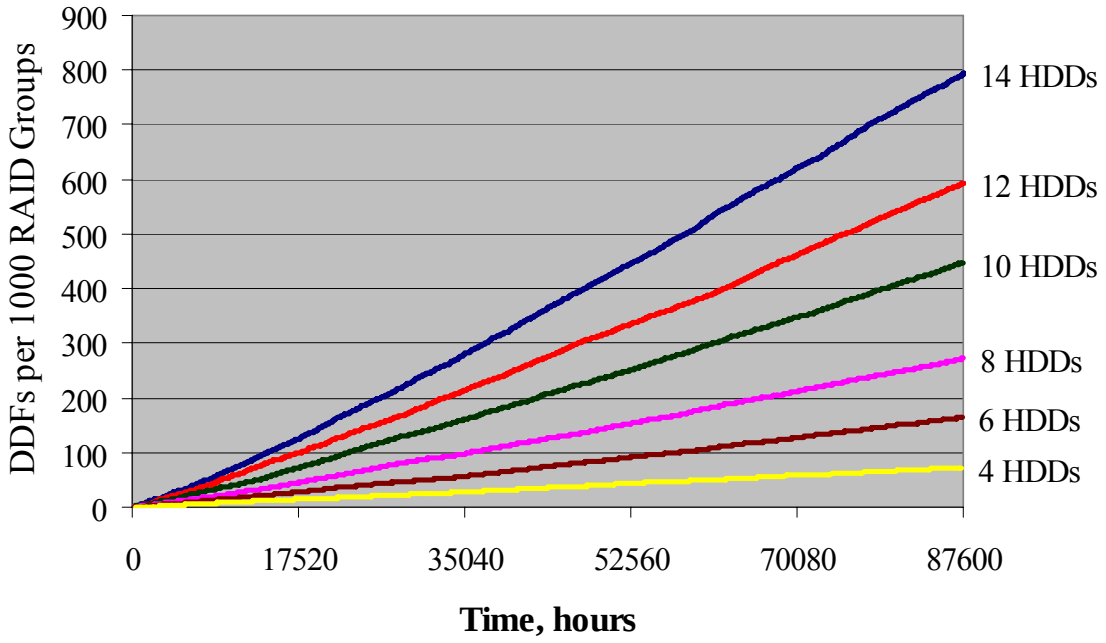


Figure 39 - RAID Group Effects for Case #11: 336 hr TTScrub & 9259 hr TTLd

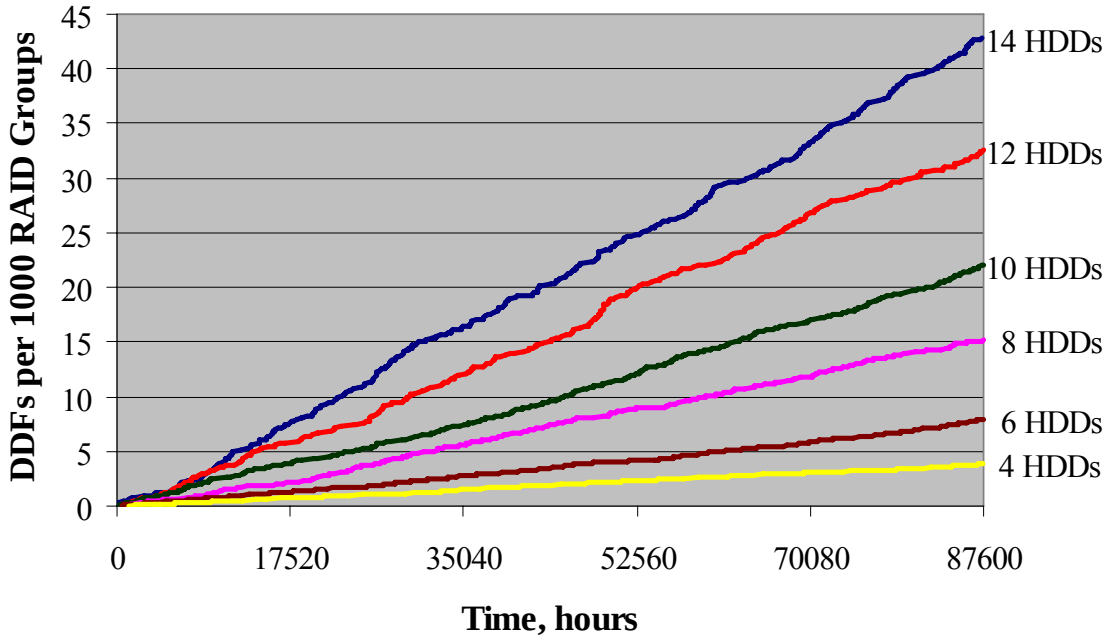


Figure 40 - RAID Group Effects for Case #12: 12 hr TTScrub & 9259 hr TTLd

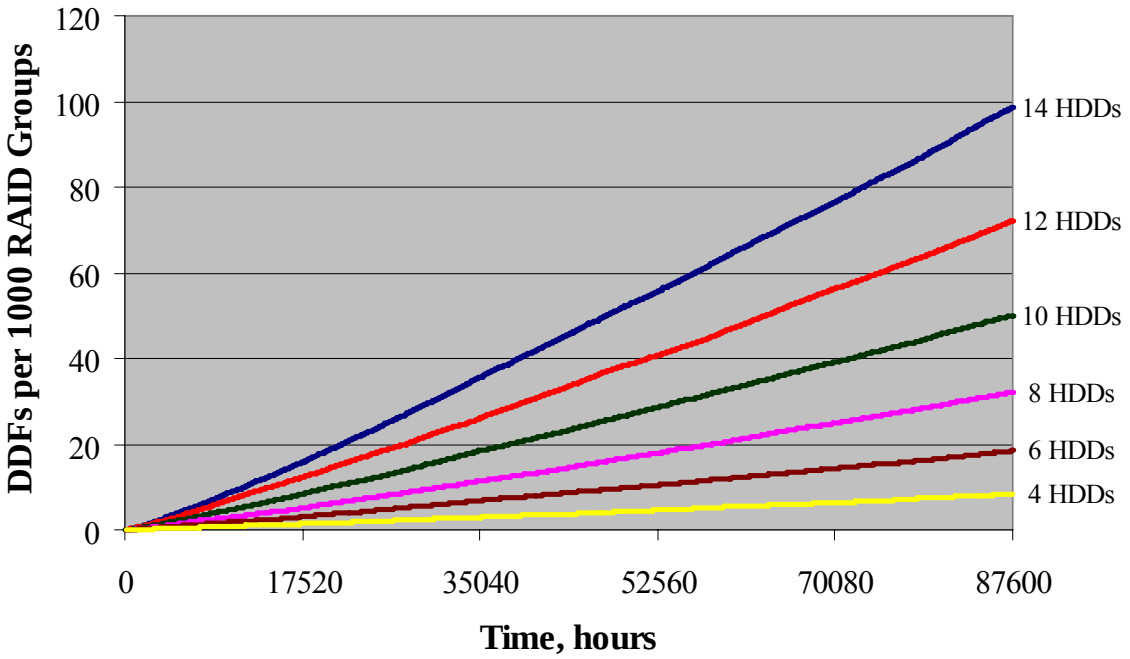


Figure 41 - RAID Group Effects for Case #13: 336 hr TTScrub & 92590 hr TTLd

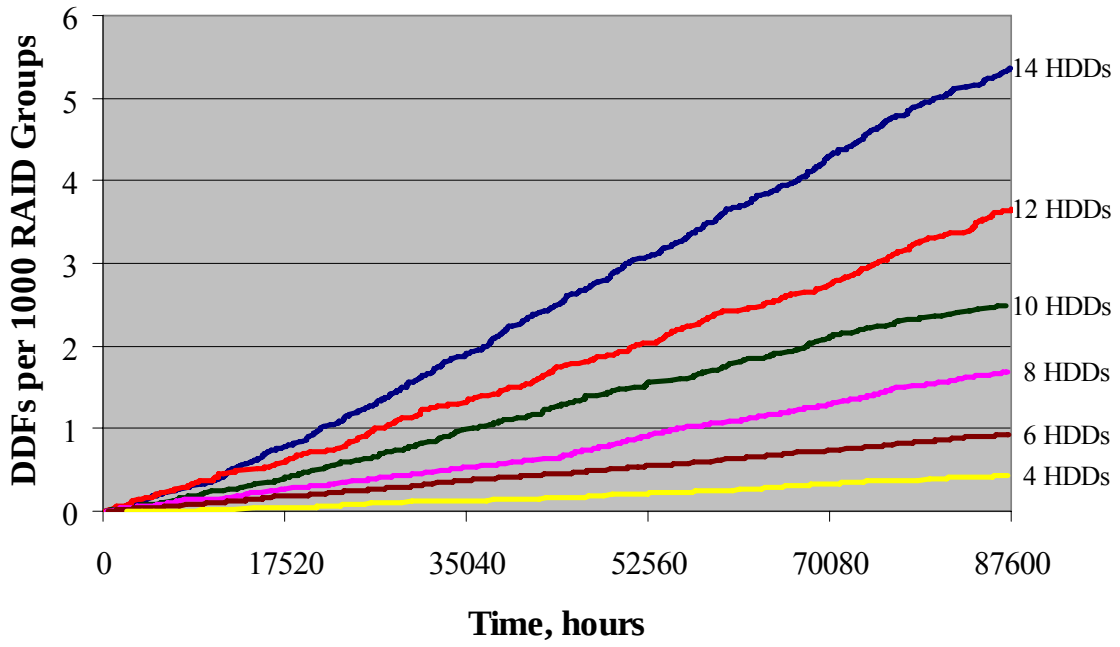


Figure 42 - RAID Group Effects for Case #14: 12 hr TTScrub & 92590 hr TTLd

The number of combinations is governed by Binomial coefficients using the equation

$$ {}_nC_r = \frac{n!}{(n-r)!r!} \qquad \textbf{eq. 19} $$

where n = number of combinations of interest and r = the number in the group. For *N+1* RAID, *n* = 2 and *r* = number of HDDs in the RAID group. It has already been shown that on an absolute basis, the number of DDFs is highly dependent on the governing distributions and the inclusion of latent defects, so this analysis examines estimation of the number of DDFs based on the binomial coefficients and the number calculated from a RAID group that is a different size. That is, if the number of DDFs for a RAID group of, say, 14 is calculated, is there a "quick and dirty" way to determine the number that would occur in a RAID group of, say, 8 if all the distributions are exactly the same as in the RAID group of 14? Table 8 is used extensively in this discussion.

First, the number of combinations (binomial coefficients) is determined for RAID group sizes 14, 12, 10, 8, 6, and 4. Then the ratios for all combinations are calculated. The "amount of separation" must be considered in the calculating the ratios.[7] The ratios for all separations are shown in the lower section of Table 8. For each separation, the ratio of the binomial combinations is calculated under the sub-heading "Ratios" in the third column of the table (Binomial Coefficients). Similarly,

[7] Separation is the relative difference in quantities of two RAID groups. A group of 14 is "2 apart" from a group of 12. A group of 10 is "4 apart" from a group of 6.

the ratios are calculated for all amounts of separation for Cases #11 through #14, which are in columns 4-7.

Table 8 - Binomial Ratios and DDF Ratios for Various Group Sizes

	RAID Group Size	Binomial Coefficien	DDFs Case #11	Case #12	Case #13	Case #14
	14	91	792	46	98.5	5.4
	12	66	585	33	72	3.7
	10	45	433	20	50	2.5
	8	28	283	12.6	32	1.7
	6	15	157	7	18.5	1
	4	6	74	3.5	8	0.43
	Ratios					
2 apart	14:12	1.38	1.35	1.39	1.37	1.46
4 apart	14:10	2.02	1.83	2.30	1.97	2.16
6 apart	14:8	3.25	2.80	3.65	3.08	3.18
8 apart	14:6	6.07	5.04	6.57	5.32	5.40
10 apart	14:4	15.17	10.70	13.14	12.31	12.56
2 apart	12:10	1.47	1.35	1.65	1.44	1.48
4 apart	12:8	2.36	2.07	2.62	2.25	2.18
6 apart	12:6	4.40	3.73	4.71	3.89	3.70
8 apart	12:4	11.00	7.91	9.43	9.00	8.60
2 apart	10:8	1.61	1.53	1.59	1.56	1.47
4 apart	10:6	3.00	2.76	2.86	2.70	2.50
6 apart	10:4	7.50	5.85	5.71	6.25	5.81
2 apart	8:6	1.87	1.80	1.80	1.73	1.70
4 apart	8:4	4.67	3.82	3.60	4.00	3.95
2 apart	6:4	2.50	2.12	2.00	2.31	2.33

There appears to be a strong influence of the possible combinations on the expected number of DDFs using the model.[8] This is determined by looking at the ratio for a separation of 2 (14:12) for both the binomial based and the model based.

[8] There is obviously a relationship between the binomial coefficients and the expected number of DDFs if the process is HPP. This analysis examines the results to see if that relationship still holds true in this model.

The binomial ratio for 14:12 is 91/66 = 1.38. For Case #11, the ratio is 792/585 = 1.35.

Notice that, regardless of the case number and regardless of the absolute number of HDDs in the RAID group, there is a fairly consistent relationship within any separation and across all cases. Continuing this example, the ratios for Cases #11 through #14 are 1.35, 1.39, 1.37 and 1.46. Remember that the absolute number of DDFs ranges from 792 down to 5.4 for all the Cases with raid group size of 14. It appears that if the model based estimate of DDFs is simulated for a specific RAID group size, a rough estimate of the number of DDFs for any other RAID group size can be calculated using the ratio derived from the binomial coefficients.

Consider this example. Suppose the simulation of a RAID group of 10 estimated that 20 DDFs will occur in the 10 year life. For the same distribution parameters, the number that may be calculated for a RAID group of 6 is 6.67, calculated as follows:

- The binomial base ratio for the ratio of 10:6 is 3.00
- The number of DDFs estimated from the simulation for RG=10 is 20
- 20 / 3 = 6.67
- The estimate for a RAID group of 6 is 7, a fairly close estimate

A perusal of the ratios for any amount of separation from the simulation is fairly consistent with that calculated from the binomial coefficients for the same separation and the same maximum. This analysis is empirical and has predictable errors. However, this clearly shows that even in this new simulation model, the effect of the number of HDDs has a rather predictable impact on the number of expected DDFs.

9.6 Effect of Shape Parameter

This last sensitivity study (Figure 43) examines the effects that the shape parameter, β, for the operational failure distribution (TTOp) has on the number of DDFs. Case #8 from Table 5 is the basis for the study. Again, the somewhat counter-intuitive results are produced in that the number of DDFs is greater when the hazard rate is decreasing than when it is constant or increasing.

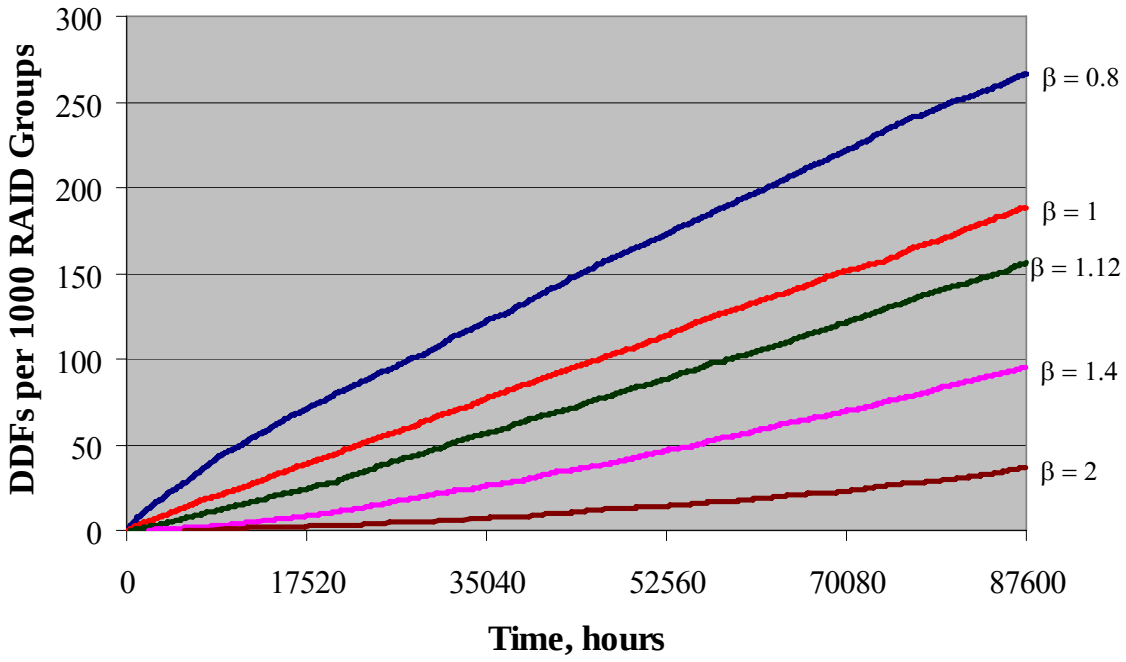


Figure 43 - Shape Parameter Effects

This plot shows the effects of the shape parameter, β, on the number of DDFs. All parameters are the same as for Case #8 in Table 5 *except the beta for the TTOp failure distribution.*

However, the rate of change, *dN/dt*, or ROCOF is steadily decreasing for β = 0.8, and appears to be continuously increasing ROCOF when β = 2.0. These results, as well as other plots that show non-linearity, indicate the ROCOF is not a HPP. Clearly,

the hazard rate of the individual HDDs impacts the ROCOF and should be properly accounted for in a system analysis. In this study, a simplistic assumption of $\beta = 1$ produced DDF estimates 121% higher than if the true beta is 1.12, and estimates only 70% of the DDFs if the $\beta = 0.8$ while the characteristic life remains unchanged. The ROCOF for the cases when $\beta = 0.8$ and $\beta = 2.0$ are shown in Figure 44.

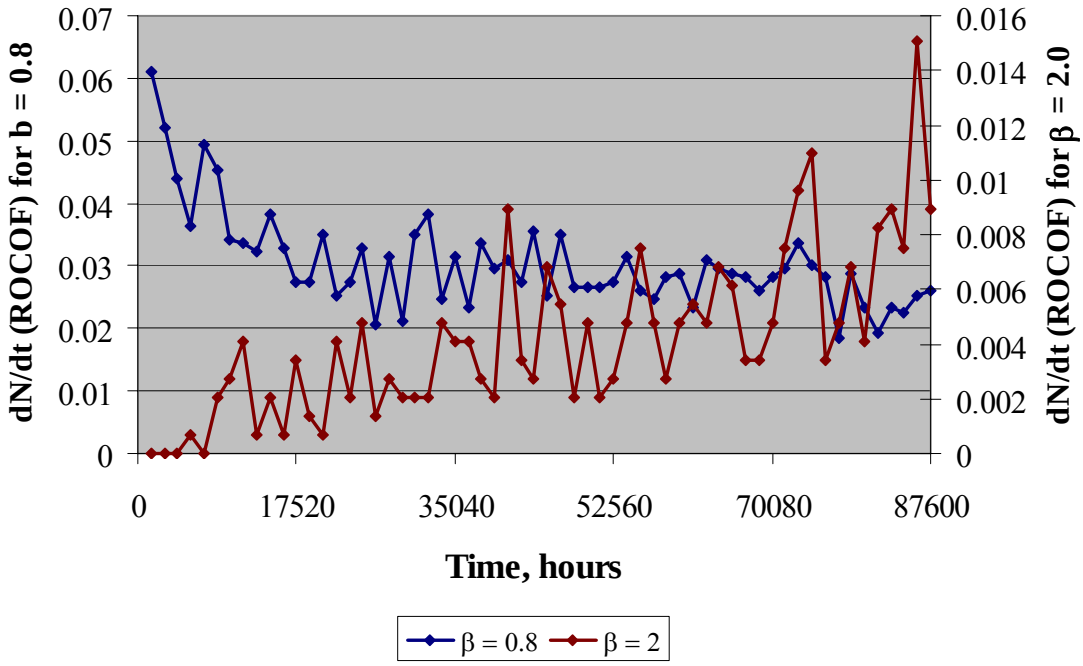


Figure 44 - ROCOF for Case #8 with Decreasing and Increasing Failure Rates

9.7 Comparison to Field Data

The last two studies compare predicted number of DDFs to the field data for Network Appliance systems. Whereas the simulations represented between 10,000 and 500,000 systems operating for 10 years, the subpopulations from Network Appliance field are on the order of 10,000 RAID groups for 18 months. Two sets of charts are presented based on the HDD's data bus type, ATA and fibre-channel (FC).

Section 9.7.1 presents comparison for 10k rpm HDDs with the FC interface. Section 9.7.2 presents comparisons for HDDs with the ATA interface. For both the FC and ATA data sets, the RAID groups were operating in RAID-6[9] mode *(N+2)*. During reconstruction for an operational failure (TTOp), the times to discovery of a latent defect were recorded. These are equivalent to DDFs if the system were operated as RAID-4 (or RAID-5) using *N+1* redundancy. The exact scrubbing algorithm used by NetApp is proprietary and confidential, and cannot be exposed here. However, scrubbing cannot be completed in less than 3 hours due to HDD capacity and data-bus bandwidth, and will be completed in less than 1 week. For bounding purposes, scrubs of 12 and 168 hours are used with a 48 hour scrub as an intermediate modeling point.

9.7.1 Fibre-Channel HDDs

Data collected for FC studies includes all HDD manufacturers, all models and all vintages combined. The actual TTOp and TTLd are quantity weighted combinations of the individual distributions shown in Chapter 5. Field data show that RAID groups size 14 and 16 have about the same number of DDFs per 1,000 (Figure 45). This is a consequence of statistical variability in the data. If one considers only the MCF up through 13,140 hours, it appears that the number of DDFs for RAID group size 16 is greater than that at of size 14, as expected. The number of size 14 RAID groups that have been in the field longer than 13,140 hours diminishes to about 800, significantly smaller than the 4,600 that have been in the field for the shorter durations (Figure 46). RAID groups of size 16 diminish from over 10,000 for the shorter field lives, to fewer than 1,000 that have been in the field for over 13,140 hours (Figure 47). Only 3 DDFs

[9] Because these RAID groups were run in RAID-6 mode, systems did not fail, but DDFs were recorded

occur after 13,140 hours size 14 and 2 for size 16. The paucity of failures after 13,140 brings into question any measurement or statistic that uses data after 13,140 hours.

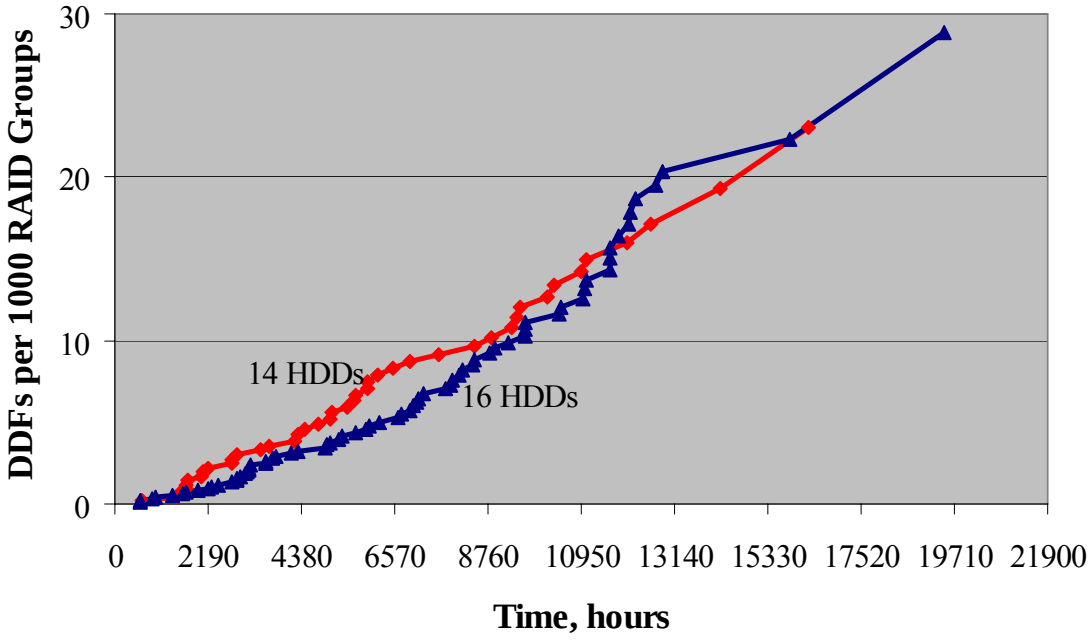


Figure 45 - DDF Field Data for 10k RPM, FC HDDs of Group Sizes 14 and 16

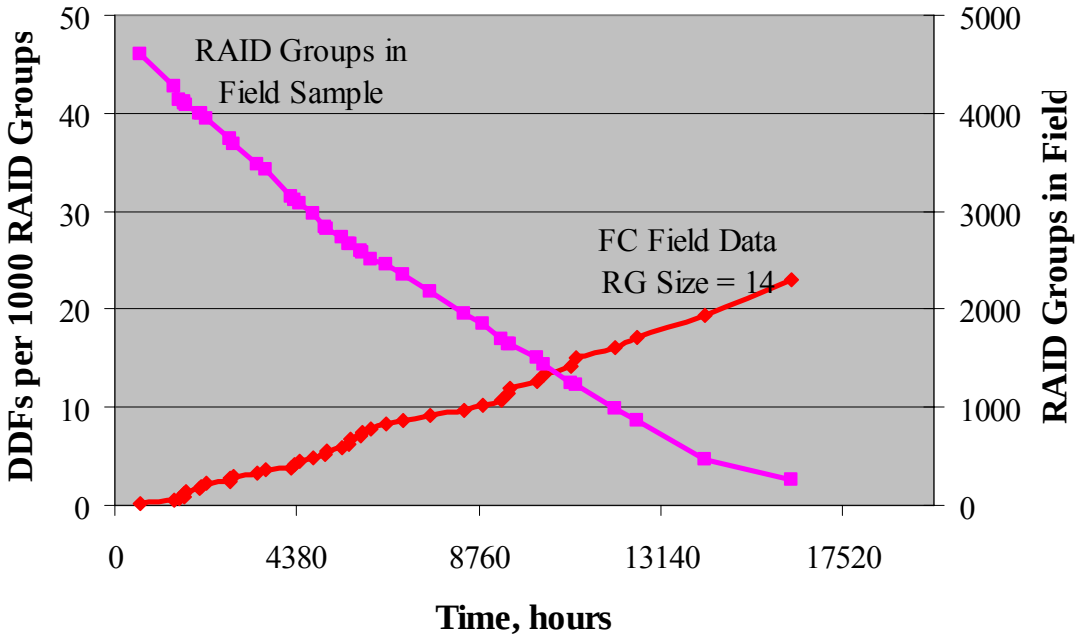


Figure 46 - FC RAID Group Size 14: Number of RAID Groups versus DDFs

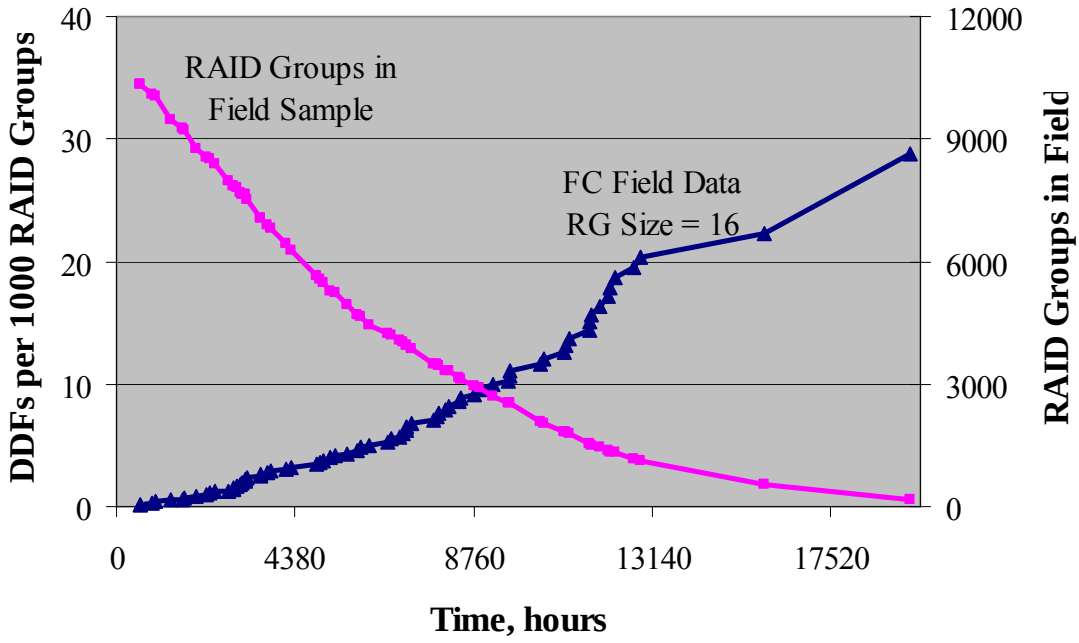


Figure 47 - FC RAID Group Size 16: Number of RAID Groups versus DDFs

The ROCOFs for both RAID group size 14 and 16 show stability up to about 13,140 hours. After that time, the lack of data causes significant fluctuations in the ROCOFs (Figure 48 and Figure 49). The conditions surrounding Network Appliance field data are most similar to Case # 8 or Case #9, so they are used for comparison to the measured field data. If a TTScrub of 48 hours is used, the model and the field data are essentially the same for RAID group size 14 and extremely close for size 16 (Figure 50 and Figure 51).

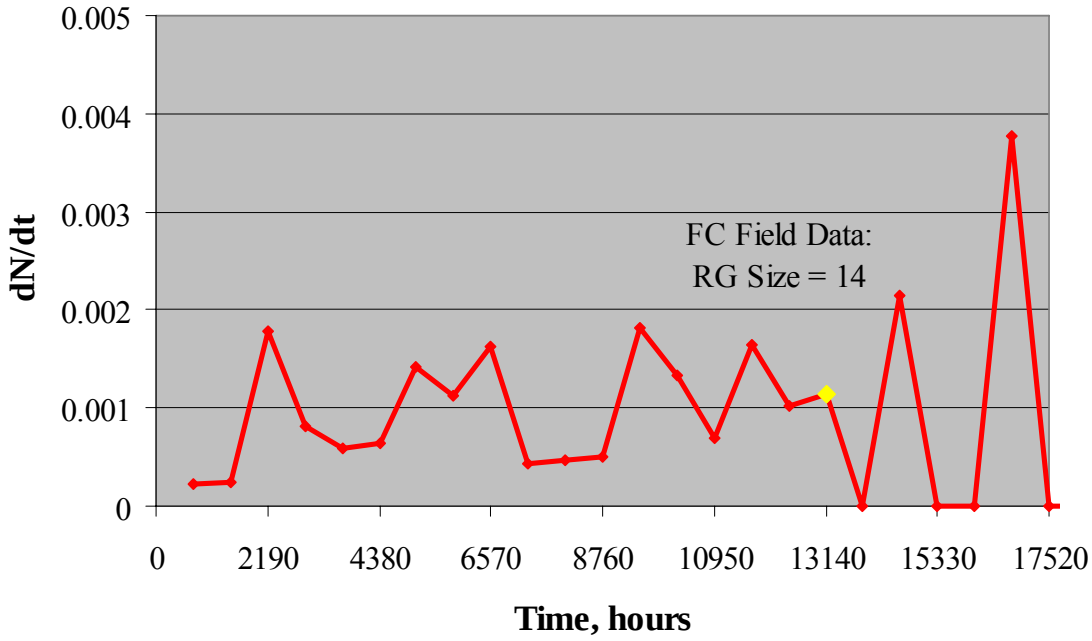


Figure 48 - ROCOF for FC RAID Group Size 14

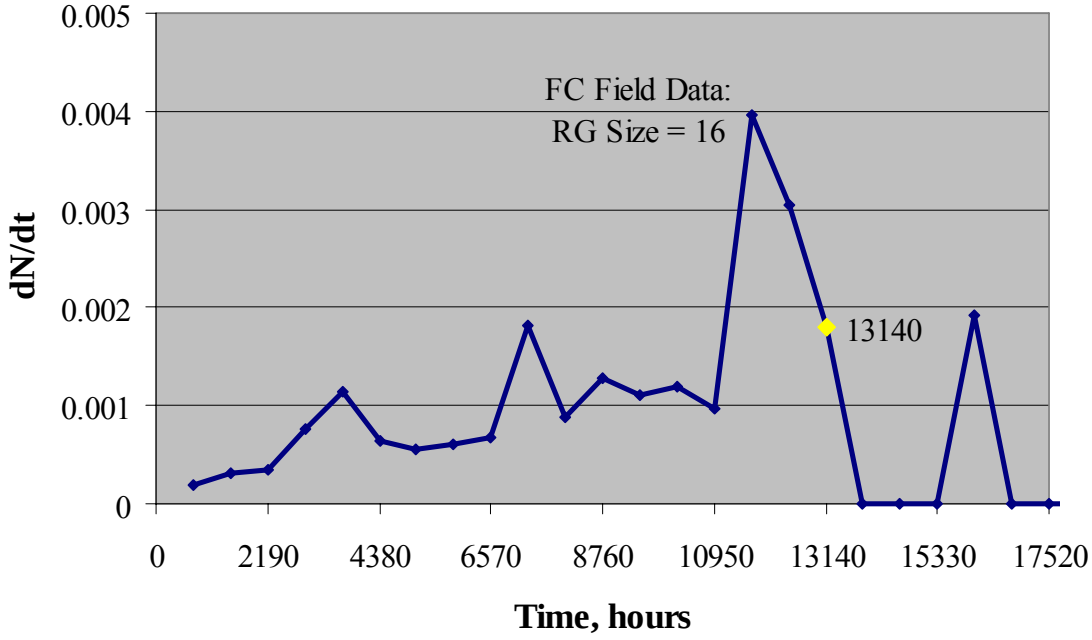


Figure 49 - ROCOF for FC RAID Group Size 16

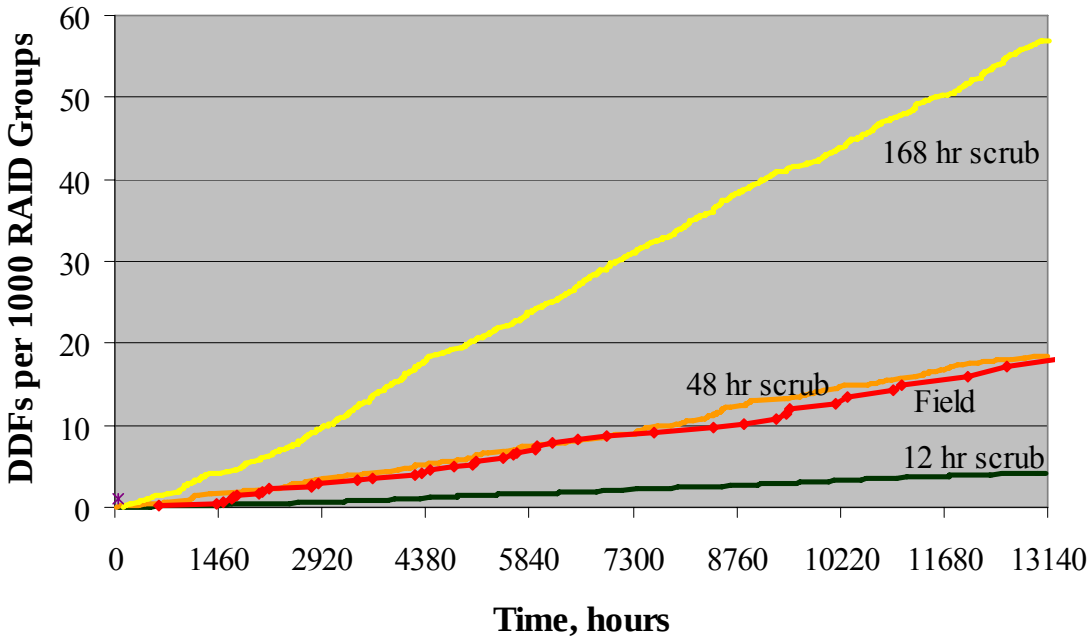


Figure 50 - Fibre Channel 10k rpm, RAID Group Size 14: Model versus Field

The number of DDFs from NetApp field experience is in between the two modeled cases. Since the exact conditions for the field population are not known and knowing how sensitive the model is to the input parameters, these bounds are interpreted as extremely good correlation between the predicted model and the field results.

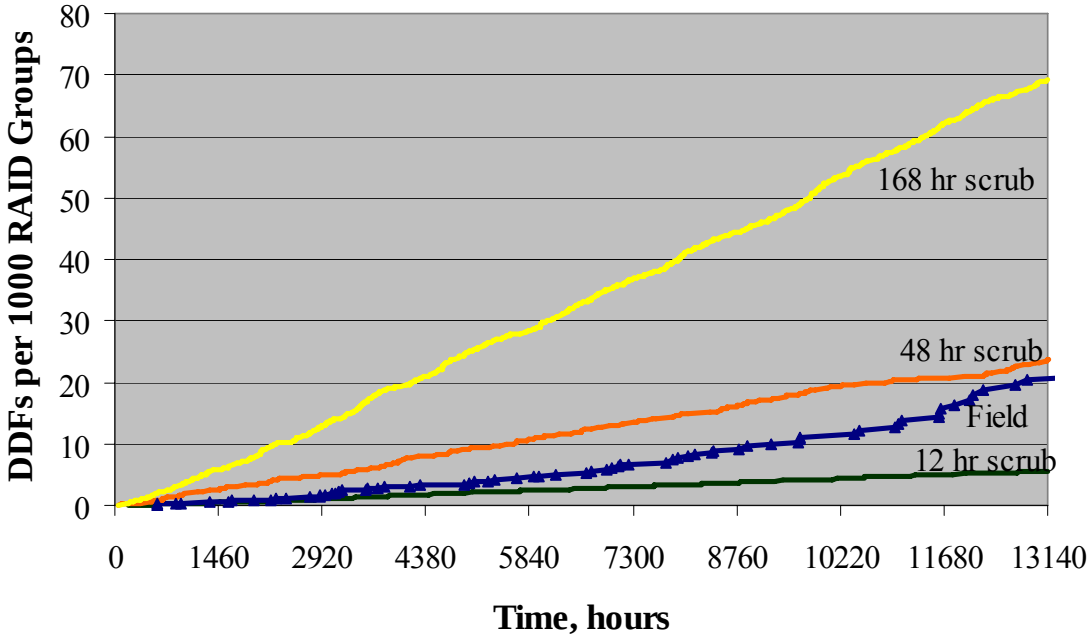


Figure 51 - Fibre Channel 10k rpm, RAID Group Size 16: Model versus Field

As in Figure 50, the number of DDFs from NetApp field experience falls near the 48 hour TTScrub assumption. The exact conditions for the field population are not known but these plots must be interpreted as extremely good correlation between the predicted model and the field results.

9.7.2 ATA Interface HDDs

The population of ATA HDDs used in this analysis contains several capacities, many vintages and is grouped based on RAID group size. Subpopulations of size 14 and 16 were selected based on overall quantity of data available. The ROCOF for both groups show an increasing trend starting at initial turn-on in the field (Figure 52). This RAID group behavior reflects that of the individual HDDs, which demonstrated a high degree of "wear out" after 10,000 hours, similar to that in Figure 12 in Chapter 5. For both size 14 and 16, the number of RAID groups in the field diminishes rapidly in time (Figure 53 and Figure 54). Nearly 6,000 RAID groups size 16 operated for a short duration, but that number decreases to about 2,000 at 13,870 hours. For RAID group size 14, the quantity decreased from over 8,000 to about 1,000 by the same 13,870 hours. The MCF calculation accounts for the change in system quantity as described in [54]. For both RAID group sizes the ROCOF (dN/dt) increases from time zero, but becomes erratic at 13,870 hours due to the small number of RAID group failures. Plots of dN/dt (Figure 55 and Figure 56) quantify the increase in ROCOF that is apparent by observation of Figure 52.

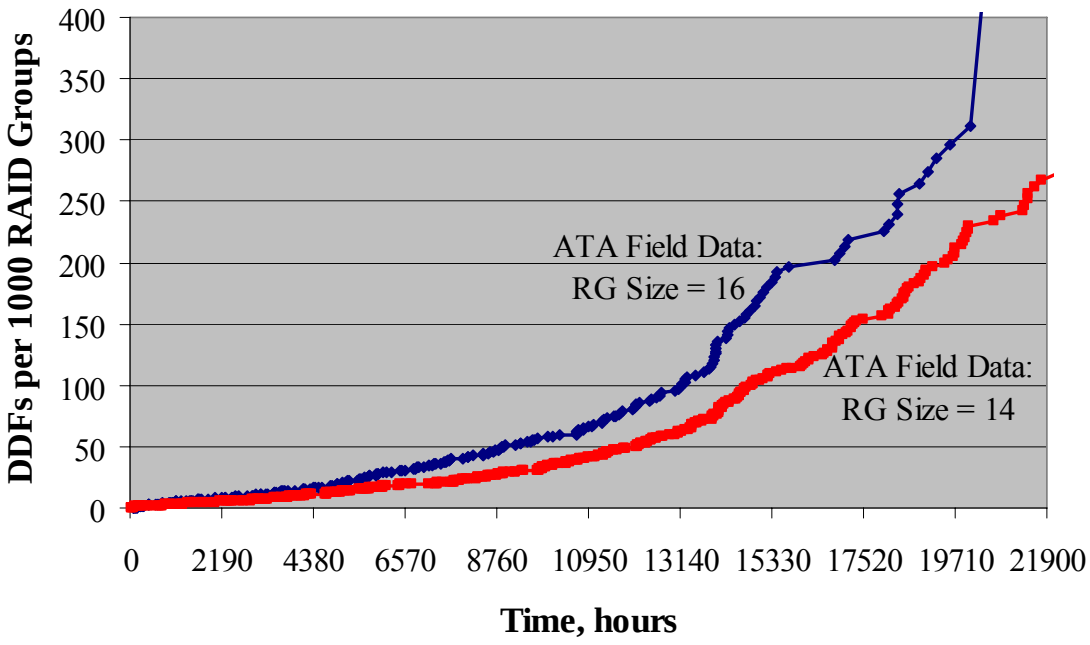


Figure 52 - DDFs for all ATA HDDs: All Manufacturers, Families, Capacities and Vintages

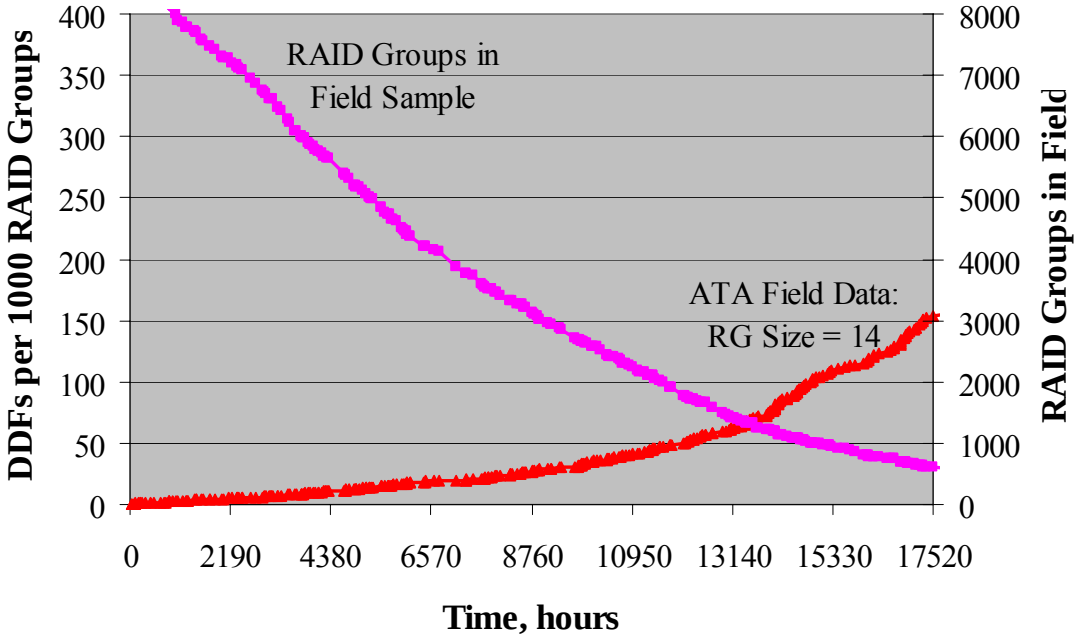


Figure 53 - ATA RAID Group Size 14: Number of Groups versus DDFs

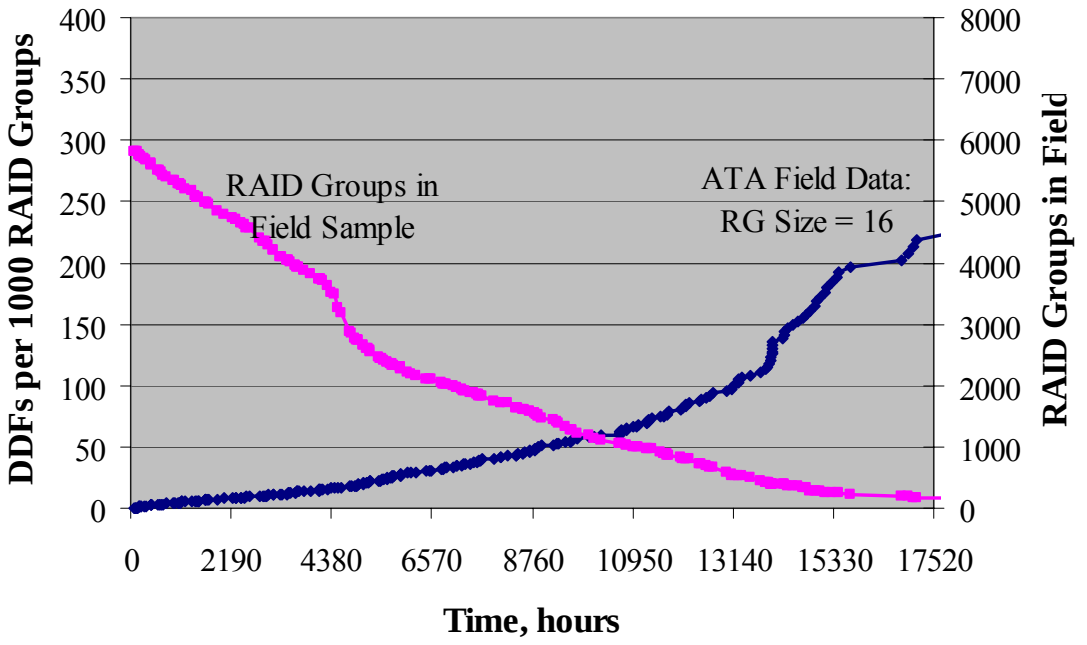


Figure 54 - ATA RAID Group Size 16: Number of Groups versus DDFs

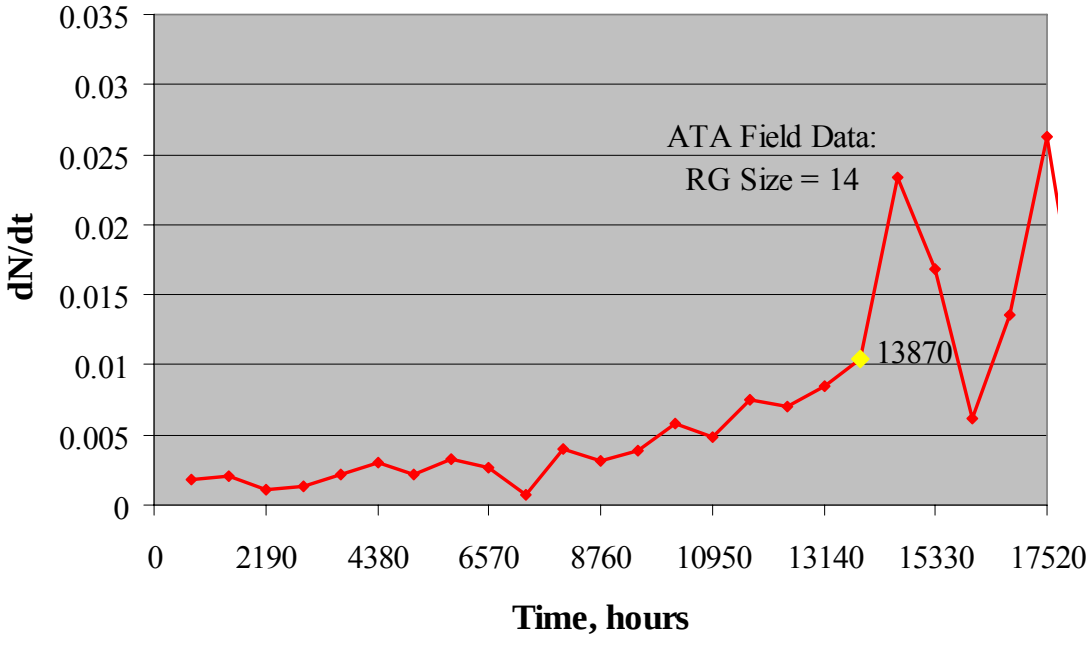


Figure 55 - ROCOF for ATA RAID Group Size 14

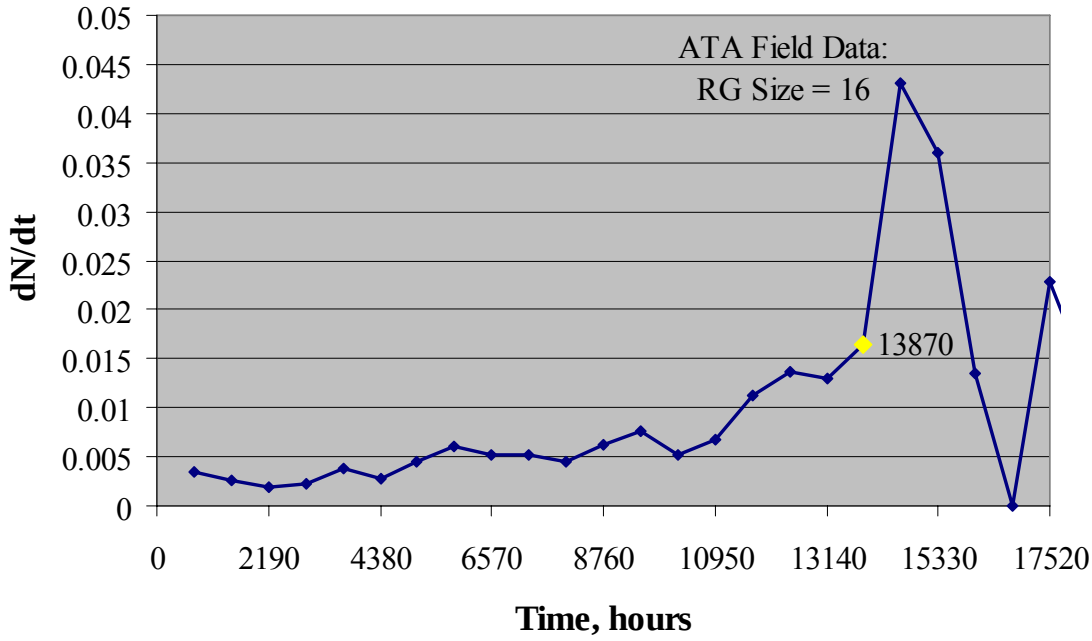


Figure 56 - ROCOF for ATA RAID Group Size 16

As with the fibre-channel HDDs, the scrub algorithm for the ATA HDDs in Network Appliance systems is proprietary, but is in the range of 12 to 168 hours. Thus, exact correlations between the model and the field data are not available. However, Figure 57 and Figure 58 show the accuracy of the prediction using the new model as compared to the field data. When a 48 hour time to scrub is assumed, the model tracks the field results very well for both RAID group sizes. The three model lines use the parameters shown in Table 9 and change only the TTScrub.

Table 9 - Parameters for Field Comparison

Operational						Latent					
TTOp			TTR			TTLd			TTScrub		
γ	η	β	γ	η	β	γ	η	β	γ	η	β
0	75000	1.49	6	12	2	0	9259	1	3	48	3

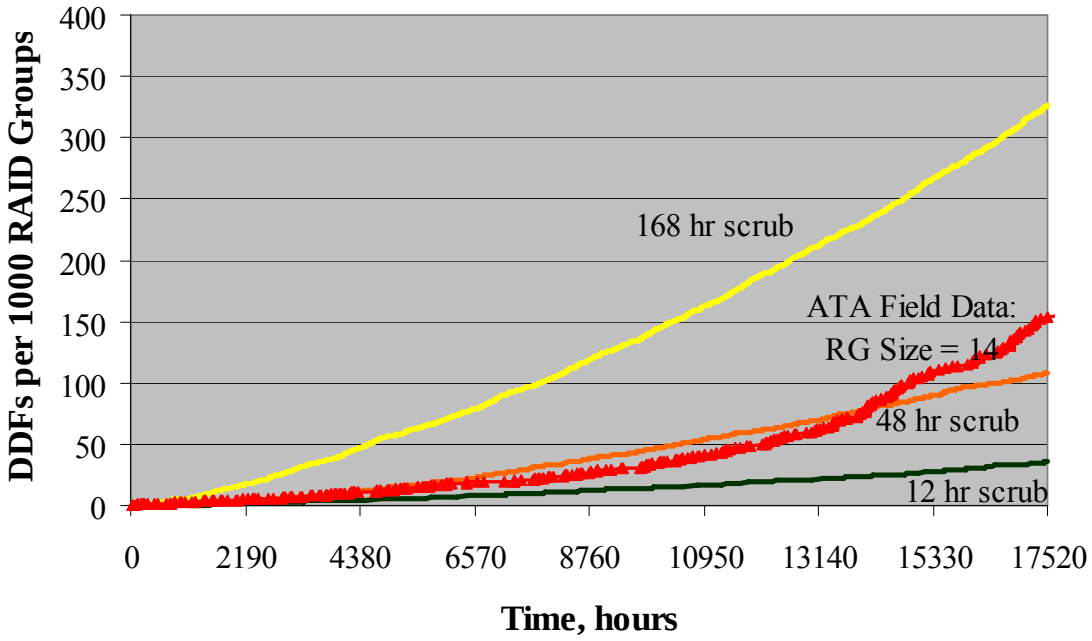


Figure 57 - ATA RAID Group Size 16: Model versus Field data

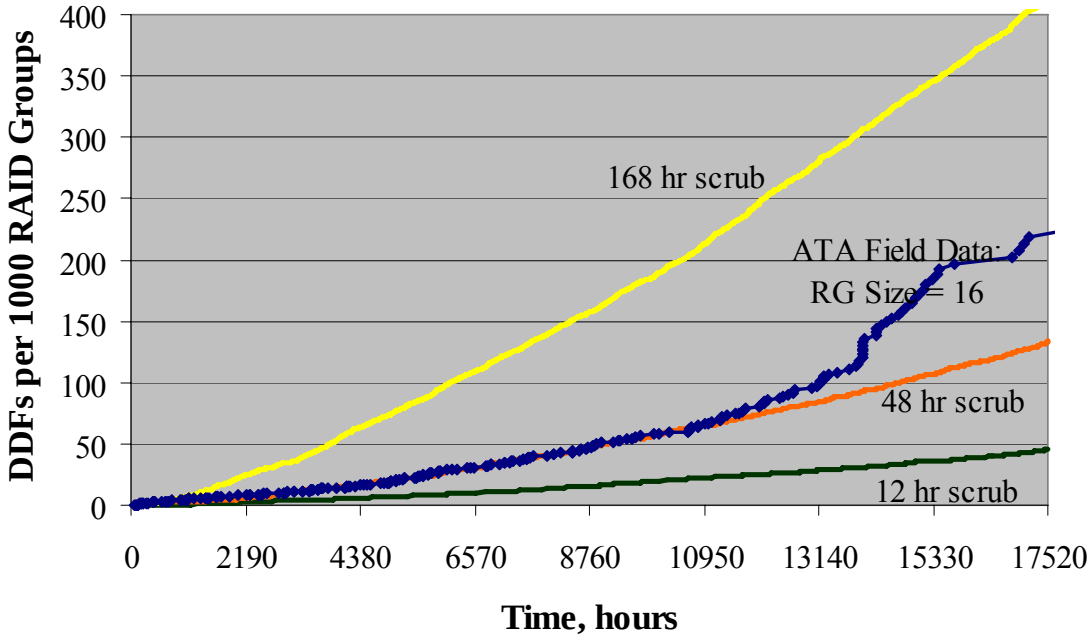


Figure 58 - ATA RAID Group Size 14: Model versus Field Data

The model fit for RAID group size 14 assuming 48 hour scrub is not quite as good a match to the field data as for RAID group size 16, but is still extremely close. Because the size 14 RAID group plots below the 48 hour line, it most likely experiences scrubs more frequently than for the size 16 RAID group. Since scrubbing frequency is an indirect function of the number of HDDs in the RAID group, this plot is consistent with expectations relative to the larger RAID group plot. These plots show the potential accuracy of the model, especially when compared to the estimates derived from the MTTDL method or Markov models. The MTTDL models cannot account for latent defects or increasing HDD failure rates. When the operational failure rate (TTOp) and repair rate (TTR) are assumed constant with mean values 75,000 hours and 12 hours respectively, the MTTDL model calculates 6.6 DDFs per 1000 RAID groups at 13,140 hours (latent defects and scrubbing cannot be included).

9.8 Results Summary

The results of these analyses are quite extensive and warrant a summary.

- Using simplifying assumptions of constant failure rate and replacement rate with no latent defects and no scrubbing, the results from the new model are the same as for the MTTDL and Markov processes, as they should be. (Section 9.2)
- The model is sensitive to modifications of the input distributions even without latent defects or scrubbing (Section 9.2)
- Latent defects, if not scrubbed, may cause the number of DDFs to increase by 5000 times over the number estimated when they are not considered (Section 9.3)
- Increasing HDD failure rates can result in a NHPP ROCOF without scrubbing (Section 9.3)

- Scrubbing can reduce the number of DDFs by multiple orders of magnitudes (Section 9.4)
- Model results are sensitive to Read Error Rate (latent defect distribution) and accurate results depend on accurate understanding of the latent defect distribution (Section 9.4)
- The TTLd distribution can result in an ROCOF that is either HPP (constant ROCOF) or NHPP depending on the value selected. (Section 9.4)
- RAID group size has a loosely predictable impact for a given set of input distributions. Ratios of binomial coefficients may provide "ball park" estimates (Section 9.5)
- The shape parameter (β for Weibull) can greatly affect the ROCOF. Assuming a constant failure rate (β = 1) when it isn't can yield inaccurate results (Section 9.6)
- Field data match the model for two different sets of distributions (Section 9.7)
    - Fibre-channel HDDs which have a moderate characteristic life (461,386 hours) and slightly increasing failure rate (shape parameter = 1.14)
    - ATA HDDs, which have a low characteristic life (75,000 hours) and significantly increasing failure rate (shape parameter = 1.4)
- The model clearly shows the potential to predict the number of DDFs expected in the field much better than the MTTDL or Markov models (Section 9.7)
- The model generally predicts greater number of DDFs than for the MTTDL method due to the inclusion of latent defects and time dependent input distributions (All of Chapter 9)

- Numeric comparisons of the number of DDFs for the various cases show that in all cases, the number of DDFs estimated from the model is greater than the number estimated from the MTTDL.

Table 10 - Comparison of FC Results: Model versus MTTDL

	Scrub Time	RG = 14	RG = 16
MTTDL	n/a	0.13	0.18
New Model	12 hrs	4.2	5.5
	48 hrs	18	24
	168 hrs	57	69

Table 11 - Comparison of ATA Results: Model versus MTTDL

	Scrub Time	RG = 14	RG = 16
MTTDL	n/a	6.8	9
New Model	12 hrs	7	36
	48 hrs	27	108
	168 hrs	95	326

Chapter 10 Conclusions

Conclusions are separated into my contribution and added value of my research, and the impact of my research.

10.1 Significant New Research

While some of the research is well known but not published (HDD failure modes, mechanisms), the impact of these failure modes on the model is novel. Previous researchers have usually ignored latent defects. Data can become corrupt simply by having the head fly over the disc surface. Reading or writing is not necessary and "bit rot" is negligible as a failure mode. My contribution is that by clearly identifying the failure modes and mechanisms of interest, dividing them into operational and latent, I was able to collect appropriate and accurate data for the model.

One of the greatest contributions is the understanding of failure distributions for HDDs used in RAID usage environments. These studies have identified data attributes critical to developing an accurate model. These attributes include recognition that HDD failure rates are rarely constant and are highly dependent on manufacturer and vintage.

The read error rates reported are also a significant contribution. These were developed from large populations of HDDs and provide a better understanding of read error rates than other researchers.

10.2 Added Value in Modeling

This research has resulted in a very flexible and comprehensive model to assess the expected number of double disk failures. The modeling technique addresses many of the erroneous assumptions that prevent the MTTDL and Markov based estimates from being accurate. I have corrected two dominant errors that have not been addressed to date:

- System ROCOF is not the same as the failure rate: ROCOF is sequence dependent whereas the commonly used "failure rate" is not sequence dependent
- Input distributions are not homogeneous Poisson processes
    - HDD failures are time dependent
    - Restoration distributions are time dependent
    - Restoration distributions may have a minimum time to restore (non-zero location parameter)
    - Once written, data can become corrupt (change)

10.3 New Model

The culmination of this research is a new model method. I developed a sequential Monte Carlo method that mimics the sequential nature of the failure process and restoration processes. Each HDD slot carries its own failure, restoration, latent defect and scrub distributions. This allows some fraction of the HDDs to possess different distributional characteristics than the remaining, as is the case for multiple vintages or mixtures of HDDs from different manufacturers.

The model recognizes and models the conditionality between the operational and latent defects. That is, concurrent latent defects and operational failures in which the

latent defect occurs first will result in a DDF. Operationally, the probability of a latent defect occurring after the operational failure is only possible by getting a data erasure (a subset of the read error rate) in a section of the unrecovered data blocks on the media of the other HDDs during the 12 to 168 hours of reconstruction. Since this probability is extremely small, this sequence of events is excluded from the model. However, the model does include data corruption as a function of time and usage as the basis for the latent defects.

The model can accept RAID groups of size 2 to 25 and any user defined mission time. Latent defects can be excluded if desired. This feature was used in the debugging and troubleshooting of the visual basic macro. The model has demonstrated sensitivity to all input parameters and is consistent with the MTTDL and Markov models when the inputs are set to produce HPP ROCOF. The model has shown its ability to very accurately model RAID groups of size 14 and 16 that have been deployed in the field for over 18 months. The model and field data, confirm the superiority of this model over any other developed to date.

10.4 Impact of the Research

The impact of this research is evident in several ways. This paper has been accepted for presentation at the 37th Annual IEEE/IFIP Conference on Dependable System and Networks, June, 2007. It will be published in the IEEE-DSN/DCCS Proceedings. The field data has been the subject of 6 publications [58], [59], [60], [61], [62] and [63]. Additionally, other researchers have referenced one or more of these publications.

Chen, Y., et al., "Managing Server and Energy and Operational Costs in Hosting Centers," *Sigmetrics '05*, Jun 6-10, Banff, Alberta Canada.

G. F. Hughes, J. F. Murray, K. Kreutz-Delgado, and C. Elkan, "Improved Disk Drive Failure Warnings," *IEEE Trans. on Reliability*, vol. 51, no. 3, Sep. 2002, pp350-357.

B. Schroeder and G. Gibson, "Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?" *Proc. of 5th USENIX Conference on File and Storage Technologies* (FAST), Feb. 2007.

E. Pinheiro, W.-D. Weber, and L. A. Barroso, "Failure Trends in a Large Disk Drive Population," *Proc. of 5th USENIX Conference on File and Storage Technologies* (FAST), Feb. 2007.

D. Anderson, J. Dykes, and E. Riedel, "More than an Interface," *Proc. of 2nd USENIX Conference on File and Storage Technologies* (FAST '03), Feb. 2003, pp245-257.

T. J. E. Schwarz et al., "Disk Scrubbing in Large Archival Storage Systems," *IEEE Computer Society Symposium,* MASCOTS, 2004, pp1161-1170.

Q. Xin, E. L. Miller and T. J. E. Schwarz, "Evaluation of Distributed Recovery in Large-Scale Storage Systems," *13th IEEE Symp. on High Performance Distributed Computing*, Jun. 2004.

Q. Xin, T. J. E. Schwarz and E. L. Miller, "Disk Infant Mortality in Large Storage Systems," *IEEE 13th Annual Modeling, Analysis, and Simulation of Computer and Telecommunication Systems,* (MASCOTS '05), 2005.

J-F Paris and D. D. E. Long, "Using Device Diversity to Protect Data against Batch Correlated Disk Failures," Storage '06, Alexandria, VA, 2006, pp47-51.

RIAC Journal, Q2, 2005.

10.5 Future work

After discussions with design engineers at Maxtor, Hitachi, Seagate, Fujitsu and Western Digital, all assert that usage affects reliability. Not so much in terms of bulk HDD temperature, but the localized head temperature. This is greatly affected by the

profile of reading and writing, the length of the blocks written and whether the writes are sequential or random. However, a recent paper [67] asserts that the relationship between usage and failures is "weaker than previous work has suggested." This apparent discrepancy between the HDD manufacturers and a respectable data analysis needs further investigation. The significance is great in that the usage/duty cycle can affect the operational time to failure and, more worrisome, the generation of latent defects. The significant questions relate not to the macro level duty cycle of percent I/O, but the specific head usage.

The second area is in expanding the model to include RAID-6, the N+2 redundancy configuration. Knowing that HDDs will continue to grow in capacity (750 GB is now available and a 1TB will be available soon) means the number of latent defects will also increase. While this can be remedied in part by scrubbing, additional scrubbing impacts the system performance. An obvious question is, should all high reliability RAID systems be made with RAID-6? Can RAID-4 and RAID-5 continue to provide the necessary reliability without affecting performance? There are some people who believe all systems should be RAID-6 today!

A last area, although of lower importance, is developing a better understanding of whether there is any noticeable predicted difference between RAID-4 and RAID-5. Some contend that RAID-4 is more stressful on the parity HDD than RAID-5. Tentative data doesn't support that contention, but modeling can investigate the impact if there is a difference.

Chapter 11 Appendices

These appendices document important aspects of this research, including creation of a MCF plot in Section 11.1, the Excel input page in 11.2.1, the VBE code in 11.2.2, and the output necessary "post-processing" in 11.3.

11.1 Mean Cumulative Failure function

Results of the new model and the Monte Carlo simulations are presented as mean cumulative failure plots (MCFs). These plots are non-parametric (do not assume an underlying failure distribution), can plot values greater than unity and are, therefore, excellent for presenting RAID system results without assuming a HPP. Each plot represents 1,000 identical RAID groups only to allow the ordinate to exceed unity for plotting purposes. These plots can be physically interpreted as having been generated in the following manner:

- 1000 systems of RAID N+1 are placed in a room and run for 10 years (87,600 hours).
- Operating time for all systems combined is tracked periodically (hourly), as is the cumulative number of failures.
- When a failure occurs, the cumulative number of failures is divided by the cumulative number of total hours and a point is plotted.

Since the system is restored (per the model assumptions) it can experience multiple failures in the 10 years. This is very similar to a CDF in the probability space, but a CDF, by definition, never exceeds unity. MCF plots are usually presented as "stair-steps". As an example of how they are created, consider a group of 5 systems that all begin operation at the same time and run for 2 years (17,520 hours). The time to failure and the associated system number and the mean cumulative failure percent are shown in Table 12. Systems 1, 2, 3 and 5 experienced failures, but system 4 did not.

Table 12 - Example Data for MCF Plot

Failure Number	Time to Failure, hrs.	System Number	Cumulative failure percent
1	500	1	20%
2	6430	5	40%
3	10587	3	60%
4	12249	2	80%
5	13280	3	100%
6	14617	1	120%

The times to failure are plotted in Figure 59. The left axis indicates the cumulative number of failures per the population of 5 systems. Notice that the fraction of failures, plotted on the right axis and expressed as a percentage, is 120%. This means that, on the average, there were 1.2 failures per system after 2 years operation. From this plot, the general trend of the ROCOF can be deduced. A straight line (constant slope) means the ROCOF is neither increasing nor decreasing. In this plot, the rate is generally increasing (has a positive slope). Ascher [52] cautions

against attempting to fit a distribution to this curve, however, because there is no basis for a single underlying distribution. The shape may decrease or oscillate after the plotted points.

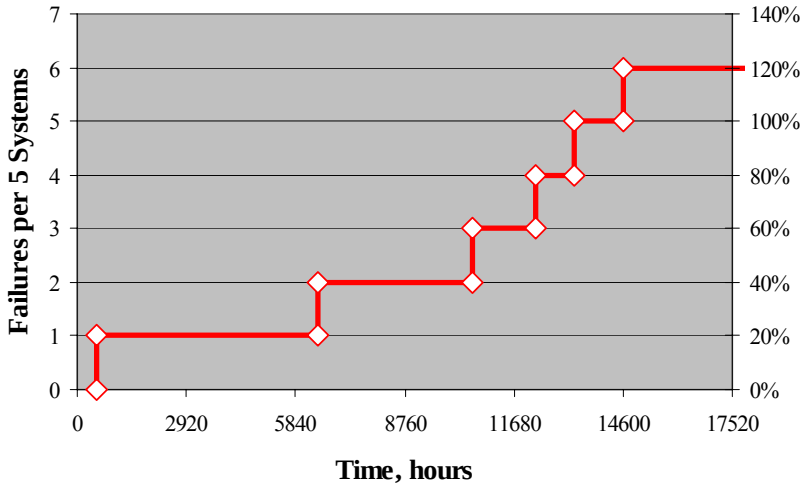


Figure 59 - Plot of MCF example
Notice from this MCF that the ROCOF for this system is increasing (in general)

11.2 Sequential Monte Carlo Simulation

The inputs for the Monte Carlo simulation are extremely flexible, but require attention to make sure they are really what you want them to be. There is no error checking for bizarre inputs or accidental errors.

11.2.1 Inputs

The input example in Figure 60 illustrates this. The model conditions are shown in the upper left section of the figure in lavender color. These include the total number of HDDs in the RAID group ($N + 1 = 8$), the mission duration (87,600 hours), the number of parity disks (1), the number of "runs" (10,000), and whether the

simulation includes latent defects (y/n). The number of parity disks is included as a variable to permit future modification, so it can evaluate *N + 2* (RAID-DP) systems, but currently it is capable only of *N+1* systems. The "screen" is set to "off" so the computer screen will not be updated as the simulation progresses. When "on", the screen updates slow the simulation greatly, but is useful for debugging.

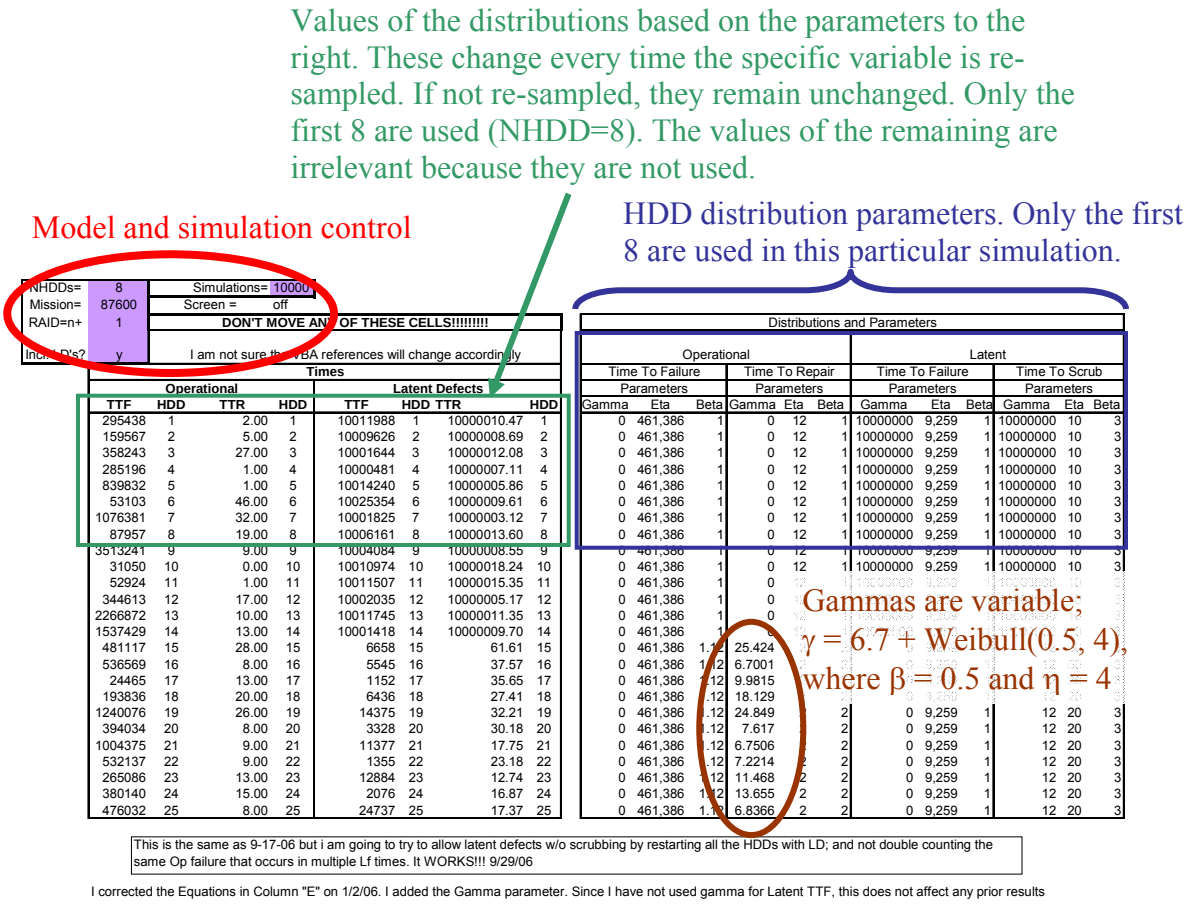


Figure 60 - Example of Monte Carlo inputs

The distribution parameters for the HDDs are on the right side of Figure 60. There are four distributions associated with each HDD slot. This permits one to

model mixtures of HDD manufacturers or vintages by using different parameters. The sampled values for the distributions are shown on the left half of the figure. Initially, all are sampled once. Subsequently, new values are sampled only when needed, per the description in Chapter 7. The @RISK™ add-in for Microsoft's Excel™ is used to generate the sampled values from the parameters.

11.2.2 Excel™ Visual Basic Macro

A macro based on Excel's Visual Basic (VBE) is used to determine when two HDDs have failed simultaneously. The flow chart for the VBE macro is shown in Figure 61.

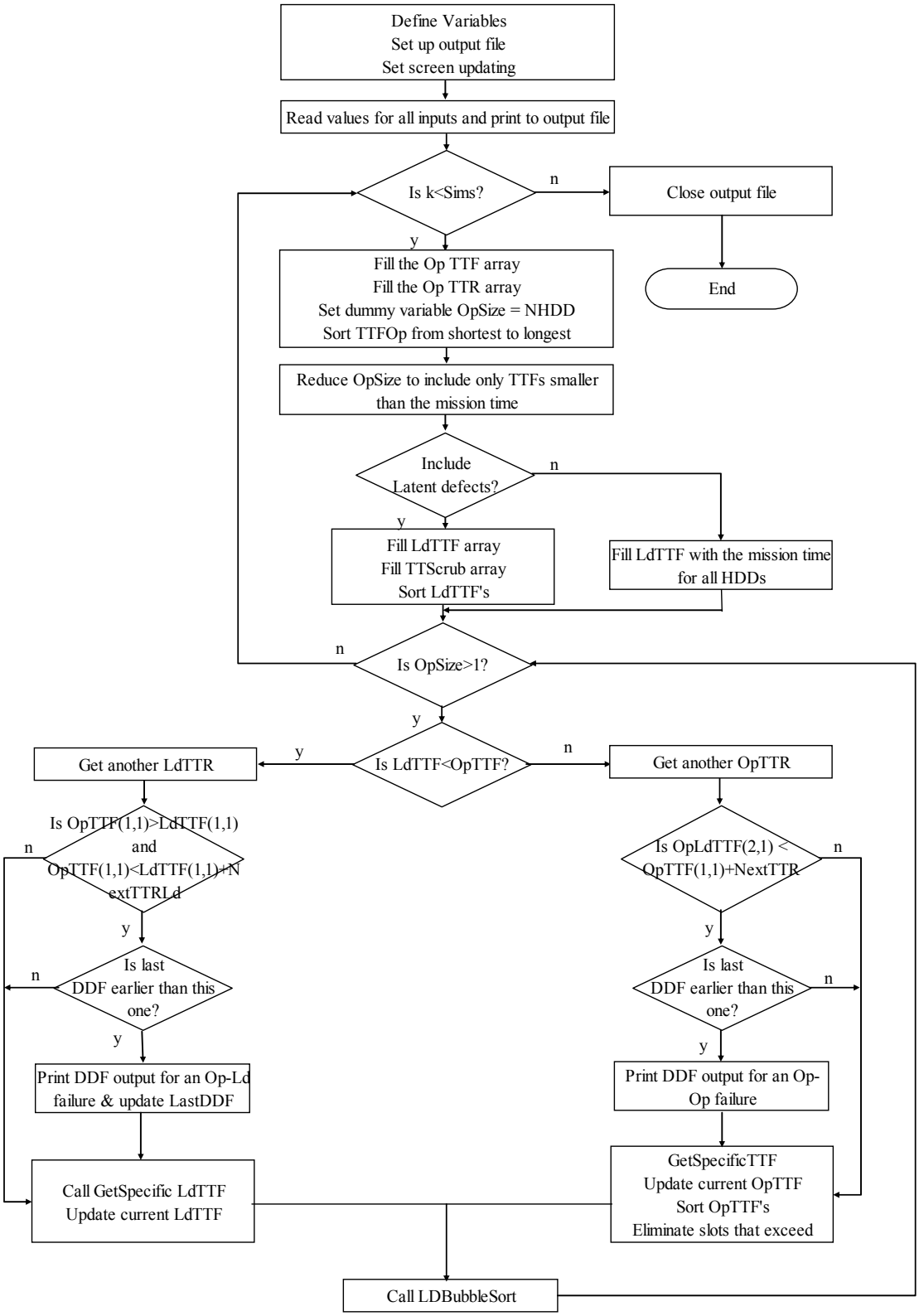


Figure 61 - Flow chart for VBE macro

11.2.3 Output Format

An example output file is shown in Figure 62. At the top of the file are all the input parameters. These are followed by the results. Included in the results are the simulation number that generated the DDF. To the right of that is the earliest time of the interval that the system was at risk (had its first failure or latent defect). The second column shows the time of the second failure (that resulted in the DDF) and the fourth column shows the end of the risk-interval. Finally, the last column indicates whether the failure was a result of two operational failures, denoted OP, or the result of a latent defect followed by an operational failure, denoted LD. The example provided was actually run for 100 runs, but to fit the page, the figure was cropped and the last few results deleted.

RAID Group Size =	8
Parity Disks =	1
Mission Duration =	87600
Using Latent Defects? y	
Simulations =	100

Operational						Latent					
Time to Failure			Time to Repair			Time to Failure			Time to Repair		
Gamma	Eta	Beta	Gamma	Eta	Beta	Gamma	Eta	Beta	Gamma	Eta	Beta
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3

Sim #	TTF1	TTF2	TTF1+TTR1	
1	790	20508	10000801	Ld
1	20757	27494	10020766	Ld
1	27748	32975	10027760	Ld
3	1637	37602	10001643	Ld
5	535	27325	10000549	Ld
8	278	32602	10000292	Ld
14	1875	60753	10001877	Ld
14	61424	68569	10061433	Ld
14	68621	84653	10068626	Ld
23	280	1966	10000294	Ld
24	129	9082	10000141	Ld
24	9951	34532	10009959	Ld
26	257	8731	10000265	Ld
28	1487	60870	10001495	Ld
29	39	40965	10000046	Ld
32	913	45581	10000924	Ld
32	50649	76951	10050664	Ld
44	705	43697	10000717	Ld
45	188	2193	10000193	Ld
45	3763	8437	10003774	Ld
45	8979	36600	10008984	Ld
46	330	26179	10000335	Ld
46	26293	43937	10026304	Ld
46	45637	46383	10045646	Ld
50	851	21223	10000856	Ld
50	23226	56786	10023236	Ld
51	2303	47856	10002310	Ld
51	48631	60409	10048641	Ld
56	749	4675	10000759	Ld
58	359	54537	10000366	Ld
58	55114	57124	10055129	Ld
62	291	74727	10000303	Ld
65	644	29918	10000654	Ld
65	30350	32021	10030360	Ld
68	184	24924	10000192	Ld
79	1939	5186	10001952	Ld
82	5724	43841	10005738	Ld
83	74	8335	10000083	Ld
83	8636	41190	10008645	Ld
84	743	15840	10000750	Ld
84	16531	28815	10016545	Ld
84	28850	51654	10028854	Ld
85	74	25506	10000079	Ld
85	25853	46023	10025863	Ld
85	46118	62746	10046126	Ld
87	989	30964	10000999	Ld
90	1163	47620	10001180	Ld
90	47954	51433	10047964	Ld

Figure 62 - Example of simulation output

11.3 ***Post-Processing***

The results in Figure 62 need a bit of post processing to create the MCF plots. First all the times to DDF must be sorted from smallest to largest. The, the plotting positions must be determined, based on the number of simulations needed to create the data points. The post-processing required for the example is shown in Figure 63.

RAID Group Size =	8
Parity Disks =	1
Mission Duration =	87600
Using Latent Defects? y	
Simulations =	100

Operational						Latent					
Time to Failure			Time to Repair			Time to Failure			Time to Repair		
Gamma	Eta	Beta	Gamma	Eta	Beta	Gamma	Eta	Beta	Gamma	Eta	Beta
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3
0	461386	1	0	12	1	6	9259	1	10000000	10	3

Sim #	TTF1	TTF2	TTF1+TTR1		Example Results	
1	790	20508	10000801	Ld	1966	0.01
1	20757	27494	10020766	Ld	2193	0.02
1	27748	32975	10027760	Ld	4675	0.03
3	1637	37602	10001643	Ld	5186	0.04
5	535	27325	10000549	Ld	5799	0.05
8	278	32602	10000292	Ld	8335	0.06
14	1875	60753	10001877	Ld	8437	0.07
14	61424	68569	10061433	Ld	8731	0.08
14	68621	84653	10068626	Ld	9082	0.09
23	280	1966	10000294	Ld	12184	0.1
24	129	9082	10000141	Ld	15840	0.11
24	9951	34532	10009959	Ld	16633	0.12
26	257	8731	10000265	Ld	20508	0.13
28	1487	60870	10001495	Ld	21223	0.14
29	39	40965	10000046	Ld	24924	0.15
32	913	45581	10000924	Ld	25392	0.16
32	50649	76951	10050664	Ld	25506	0.17
44	705	43697	10000717	Ld	26179	0.18
45	188	2193	10000193	Ld	27325	0.19
45	3763	8437	10003774	Ld	27494	0.2
45	8979	36600	10008984	Ld	28815	0.21
46	330	26179	10000335	Ld	29918	0.22
46	26293	43937	10026304	Ld	30964	0.23
46	45637	46383	10045646	Ld	32021	0.24
50	851	21223	10000856	Ld	32602	0.25
50	23226	56786	10023236	Ld	32975	0.26
51	2303	47856	10002310	Ld	34532	0.27
51	48631	60409	10048641	Ld	36600	0.28
56	749	4675	10000759	Ld	37602	0.29
58	359	54537	10000366	Ld	40965	0.3
58	55114	57124	10055129	Ld	41190	0.31
62	291	74727	10000303	Ld	43697	0.32
65	644	29918	10000654	Ld	43841	0.33
65	30350	32021	10030360	Ld	43937	0.34
68	184	24924	10000192	Ld	45581	0.35
79	1939	5186	10001952	Ld	46023	0.36
82	5724	43841	10005738	Ld	46383	0.37
83	74	8335	10000083	Ld	47620	0.38
83	8636	41190	10008645	Ld	47856	0.39
84	743	15840	10000750	Ld	51433	0.4
84	16531	28815	10016545	Ld	51654	0.41
84	28850	51654	10028854	Ld	54537	0.42
85	74	25506	10000079	Ld	56786	0.43
85	25853	46023	10025863	Ld	57124	0.44
85	46118	62746	10046126	Ld	60409	0.45
87	989	30964	10000999	Ld	60753	0.46
90	1163	47620	10001180	Ld	60870	0.47
90	47954	51433	10047964	Ld	62746	0.48
92	4878	25392	10004887	Ld	68569	0.49
93	1053	69381	10001063	Ld	69381	0.5
98	516	5799	10000524	Ld	74727	0.51
99	941	12184	10000950	Ld	76951	0.52
99	12331	16633	10012341	Ld	84653	0.53

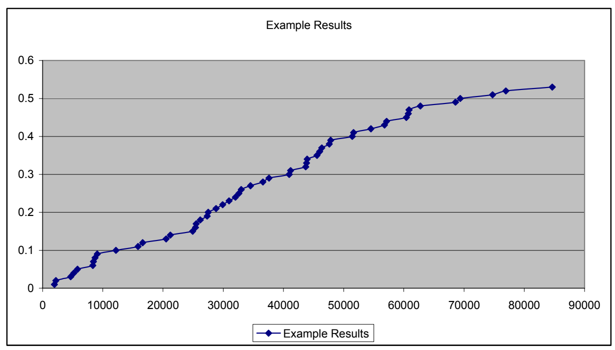


Figure 63 - Output Example with Post-processing

The post-processing is shown in the red rectangle and the chart to the right. This is done manually.

Glossary

A - Amp
Å - Angstrom
A/cm^2 - Amps per square centimeter
AFM - anti-ferromagnetic
AMR - Anisotropic magneto-resistive
BER - Bit error rate
C = Degrees centigrade
DDF - Double disk failure
DRAM - Dynamic random access memory
ECC - Error correcting codes
EM - Electro-migration
EOS - Electrical overstress
ESD - Electro-static discharge
eV - electron volt
FM - ferromagnetic
GB - Giga-byte
GMR - giant magnetoresistance
HBM - Human body model
H_c - Coercivity field
HDD - Hard disk drive
H_{ex} - Exchange field
HP - Homogeneous Poisson
HPP - Homogeneous Poisson process
I/O - Input/Output
LMR - Longitudinal magnetic recording
ma - milliamp
MB - Mega-byte
μ-in - micro-inch
MKV - Markov
mm - millimeter
MTTDL - Mean time to data loss
NHP - Non-homogeneous Poisson
NHPP - Non-homogeneous Poisson process
nJ - nano-Joule
NRRO - Non-repeatable run-out
ns - nano-second
PES - Position error sensing
PMR - Perpendicular magnetic recording
PWA - Printed wiring assembly
PWBA - Printed wiring board assembly
R - Resistance
RAID - Redundant array of inexpensive disks
RER - Read error rate

RRO - Repeatable run-out
T/A - Thermal asperity
T_b - Blocking temperature
TPI - Tracks per inch
TTLd - Time to latent defect
TTOp - Time to operational failure
TTR - Time to restore
TTScrub - Time to scrub
V_{HBM} – Volts using the human body model

References

[1] "Worldwide Disk Storage Market Surges Ahead on Strong Third Quarter Demand, According to IDC," Data Storage News, Dec. 7, 2006, www.datastorex.com/content/edit-ceonews.asp

[2] G. Forest, "Tiered Storage School 201," Storage Visions 2006, Keynote presentation, http://wp.bitpipe.com/resource/org_1116538445_185/34334_Forest201.pdf?site_cd=bp&asrc=ORG_OSE_GOOGUS

[3] R. De, "Storage: Enterprise Market Blockbuster," Dataquest, http://www.dqindia.com/content/DQTop20_2006/giants06/2006/106072513.asp, July 25, 2006.

[4] H. H. Kari, "Latent Sector faults and Reliability of Disk Arrays," Ph. D. Dissertation, TKO-A33, Helsinki University of Technology, Espoo, Finland, 1997 http://www.cs.hut.fi/~hhk/phd/phd.html

[5] T. J. E. Schwarz et al., "Disk Scrubbing in Large Archival Storage Systems," *IEEE Computer Society Symposium,* MASCOTS, 2004, pp1161-11.

[6] G. A. Gibson, "Redundant Disk Arrays: Reliable, Parallel Secondary Storage," Ph. D. Dissertation, Dept of Computer Science, UC Berkeley, April 1991. T7.6 1991 G52 ENGI.

[7] V. Prabhakaran, "IRON File Systems," *SOSP '05*, Oct. 2005, Brighton, UK.

[8] A. Paranjpe, "Thin Film Fabrication in the PMR Era," *IDEMA Perpendicular Recording Symposium*, IDEMA, Dec. 6, 2006.

[9] C. Ruemmler and J. Wilkes, "An Introduction to Disk Drive Modeling," *IEEE Computer*, Vol. 27, No. 3, March 1994, pp17-29.

[10] J. Best et al., "The Femto Slider in Hitachi Hard Disk Drives," http://www.hitachigst.com/tech/techlib.nsf/techdocs/AE7AEDB327B2E21186256D330078799B/$file/Femto_white_paper_FINAL_082505.pdfC.

[11] "Interface Materials," Hitachi Global Storage Technologies, http://www.hitachigst.com/hdd/research/storage/im/index.html

[12] "Recording Head/Adv. Head Processing," Hitachi Global Storage Technologies, http://www.hitachigst.com/hdd/research/recording_head/headprocessing/index.html

[13] S. Gider, L. Baril and D. Mauri, "Kinetics of Thermal Decay in NiMn and PtMn Spin Valve Devices," *IEEE Transactions on Magnetics*, Volume 37, Issue 4, Part 1, July 2001, pp1704 – 1706.

[14] C. Lee, A. J. Devasahayam, C. Hu, Y. Zhang, M. Mao, J. C. S. Kools an K. Rook, "Critical Thickness Effects of NiFeCr-CoFe Seed Layers for Spin Valve Multilayers," *IEEE Transactions on Magnetics*, Volume 40, Issue 4, Part 2, July 2004, pp2209 – 2211.

[15] T. Lin, D. Mauri, B. York and P. Rice, "Crystalline Reconstruction in Ni-Cr-FeNi-Fe Films," *Applied Physics Letters*, Amer. Inst. of Phys., Volume 84, Number 3, January 19, 2004, pp386 – 388.

[16] A. K. Petford-Long, X. Portier, E. Y. Tsymbal, T. C. Anthony and J. A. Brug, "In-Situ Lorentz Microscopy Studies of Spin-Valve Structures," *IEEE Transactions on Magnetics*, Volume 35, Issue 2, March 1999, pp788 – 793.

[17] A. Al Mamum and S. S. Ge, "Precision Control of Hard Disks," *IEEE Control Systems Magazine*, 2005, pp 14-19.

[18] M. A. Russak, G. Bertero, "Status and Future Direction of PMR Media," *IDEMA Perpendicular Recording Symposium*, IDEMA, Dec. 6, 2006.

[19] S. Prakash, K. Pentek and Y. Zhang, "Reliability of PtMn-based Spin Valves," *IEEE Transactions on Magnetics*, Volume 37, Issue 3, May 2001, pp1123 – 1131.

[20] I. Tsu; G. A. Burg and W. P. Wood, "Degradation of Spin Valve Heads Under Accelerated Stress Conditions," *IEEE Transactions on Magnetics*, Volume 37, Issue 4, Part 1, July 2001, pp1707 – 1709.

[21] Y. Yang, S. Shojaeizadaeh, J. A. Bain, J. G. Zhu and M. Asheghi, "Detailed Modeling of Temperature Rise in Giant Magnetoresistive Sensor During Electrostatic Discharge Event," *Journal of Applied Physics*, Volume 95, Number 11, June 1, 2004, pp6780 – 6782.

[22] A. Wallash and W. Wang, "A Study of Diode Protection for Giant Magnetoresistive Recording Heads," *Proc. Electrical Overstress/Electrostatic Discharge Symp.*, 1999, Sept 28-30, 1999, pp385 – 390

[23] A. Wallash, "Understanding ESD Damage to Magnetoresistive (MR) Recording Heads," Quantum Corp., http://www.wallash.com/datastorage.pdf

[24] A. Wallash, "New Early Failure Phenomenon in Electrostatic Discharge Damages Giant Magnetoresistive Recording Heads," *Journal of Applied Physics*, Volume 93, Number 10, May 15, 2003, pp7319 – 7321.

[25] C. Moore and A. Wallash, "ESD Testing of GMR Heads as a Function of Temperature," *Proc. Electrical Overstress/Electrostatic Discharge Symp.*, 1999, Sept 28-30, 1999, pp309 – 314.

[26] D. A. Patterson, P. Chen, G. Gibson, R. H. Katz, "Introduction to Redundant Arrays of Inexpensive Disks (RAID)," *Thirty-Fourth IEEE Computer Society International Conference: Intellectual Leverage*, COMPCON, 27 Feb.-3 March 1989, pp112 – 117.

[27] M. L. Shooman, *Reliability of Computer Systems and Networks*, Wiley, New York, 2002.

[28] P. Corbett et al., "Row Diagonal Parity for Double Disk Failure Correction," *Proc. of 3rd USENIX Conference on File and Storage Technology*, San Francisco, 2004.

[29] D. A. Patterson, G. A. Gibson, R. H. Katz, "A Case for Redundant Arrays of Inexpensive Disks (RAID)," *Proc., ACM Conference on Management of Data (SIGMOD)*, Chicago IL, June 1988, pp109-116.

[30] W. A. Burkhard, J. Menon, "Disk Array Storage System Reliability," *The Twenty-Third International Symposium on Fault-Tolerant Computing*, FTCS-23, 22-24 June 1993, pp432 – 441.

[31] P. M. Chen, E. K. Lee, G. A. Gibson, R. H. Katz, D. A. Patterson, "RAID: High-Performance, Reliable Secondary Storage," *ACM Computing Surveys*, 1994, pp145-186.

[32] W. V. Courtright, II, "A Transactional Approach to Redundant Disk Array Implementation," Ph. D. Thesis, CMU-CS-97-141, School of Computer Science, Carnegie Mellon University, 15 May 1997.

[33] "IDEMA Standard 98-2, Specification of Hard Disk Drive Reliability," International Disk Equipment Manufacturers Association, Sunnyvale, CA. 1998. http://www.idema.org

[34] R. Geist and K. Trivedi, "An Analytic Treatment of the Reliability and Performance of Mirrored Disk Subsystems," *Twenty-Third Inter. Symp. on Fault-Tolerant Computing*, FTCS-23, June 1993, pp442 - 450.

[35] M. Holland, "On-Line Data Reconstruction In Redundant Disk Arrays," Ph. D. Dissertation, CMU-CS-94-164, School of Computer Science, Carnegie Mellon University, May 1994.

[36] M. Malhotra, "Specification and Solution of Dependability Models of Fault Tolerant Systems," Ph. D. Dissertation, CS-1993-12, Dept of Computer Science, Duke University, May 14, 1993.

[37] M. Schulze, G. Gibson, R. Katz, D. A. Patterson, "How reliable is a RAID?" *Thirty-Fourth IEEE Computer Society International Conference: Intellectual Leverage*, COMPCON, 27 Feb.-3 March 1989, pp118 – 123.

[38] T. J. E. Schwarz, "Reliability and Performance of Disk Arrays," Ph. D. Dissertation, UCSD, Computer Science Dept., 1994.

[39] T. J. E. Schwarz, W. A. Burkhard, "Reliability and Performance of RAIDs," *IEEE Transactions on Magnetics*, vol. 31, no. 2, March 1995, pp1161 – 1166.

[40] J. Changlong, M. Cheng, H. Ning, W. Chongyang, J. Huibo, "Reliability Analysis of High-speed and Sustained Data Recording System Based on Disk Array," *2002 International Conference on Communications, Circuits and Systems*, IEEE, vol. 1, 29 June-1 July 2002, pp105 – 108.

[41] Y. Chen, L. M. Ni, M. Yang, P. Mohapatra, "CoStore: A Serverless Distributed File System Utilizing Disk Space On Workstation Clusters," *21st IEEE International Performance, Computing, and Communications Conference*, 3-5 April 2002, pp393 – 39.

[42] W. Gang, L. Xiao-Guang, L. Jing, "Parity Declustering Data Layout For Tolerating Dependent Disk Failures In Network Raid Systems," *Proc., Fifth International Conference on Algorithms and Architectures for Parallel Processing*, 23-25 Oct. 2002, pp22 - 25.

[43] C. Jiang, J. Xiong, X. Zhang, H. Jia, D. Xu, "Sustained Data Recording System Based on Disk Array," *Aerospace and Electronic Systems Magazine*, IEEE , vol. 18, no. 11, Nov. 2003. pp31 - 34

[44] S. Savage, J. Wilkes, "AFRAID - A Frequently Redundant Array of Independent Disks," *1996 USENIX Technical Conference*, January 22–26, 1996, pp27 – 39.

[45] Q. Xin; E. L. Miller, T. Schwarz, D. D. E. Long, S. A. Brandt, W. Litwin, "Reliability Mechanisms for Very Large Storage Systems," *Proc. 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies*, (MSST 2003)7-10 April, 2003, pp146 – 156.

[46] J. B. Keats and S. P. Chambal, "Transient Behavior of Time –Between-Failures of Complex Repairable Systems," *Qual. Reliab. Engrnr. Int*, 2002, 18:293-297.

[47] H. Ascher, "[Statistical Methods in Reliability]: Discussion," *Technometrics,* Vol. 25, No. 4, Nov. 1983, pp320-326.

[48] H. E. Ascher, "A Set-of-Numbers is NOT a Data-Set", *IEEE Trans. on Reliability*, Vol. 48, No. 2, Jun. 1999, pp135-140.

[49] H. Ascher, "Reliability Modeling Via Data Analysis," Harold E. Ascher & Associates, Potomac, MD, hasher@att.net

[50] H. Ascher and H. Feingold, "The Aircraft Air Conditioner Data Revisited," *Proc. Annual Reliability & Maintainability Symp.,* January 1979, pp153 – 159.

[51] H. E. Ascher, T. Y. Lin, and D. P. Siewiorek, "*Modification of*: Error Log Analysis: Statistical Modeling and Heuristic Trend Analysis," *IEEE Transactions on Reliability*, vol. 41, no. 4, December 1992, pp599-607.

[52] H. E. Ascher and C. K. Hansen, "Spurious Exponentiality Observed When Incorrectly Fitting a Distribution to Nonstationary Data," *IEEE Transactions on Reliability*, vol. 47, no. 4, December 1998, pp451-459.

[53] W. A. Thompson, "On the Foundations of Reliability," *Technometrics*, Vol. 23, No. 1, Feb. 1981, pp1-13.

[54] W. Nelson, "Graphical Analyses of System Repair Data," *Journal of Quality Technology*, vol. 20, no. 1, Jan. 1988.

[55] L. H. Crow, “Evaluating the Reliability of Repairable Systems,” *Proc. Annual Reliability & Maintainability Symp.,* Jan. 1990, pp275-279.

[56] F. Proschan, "Theoretical Explanation of Observer Decreasing Failure Rate," *Technometrics*, Vol. 5, 1963, pp375-383.

[57] W. A. Thompson, "The Rate of Failure is the Density, Not the Failure Rate," *The American Statistician*, Editorial, vol. 42, no. 4, Nov. 1988, pp288-291.

[58] J. G. Elerath, and S. Shah, "Disk Drive Reliability Case Study: Dependence Upon Head Fly-Height and Quantity of Heads." *Proc. Annual Reliability & Maintainability Symp.,* January 2003, pp608-612.

[59] J. G. Elerath, and S. Shah, "Server Class Disk Drives: How Reliable are they?" *Proc. Annual Reliability & Maintainability Symp.,* January 2004, pp151-156.

[60] S. Shah, and J. G. Elerath, "Disk Drive Vintage and its Affect on Reliability." *Proc. Annual Reliability & Maintainability Symp.,* January 2004, pp163-167.

[61] S. Shah and J. G. Elerath, “Reliability Analysis of Disk Drive Failure Mechanisms,” *Proc. Annual Reliability & Maintainability Symp.,* January 2005, pp226-231.

[62] J. G. Elerath, “Specifying Reliability in the Disk Drive Industry: No More MTBF’s.” *Proc. Annual Reliability & Maintainability Symp.,* January 2000, pp194-199.

[63] J. G. Elerath and S. Magie, “Field Reliability from Post-GA Manufacturing Process and Design Changes,” IDEMA, DISKCON Asia-Pacific, May 2006. http://www.idema.org/_smartsite/modules/local/data_file/show_file.php?cmd=download&data_file_id=1441,

[64] J. Gray, C. van Ingen, “Empirical Measurements of Disk Failure Rates and Error Rates,” Microsoft Research Technical Report, MSR-TR-2005-166, Dec. 2005.

[65] D. Trindade and S. Nathan, “Simple Plots for Monitoring Field Reliability of Repairable Systems,” *Proc. Annual Reliability & Maintainability Symp.,* Jan. 2005, pp539-544.

[66] C. L. T. Borges, et al., “Composite Reliability Evaluation by Sequential Monte Carlo Simulation on Parallel and Distributed Operating Environments,” *IEEE Trans. on Power Systems*, vol. 16, no. 2, May 2001, pp 203-209.

[67] E. Pinheiro, W. D. Weber and L. A. Barroso, "Failure Trends in Large Disk Drive Population," *Proc. 5th USENIX Conference on File Storage Technologies* (FAST '07), Feb. 2007.