Dallmeier

REV. 2.0.0 | 2019-08-01

WHITEPAPER

BASIS-INFORMATION
HDD & RAID

INFORMATIONEN ZUR DER LEBENSDAUER UND FUNKTIONSWEISE
VON FESTPLATTEN UND REDUNDANTEN SPEICHERSYSTEMEN

INHALT

ZUSAMMENFASSUNG

Festplatten (hard disk drives / **HDD**) sind Speichermedien, die binäre Daten unter Verwendung einer ausgefeilten Mechanik auf rotierenden magnetischen Scheiben lesen und schreiben [1]. Sie stellen eine zentrale Komponente aller digitalen Aufzeichnungssysteme dar und werden für die kontinuierliche Speicherung von Audio- und Video-Daten verwendet. Trotz höchster Qualitätsanforderungen muss ein natürlicher Verschleiß von HDDs beachtet werden, der bedingt durch einen in der Regel ununterbrochen Betrieb (24/7) noch verstärkt wird.

Ein **RAID**-System ist eine Speichertechnologie, die mehrere HDDs zu einer logischen Einheit zusammenfasst, in der sämtliche Daten redundant gespeichert werden [2]. Diese Redundanz erlaubt den Ausfall und Ersatz einzelner HDDs ohne Datenverlust. In digitalen Aufzeichnungssystemen werden RAID-Systeme eingesetzt, um den natürlichen Verschleiß und Ausfall von Festplatten abzufangen.

Selbst fortschrittlichste RAID-Systeme in Verbindung mit HDDs höchster Qualität können eine vollkommene Datensicherheit nicht garantieren. Diese Feststellung basiert auf den Grenzen der RAID-Technologie [3] und der in der Praxis beobachteten Lebensdauer von Festplatten [4]. Die geläufige Ansicht, ein RAID-System biete die gleiche Sicherheit wie ein Backup (Datensicherung), ist nicht zutreffend [5]. Wichtige Aufzeichnungen sollten immer durch ein Backup gesichert werden.

Dieses Dokument enthält verschiedene Erklärungen zu Begriffen, die bei der Beurteilung der Zuverlässigkeit von Festplatten relevant sind. Anschließend wird deren tatsächliche Relevanz für die Praxis basierend auf zwei viel beachteten Studien der Google Inc. [4] und der Carnegie Mellon Universität[6] abgeleitet. Darauf aufbauend wird die Funktionsweise verschiedener RAID-Systeme und ihre Vorteile und Grenzen betrachtet.

HDD LEBENSDAUER

Für die Beurteilung der Zuverlässigkeit von Festplatten werden in der Regel verschiedene Angaben (AFR, MTBF, MTTF) aus den entsprechenden Datenblättern der Hersteller verwendet. Zudem werden auch verbreitete Methoden (SMART) zur Ausfallfrüherkennung beachtet.
Wie im Folgenden gezeigt wird, eignen sich diese theoretischen Betrachtungen nur eingeschränkt zur Beurteilung und Planung von Speichersystemen. Sie müssen um Beobachtungen und Erfahrungen aus der Praxis erweitert werden.

2.1 AFR, MTBF UND MTTF

AFR, MTBF und MTTF sind die gängigsten Herstellerangaben zur Zuverlässigkeit von Festplatten. Sie basieren auf Erfahrungswerten, Schätzungen und Hochrechnungen. Daher können sie nicht als absolute Angaben, sondern nur als Erwartungs- oder Wahrscheinlichkeitswerte interpretiert werden.

2.1.1 Definition

2.1.1.1 AFR

Die **AFR** (Annualized Failure Rate / auf ein Jahr hochgerechnete Ausfallrate) ist der prozentuale Ausfallanteil einer bestimmten Menge an Festplatten, der aufgrund von Erwartungswerten auf ein Jahr hochgerechnet wird [6].

2.1.1.2 MTBF

Die **MTBF** (Mean Time Between Failures / Mittlere Betriebsdauer zwischen Ausfällen) gibt die erwartete Betriebsdauer zwischen zwei aufeinanderfolgenden Ausfällen eines Gerätetyps in Stunden an (Definition nach IEC 60050 (191)) [8].

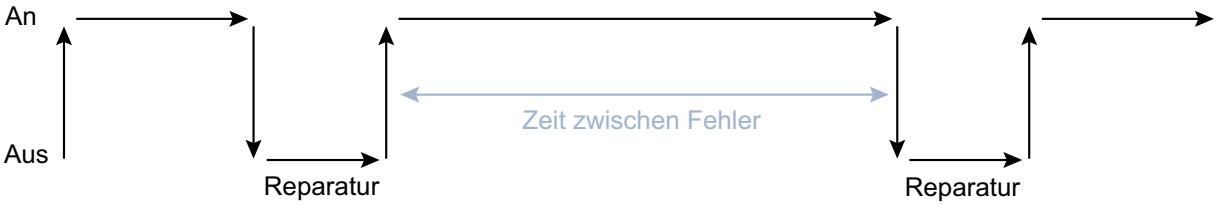

Abb. 2-1 MTBF - Zeit zwischen Fehler

Die MTBF betrachtet also den Lebenszyklus eines Gerätes, das wiederholt ausfällt, repariert und wieder in Betrieb genommen wird.
Die Betrachtung kann aber auch auf einen Ausfall ohne Reparatur bezogen werden, wie er typischerweise bei Festplatten unterstellt wird. In diesem Fall wird also die mittlere Betriebsdauer bis zum Ausfall des Gerätes betrachtet.

2.1.1.3 MTTF

Die **MTTF** (Mean Time To Failure / Mittlere Betriebsdauer bis zum Ausfall) gibt die erwartete Betriebsdauer bis zum Ausfall eines Gerätetyps in Stunden an (Definition nach IEC 60050 (191)) [8].

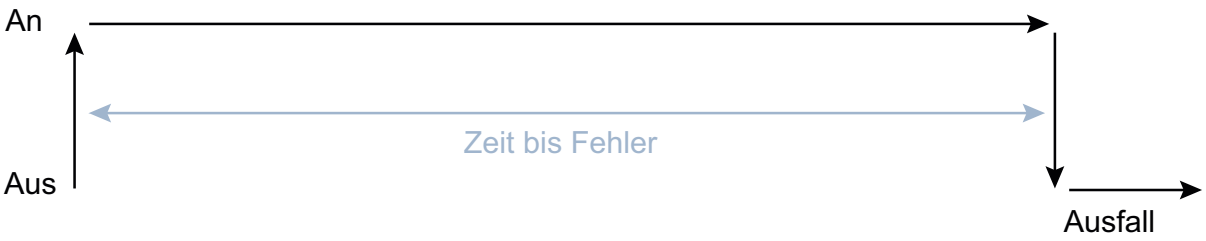

Abb. 2-2 MTTF - Zeit bis Fehler

Die Begriffe MTBF und MTTF werden in Bezug auf Festplatten oft synonym verwendet. Für die MTBF hat sich sogar das Backronym meant time before failure eingebürgert [7].

Die Seagate Technology LLC z.B. schätzt die **MTBF** ihrer Festplatten als die Anzahl der Betriebsstunden pro Jahr geteilt durch die hochgerechnete Ausfallrate **AFR** [9]. Diese Betrachtung basiert auf einem Ausfall ohne Reparatur. **MTTF** wäre also die korrekte Bezeichnung.

2.1.2 Zusammenhang

AFR, MTBF und MTTF sind Angaben, die auf Erfahrungswerten, Schätzungen und Hochrechnungen basieren. Sie können basierend auf einer singulären Betrachtung festgelegt werden oder im Zusammenhang und unter Verwendung verschiedenster Schätzmethoden (z.B. Weibull, WeiBayes) errechnet werden [9].
Vereinfacht, und für die Betrachtungen in diesem Dokument ausreichend, gibt die **MTTF** die Anzahl der Betriebsstunden pro Jahr geteilt durch die hochgerechnete **AFR** an [6].

2.1.3 Interpretation

Wie Seagate spezifizieren die meisten Hersteller die Zuverlässigkeit ihrer Festplatten durch die Angabe von **AFR** und **MTTF**. Wie im Folgenden gezeigt, sind diese Angaben aber immer wesentlich besser als die Beobachtungen in der Praxis.

Die Abweichungen resultieren aus der unterschiedlichen **Definition eines Ausfalls**. Eine Festplatte, die von einem Kunden aufgrund auffälligen Verhaltens ausgetauscht wird, kann vom Hersteller als voll funktionsfähig betrachtet werden. Seagate stellte z.B. fest, dass 43% der zurückgegebenen Festplatten funktionsfähig sind [6]. Für andere Hersteller lassen sich Werte zwischen 15% und sogar 60% belegen [4]. Zudem weichen die **Rahmenbedingungen** der Praxis und der herstellerspezifischen Tests in der Regel voneinander ab. Insbesondere höhere Temperatur und Luftfeuchtigkeit sowie höhere Auslastung durch permanente Schreib-Lese-Vorgänge resultieren in der Praxis in wesentlich höheren Ausfallraten [6].
Eine brauchbare Interpretation von **AFR** und **MTTF** gelingt also erst durch die Beachtung der Vorgehensweise der Hersteller. Wie auch Adrian Kingsley-Huges [10] in seinen Ausführungen feststellt, liegt der Unterschied zwischen beobachteten und angegebenen MTTFs in deren Ermittlung.

Vereinfacht kann die MTTF also wie folgt berechnet werden:

MTTF = ([Testperiode] x [Anzahl HDDs]) / [Anzahl der ausgefallenen HDDs]

Bei einer angenommenen Testperiode von 1.000 Stunden (ca. 41 Tage) mit 1.000 HDDs und einer ausgefallenen HDD ergibt sich also:

MTTF = (1.000 Stunden x 1.000 HDD) / 1 HDD = 1.000.000 Stunden

Die hochgerechnete jährliche Fehlerrate AFR ergibt sich aus dem reziproken Wert:

AFR = ([Anzahl Fehler] / ([MTTF] / [8.760 Stunden pro Jahr])) x 100%
AFR = (1 Fehler / ([1.000.000 Stunden] / [8.760 Stunden pro Jahr])) x 100%
AFR = (1 Fehler / 114,16 Jahre) x 100%
AFR = 0,86 %

Umgekehrt könnte man mit einer geschätzten AFR von 0,86% auch die MTTF errechnen:

MTTF = (1.000 HDD x 8.760 Stunden) / (1.000 HDD x 0,86%)
MTTF = 8.760.000 HDD Stunden / 8,6 HDD
MTTF = 1.018.604 ~ 1.000.000 Stunden (114 Jahre)

Was sagt eine **MTTF von 1.000.000 Stunden** (114 Jahre) also aus?

Die Annahme das eine HDD 114 Jahre betrieben werden kann ist absolut falsch! Dieser Wert besagt, dass bei **1.000 gleichzeitig gestarteten Festplatten der erste Ausfall nach 1.000 Stunden (42 Tage) zu erwarten** wäre.

Anders betrachtet könnte man auch sagen, dass **114 Festplatten ein Jahr lang betrieben werden können und dabei nur ein Ausfall zu erwarten** wäre.

2.1.4 Praxisbezug

MTTF und AFR sind hochgerechnete Erwartungswerte und unterstellen damit ein gleichmäßiges Ausfallverhalten. Beobachtungen in der Praxis zeigen aber, dass ein relativ großer Anteil der HDDs nicht wie erwartet, sondern wesentlich früher oder später ausfällt.

2.1.4.1 Ausfallrate - Eine Studie der Google Inc.

In ihrer ausführlichen Studie zu HDD Fehlertrends werten Pinheiro, Weber und Barroso (Google Inc.) mehr als 100.000 HDDs in einem Zeitraum von 9 Monaten aus. Die Testgruppe umfasst HDDs mit 80 bis 400 GB, die in verschiedenen Systemen der Google Inc.
eingesetzt waren [4].

Die Autoren beachten eine tendenzielle Uniformität bezüglich der Festplattenmodelle in bestimmten Altersgruppen. Dies könnte die absoluten AFR leicht beeinflussen, ändert aber nichts am feststellbaren Trend.
Die Festplatten wurden nach ihrem Alter in Gruppen eingeteilt. Die mit dem entsprechenden Alter ausgefallen Festplatten wurden dann in Relation mit der zugehörigen Gruppe gesetzt. [4].

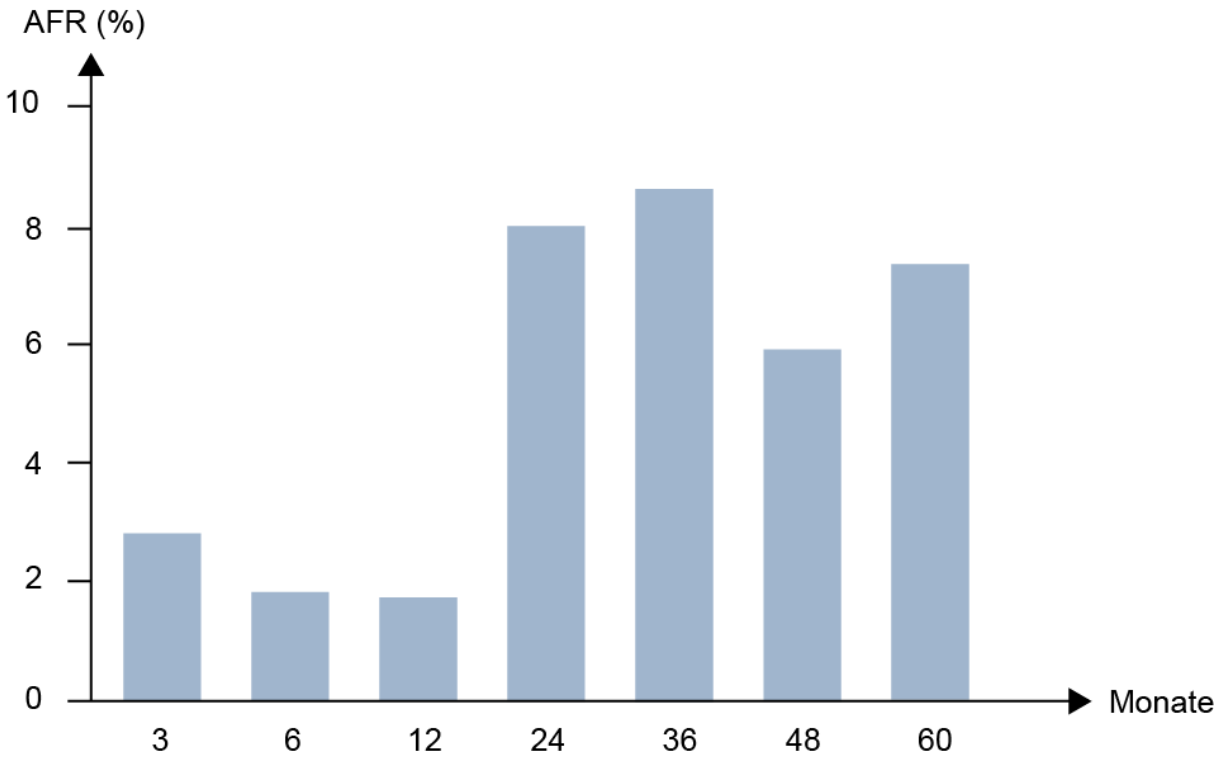

Abb. 2-33 AFR und Altersgruppen (Google Inc.)

Die **Ausfallraten** liegen zwischen 1,7% für HDDs, die mit einem Alter von 1 Jahr ausgefallen sind und 8,6% für HDDs, die mit einem Alter von 3 Jahren ausgefallen sind. Damit sind die beobachteten **AFR durchweg wesentlich höher als die Angaben der Hersteller**.

Interessant ist aber auch die Feststellung, dass ein relativ **hoher Anteil der HDDs** bereits sehr früh ausfällt, mit 3 Monaten (ca. 2,3%) oder 6 Monaten (ca. 1,9%). Dies zeigt bereits das Phänomen der „Kindersterblichkeit“ von Festplatten, auf das im Folgenden eingegangen wird.

2.1.4.2 Ausfallzeitpunkt - Eine Studie der Carnegie Mellon University

Schroeder und Gibson von der Carnegie Mellon Universität finden in ihrer Studie „Disk failures in the real world“ ähnliche Ergebnisse. Sie werten die Daten von ca. 100.000 Festplatten aus, die in verschiedenen großen Systemen eingesetzt sind [6].

Auch sie stellen eine große Abweichung der Herstellerangaben (0,58% bis 0,88%) von den beobachteten Ausfallraten (durchschnittlich ca. 3% bis 13% bei einzelnen Anlagen) fest.
Die **durchschnittliche Ausfallrate** aller Festplatten ist **3,4 mal höher als die maximal spezifizierte AFR** von 0,88% [6]. Für Systeme mit einer Laufzeit von unter 3 Jahren ist die Ausfallrate 6 mal höher, für Systeme mit einer Laufzeit von 5-8 Jahren sogar 30 mal höher [6].

Die Autoren stellen fest, dass eindimensionale Werte wie MTTF und AFR die Beobachtungen nicht abbilden. Sie betrachten daher ausführlich die zeitliche Verteilung der Ausfälle.

Zunächst verweisen sie auf die allgemein anerkannte **Theorie der Badewannenkurve**. Diese Kurve bildet die theoretische Ausfallrate von Hardwareprodukten über den gesamten Produktlebenszyklus ab [6] und könnte eine Vorhersage der Ausfalltendenz von Festplatten in einer großen Anlage ermöglichen.

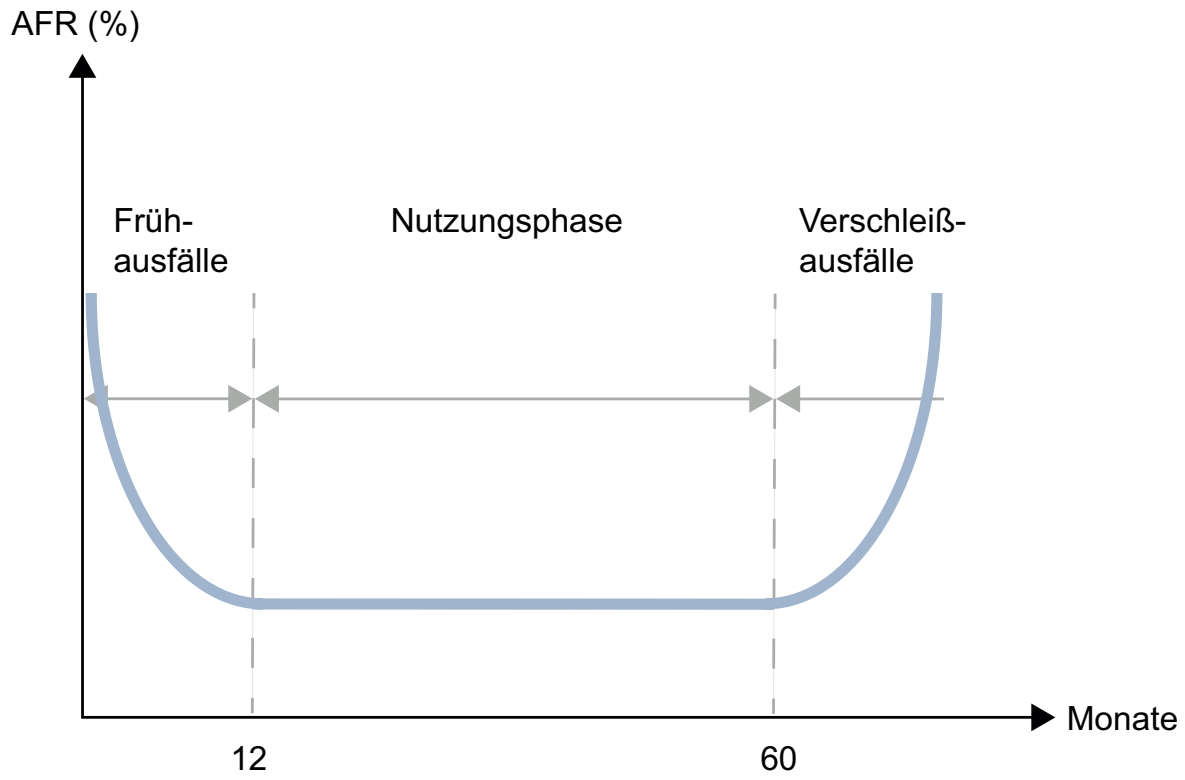

Abb. 2-4

Für Festplatten müsste demnach eine erhöhte Ausfallrate im ersten Jahr beobachtbar sein, gefolgt von einer Periode mit einer Ausfallrate auf konstantem und niedrigerem Niveau. Gegen Ende des Produktlebenszyklus wirkt sich die Abnutzung stark aus, was wieder zu einer stark ansteigenden Ausfallrate führen würde.

Diese **theoretische Überlegung** konnte in der **Praxis nur teilweise bestätigt** werden. Die folgende Grafik der monatlichen Ausfallverteilung in einer der ausgewerteten Anlagen zeigt **relativ scharf abgegrenzte Frühausfälle** (Kindersterblichkeit).

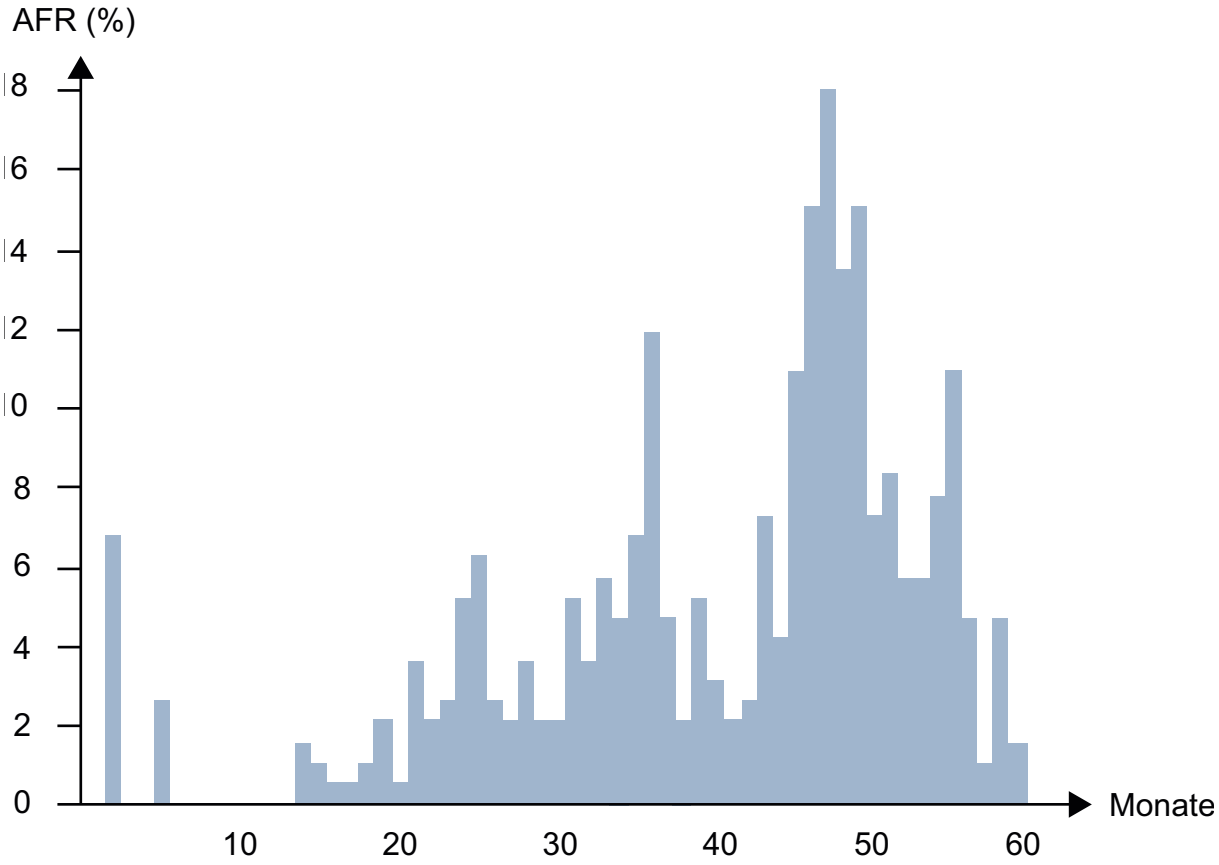

Abb. 2-5

Auffallend ist aber, dass sich die Ausfallrate in den mittleren Jahren nicht auf einen relativ niedrigen Wert einpendelt. Die AFR beginnt frühzeitig und relativ konstant zu steigen. Diese Beobachtung lässt den Schluss zu, dass sich **Abnutzung frühzeitig auswirkt** und die Ausfallrate bis zum Ende des Produktlebenszyklus linear steigt [6].

2.1.4.3 Ausfallrate - Eine Auswertung von Dallmeier electronic

Die Feststellungen der Google Inc. Studie konnten auch durch eine interne Auswertung von Dallmeier electronic bestätigt werden. Dabei wurden die Ausfälle von Festplatten ausgewertet, die zwischen Januar 2000 und Dezember 2005 in Verkehr gebracht wurden (74.000 Stück). Zunächst wurden die monatlichen Ausfälle für Wavelet- und MPEG-Systeme ermittelt.

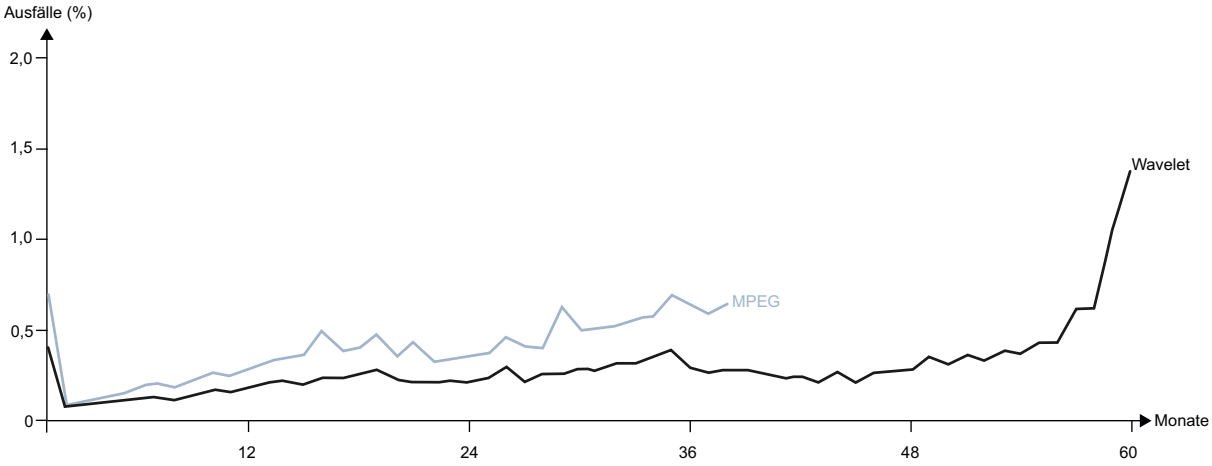

Abb. 2-6

Im Gegensatz zu der von den Herstellern durchschnittlich angegeben MTTF von 550.000 Stunden wurde ein tatsächlich beobachteter Mittelwert (87% Wavelet mit 247.000, 13% mit MPEG 111.000 Stunden) von 220.000 Stunden errechnet. Dies entspricht einer tatsächlich beobachteten AFR von ca. 3,9 %.

Basierend auf dieser Auswertung wurde die Verfügbarkeit einer Anlage mit 1.250 MPEGKanälen (auf DIS-2 Module mit insgesamt 2.500 Festplatten) betrachtet. Hierbei müsste rechnerisch mit ca. 200 Festplattenausfällen pro Jahr gerechnet werden. Für den Austausch eines DIS-2 Moduls mit defekter HDD wurden max. 2 Minuten unterstellt. Die Verfügbarkeit ergibt sich aus:

([Gesamtbetriebszeit - Reparaturzeit] / Gesamtbetriebszeit) x 100%
([8.760 Std. x 60 Min. x 1.250 Kanäle - 200 x 2 Min.] / Gesamtbetriebszeit) x 100%
([657.000.000 - 400] / 657.000.000) x 100%
99,99994% Verfügbarkeit

Bezogen auf einen Kanal dieser Anlage ergibt eine Verfügbarkeit von 99,99994% eine jährliche Uptime von 525.599,68 Minuten und eine **Downtime von nur 19,2 Sekunden** (0,32 Minuten).

Wie vorteilhaft dieser Wert ist, lässt sich anhand eines Beispiels mit einer Verfügbarkeit von nur 99.5% erkennen. Hierbei müsste mit einer Uptime von nur 522.972 Minuten und einer Downtime von bereits 2.628 Minuten (43,8 Stunden) gerechnet werden.

Wie auch dieses Beispiel verdeutlicht, ist die von den Herstellern spezifizierte MTTF wesentlich höher als die beobachtete MTTF. Dennoch kann ein **System mit ausgereifter Technik und intelligentem Aufbau hochverfügbar** realisiert werden.

2.1.5 Fazit

Die von den Herstellern durch die Angabe der **MTTF/MTBF oder AFR s**pezifizierte Ausfallrate ist **in der Regel** zu gering. Eine konservative Planung sollte eine im Durchschnitt **mindestens 3 mal höhere AFR berücksichtigen**.

Die MTTF/MTBF geben keinen Aufschluss über die zeitliche Verteilung der Ausfälle. Eine konservative Planung sollte immer eine **Kindersterblichkeit in den ersten Betriebsmonaten** und vermehrte **Abnützungsausfälle gegen Ende des Produktlebenszyklus** beachten.

Die durch die theoretische Badewannenkurve implizierte geringe und konstante Ausfallrate in den mittleren Betriebsjahren kann in der Praxis nicht bestätigt werden. Eine konservative Planung sollte bereits **ab der Mitte des Produktlebenszyklus** eine **linear ansteigende Ausfallrate** aufgrund von Abnützung berücksichtigen.

2.2 SMART

2.2.1 Definition

SMART (Self-Monitoring, Analysis and Reporting Technology / System zur Selbstüberwachung, Analyse und Statusmeldung) ist ein Industriestandard der in nahezu alle Festplatten integriert ist. SMART ermöglicht das permanente Überwachen wichtiger Parameter und damit das frühzeitige Erkennen eines drohenden Ausfalls der Festplatte [11]. Als Funktion muss SMART für jede Festplatte im BIOS aktiviert werden. Die zur Verfügung gestellten SMART-Werte werden von einer zusätzlich zum Betriebssystem installierten Software ausgewertet. Die Software kann Warnungen bei Überschreitung der herstellerspezifischen Schwellenwerte einzelner Parameter anzeigen. Nach längerer Betriebszeit können auch zu erwartende Ausfälle prognostiziert werden [11].

2.2.2 Interpretation

SMART liefert Werte für eine Vielzahl von Parametern, von denen Pinheiro, Weber und Barroso in ihrer Studie für die Google Inc. nur vier als signifikant für eine Ausfallsprognose erachtet haben [4].

Scan Errors
Festplatten prüfen als Hintergrundfunktion ständig die Oberfläche der Magnetscheiben und zählen die entdeckten Fehler. Eine hohe Anzahl an Fehlern ist ein Indikator für eine defekte Oberfläche und damit für geringere Zuverlässigkeit.

Reallocation Counts
Wenn die Festplatte einen defekten Sektor auf der Magnetscheibe findet (während eines Schreib-/Lesevorgangs oder mit Hintergrundfunktion), wird die entsprechende Sektor- Nummer einem neuen Sektor aus der Sektor-Reserve zugeordnet. Eine hohe Anzahl an neuen Zuordnungen ist ein Indikator für die Abnutzung der Magnetscheiben.

Offline Reallocations
Offline Reallocations sind eine Untergruppe der oben beschriebenen Reallocation Counts. Gezählt werden nur neue Zuweisungen von defekten Sektoren, die durch eine Hintergrundfunktion gefunden werden. Defekte Sektoren und Zuweisungen die während eines Schreib-/Lesevorgangs entdeckt werden sind nicht berücksichtigt.

Probational Counts
Festplatten können verdächtige Sektoren „auf Bewährung“ setzen, bis sie permanent ausfallen und neu zugeordnet werden oder weiterhin ohne Problem funktionieren. Eine hohe Anzahl an vorgemerkten Sektoren kann als ein schwacher Fehlerindikator betrachtet werden. Wann und ob die Werte eines Parameters eine Warnung der überwachenden Software auslösen, hängt von der Software und den Spezifikationen der Hersteller ab. Zur Veranschaulichung dieser komplexen Auswertung wird im Folgenden das stark vereinfachte und reduzierte Beispiel einer 250 GB SATA-Festplatte herangezogen [11].

Parameter	Value (normalisierter aktueller Messwert)	Worst (bisher schlechtester Wert)	Threshold (Grenzwert, Value sollte größer sein)	RAW Value (eigentlicher Messwert)	Bemerkung
Relocation Counts	100	100	005	55	55 Sektoren wurden wegen Defekts gegen Reserve-Sektoren ausgetauscht. Das Laufwerk schätzt das aber noch als problemlos ein (der Value ist nach wie vor 100).
Seek Errors	100	100	067	0	Bisher gab es keine Schreib-/Lesefehler.

Tabelle 2-1

Der normalisierte Messwert **Value** wird rückwärts gezählt und löst bei Erreichen des Grenzwertes **Threshold** eine Warnung aus. Obwohl in diesem Beispiel bereits 55 Sektor- Zuweisungen erfolgt sind, wird die Festplatte noch als absolut in Ordnung betrachtet.

Unabhängig von der Auswertung der Werte durch die SMART Software, aber maßgeblich für eine zuverlässige Ausfallprognose, ist die Fehlererkennung durch die SMART Funktion der Festplatte. Wenn die Erkennung nicht zuverlässig funktioniert, kann SMART nicht als alleiniges Instrument zur Ausfallprognose von Festplatten verwendet werden.

2.2.3 Praxisbezug

In ihrer Studie für die Gogle Inc. werten die Autoren SMART Log-Dateien von mehr als 100.000 HDDs aus [4]. Dennoch konnten sie **kein aussagekräftiges statistisches Modell** zur Ausfallvorhersage entwickeln [4].

Im Folgenden wurde die Möglichkeit betrachtet, ein **einfacheres Prognosemodell** allein auf der Basis von SMART-Parameter zu erstellen. Aber die Auswertung der entsprechenden SMART-Werte zeigte, dass keine ausreichende Genauigkeit erreicht werden konnten.

Von allen ausgefallenen Festplatten zeigten **56% keinen erkannten Fehler** bei allen vier starken SMART-Parametern. Eine Prognose auf dieser Basis könnte also nie mehr als die Hälfte der Ausfälle vorhersagen. Selbst unter Einbeziehung aller anderen SMART-Parameter zeigten 36% der ausgefallenen Laufwerke überhaupt keinen Fehler an (bei keinem Parameter!).

2.2.4 Fazit

Das Fazit der Autoren ist eindeutig: „We conclude that it is unlikely that SMART data alone can be effectively used to build models that predict failures of individual drives.“ [4]. **SMART-Daten** alleine können also **nicht verwendet** werden, um **Ausfälle einzelner Festplatten vorherzusagen**.

Hohe Werte eines einzelnen Parameters können einen unnötigen Austausch und damit Kosten verursachen. Plötzliche Ausfälle ohne vorherige Meldung könnten zu Datenverlust aufgrund fehlender Backups führen. Die Folge wäre ein Zweifel an der Zuverlässigkeit des Gesamtsystems, obwohl eigentlich nur die SMART-Funktion versagt hat.

Als Alternative verbleibt eine konservative Wartungsplanung basierend auf den Feststellungen unter Punkt 1. Für Systeme mit mehreren Festplatten kann zudem eine gewisse Absicherung durch die Verwendung von RAID-Systemen erreicht werden, wie im Folgenden beschrieben.

SPEICHERTECHNOLOGIEN

Standard-Aufzeichnungssysteme verfügen normalerweise über eine oder mehrere Festplatten (JBOD) und können den Ausfall einer HDD nicht abfangen.

Hochwertige Aufzeichnungssysteme speichern die Audio- und Videodaten unter Verwendung einer speziellen Speichertechnologie (RAID) redundant auf mehrere HDDs und können den Ausfall einer HDD in der Regel ohne Datenverlust kompensieren.

Unabhängig von der eingesetzten Speichertechnologie muss aber immer beachtet werden, dass kein System einen Ersatz für ein Backup (Datensicherung) darstellt [5].

3.1 JBOD

JBOD (just a bunch of disks / nur ein Haufen Festplatten) bezeichnet eine beliebige Anzahl von Festplatten (Array), die an ein System (Computer, Aufzeichnungssystem) angeschlossen sind. Sie können vom Betriebssystem als individuelle Laufwerke oder zusammengefasst zu einem einzigen logischen Laufwerk genutzt werden [12].

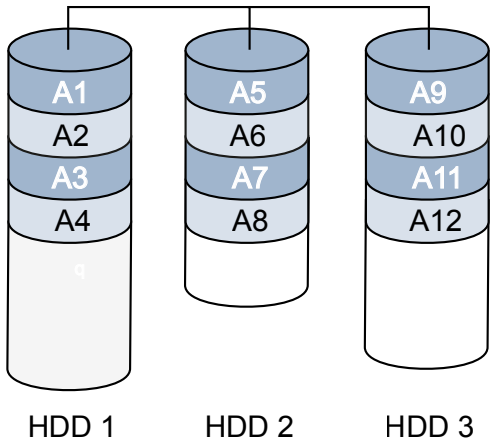

Abb. 3-1

Einem JBOD fehlt jegliche Redundanz, weshalb der Begriff oft zur Abgrenzung normaler Systeme von RAID-Systemen verwendet wird.

3.1.1 Kapazität und Kosten

Die Nettokapazität eines JBOD-Arrays ist so groß wie die Summe der Kapazitäten der einzelnen Festplatten. Die Nettokapazität entspricht also der Gesamtkapazität eines Systems. Ein System aus 8 Festplatten mit jeweils 2 TB kommt auf eine Nettokapazität von 16 TB. Ein JBOD-System ist in Bezug auf die Speicherkosten am günstigsten.

3.1.2 Sicherheit und Rebuild

Das Verhalten beim Ausfall einer Festplatte variiert bei den JBOD-Systemen verschiedener Hersteller. Dallmeier JBOD-Aufzeichnungssyteme bieten den Vorteil die Aufzeichnung fortzusetzen, wenn eine HDD ausfällt. Die Aufzeichnungen auf den verbleibenden Festplatten können nach wie vor ausgewertet und gesichert werden.

3.1.3 Fazit

JBOD ist ein einfaches und sehr kostengünstiges Speichersystem. Wenn eine einzelne Festplatte ausfällt gehen deren Aufzeichnungen aber verloren.

3.2 RAID 1

Ein RAID 1-System besteht aus einem Verbund von mindestens zwei Festplatten (RAIDArray). Die gleichen Daten werden simultan auf allen Festplatten (Spiegelung) gespeichert. Ein RAID 1-System bietet volle Redundanz [13].

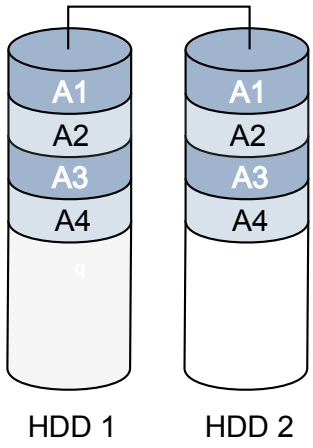

Abb. 3-2

3.2.1 Kapazität und Kosten

Die Nettokapazität eines RAID 1-Arrays ist so groß wie die kleinste Festplatte.
Durch die Spiegelung der Daten wird die Gesamtkapazität eines Systems im Idealfall halbiert, die Speicherkosten also verdoppelt. Ein System aus 8 Festplatten mit jeweils 2 TB kommt auf eine Nettokapazität von 8 TB.

3.2.2 Sicherheit und Rebuild

Wenn eine der gespiegelten HDDs ausfällt, wird die Aufzeichnung auf der verbleibenden HDD fortgesetzt. Nach dem Austausch der ausgefallenen HDD wird ein Rebuild-Prozess gestartet und die Daten werden auf die neue HDD gespiegelt.

Wenn die intakte HDD während des Austauschs oder Rebuild der defekten HDD ausfällt, kommt es unweigerlich zum Verlust der Daten (sofern nicht auf mehr als 2 HDDs gespiegelt wurde). Da bei RAID 1 nur wenige HDDs beteiligt sind, ist die Wahrscheinlichkeit für einen gleichzeitigen Ausfall relativ gering, kann aber nicht ausgeschlossen werden.

3.2.3 Fazit

RAID 1 ist ein einfaches und relativ robustes Speichersystem. Die Speicherkosten sind aber relativ hoch, da die Gesamtkapazität immer halbiert wird.

3.3 RAID 5

Ein RAID 5-System besteht aus einem Array von mindestens drei Festplatten. Die Daten werden auf allen Festplatten verteilt gespeichert. Zudem werden Paritäts-Daten erzeugt und ebenfalls verteilt gespeichert.

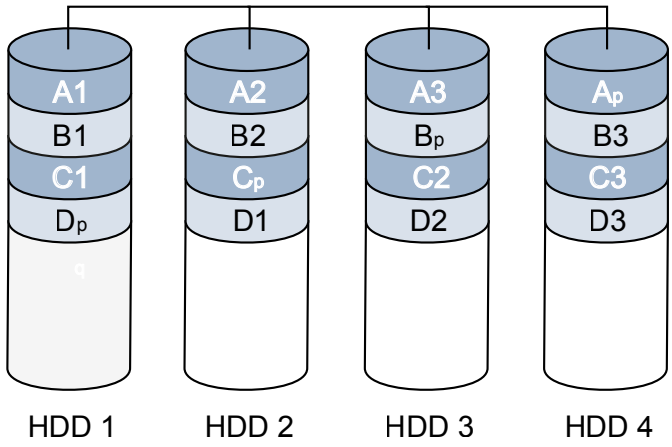

Abb. 3-3

Falls eine HDD ausfällt ermöglichen die Paritäts-Daten in Verbindung mit den verbleibenden Daten die Rekonstruktion der verlorenen Daten [14].

3.3.1 Kapazität und Kosten

Die Kapazität eines RAID 5-Arrays kann wie folgt berechnet werden:

(Anzahl der Festplatten - 1) × (Kapazität der kleinsten Festplatte)

Ein RAID 5-Array aus 8 Festplatten mit jeweils 2 TB hat also eine Nettokapazität von:

(8-1) × 2 TB= 14 TB

Wenn eine Spare-HDD verwendet wird (siehe im Folgenden), muss die Formel angepasst werden:

(Anzahl der Festplatten - 2) × (Kapazität der kleinsten Festplatte)

Im Gegensatz zu RAID 1 bietet ein RAID 5-System eine bessere Ausnutzung der Gesamtkapazität eines Systems. Damit kann eine redundante Datenhaltung bei relativ geringen Kosten realisiert werden.

3.3.2 Sicherheit und Rebuild

Wenn eine HDD ausfällt, ermöglichen die Paritäts-Daten in Verbindung mit den verbleibenden Daten die Rekonstruktion der verlorenen Daten auf einer Ersatz-HDD. Der Rebuild- Prozess startet automatisch, wenn im System bereits eine Ersatz-HDD (Spare-HDD) integriert ist. Ist dies nicht der Fall, wird er nach dem Austausch der defekten HDD gestartet.

Wenn eine weitere Festplatte ausfällt, während die defekte Festplatte ausgetauscht oder wiederhergestellt wird, kann der Rebuild-Prozess nicht abgeschlossen werden. Dies führt zum Verlust aller Daten.

Ein RAID 5 besteht in der Regel aus mehreren HDDs. Die **Ausfallwahrscheinlichkeit einer weiteren HDD steigt proportional mit der Anzahl**. Zudem muss beachtet werden, dass die Dauer eines Rebuild bei Verwendung von HDDs mit hoher Kapazität mehrere Stunden bis Tage dauern kann. Der kritische Zeitraum ist also relativ lange.

Neben dem Ausfall einer weiteren Festplatte kann auch **ein nicht korrigierbarer Lesefehle**r (unrecoverable read error, URE) das Scheitern eines Rebuild-Prozesses auslösen. Wenn ein Bruchteil der Paritäts-Daten oder der verbleibenden Daten nicht mehr lesbar ist, kann er nicht wieder hergestellt werden und der Prozess wird in der Regel gestoppt.

Die URE-Rate ist ein Druchschnittswert der von den Herstellern für ein Festplattenmodell (nicht für eine einzelne Festplatte) angegeben wird. Ein typischer **Wert von 10^{-14} Bit** bedeutet, dass es während der Verarbeitung von 100.000.000.000.000 Bits (12 TB) zu einem nicht korrigierbaren Lesefehler kommt.

Bereits bei kleineren RAID 5-Systemen (z.B. RAID 5 mit 3 × 500 GB Festplatten) führt allein die Berücksichtigung einer URE von 10^{-14} Bit zu einem statistischen **Scheitern des Rebuild-Prozesses in 8% der Fälle** [17]. Wenn größere HDDs verwendet werden, ist das Auftreten eines URE wesentlich wahrscheinlicher.

Während des Rebuilds eines RAID-Array aus 7 × 1 TB Festplatten muss der Inhalt von 6 HDDs (6 TB) gelesen werden. Bei einer URE von 10^{-14} Bit müsste das Scheitern des Rebuilds in 50% der Fälle erwartet werden [15].

3.3.3 Fazit

RAID 5 ist eine Speichertechnologie, die eine redundante Datenhaltung bei relativ geringen Kosten ermöglicht. Die Gefahr des Datenverlusts ist aber relativ hoch.

3.4 RAID 6

Ein RAID 6-System besteht aus einem Array von mindestens vier Festplatten. Die Daten werden auf allen Festplatten verteilt gespeichert. Wie bei RAID 5 werden Paritäts-Daten erzeugt und verteilt gespeichert, in diesem Fall aber doppelt.

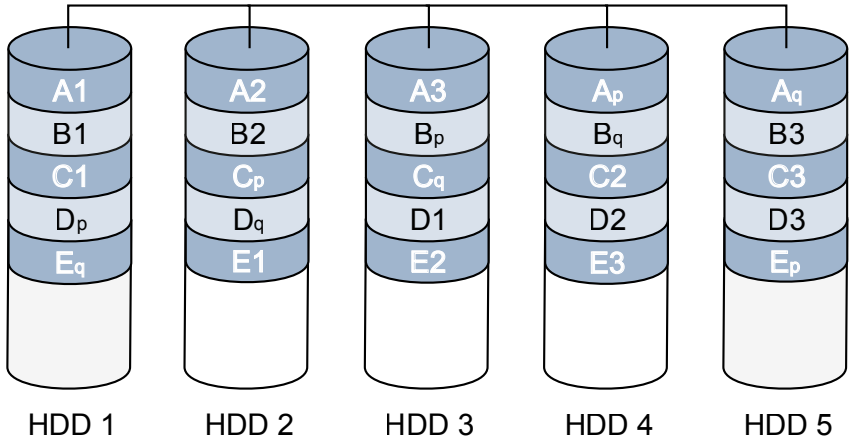

Abb. 3-4

Die doppelten Paritäts-Daten erlauben einem RAID 6, den Ausfall von bis zu zwei Festplatten abzufangen [16].

3.4.1 Kapazität und Kosten

Die Kapazität eines RAID 6-Arrays kann wie folgt berechnet werden:

(Anzahl der Festplatten - 2) × (Kapazität der kleinsten Festplatte)
Ein RAID 6-Array aus 8 Festplatten mit jeweils 2 TB hat also eine Nettokapazität von:
(8-2) × 2 TB= 12 TB

Im Gegensatz zu RAID 5 ohne Spare-HDD kann ein RAID 6-System die Gesamtkapazität eines Systems nicht ganz so gut ausnutzen. Dennoch kann eine redundante Datenhaltung bei relativ geringen Kosten realisiert werden.
Vergleicht man ein RAID 6-System allerdings mit dem in der Praxis am häufigsten verwendeten RAID 5 mit Spare-HDD, muss die Kapazitätsüberlegung relativiert werden. In diesem Fall verfügen beide Systeme über eine identische Nettokapazität (im gewählten Beispiel jeweils 12 TB) und können mit gleichen Speicherkosten realisiert werden.

3.4.2 Sicherheit und Rebuild

Generell muss bei einem RAID 6-System die Problematik von Festplattenausfällen oder nicht korrigierbaren Lesefehlern während des Rebuilds auch beachtet werden. Der große Vorteil eines RAID 6 ist aber seine Toleranz von zwei Ausfällen.
Wenn eine HDD ausfällt, ermöglichen die Paritäts-Daten in Verbindung mit den verbleibenden Daten die Rekonstruktion der verlorenen Daten auf einer Ersatz-HDD. Wenn eine weitere HDD während des Austauschs der defekten HDD oder während ihres Rebuilds ausfällt, führt dies nicht zum Verlust der Daten. Einfach ausgedrückt ermöglicht der zweite Satz an Paritätsdaten jetzt die Rekonstruktion der verlorenen Daten auf einer zweiten Ersatz-HDD.

Wie bei RAID 5 steigt die Ausfallwahrscheinlichkeit einer weiteren HDD mit deren Anzahl und der Dauer des Rebuilds, der bei RAID 6 aufgrund der doppelten Paritäts-Berechnung länger dauern kann.
Die Dauer des Rebuilds hängt bei allen RAID-Systemen von einer Vielzahl von Faktoren ab. Entscheidend sind natürlich die Anzahl und die Kapazität der Festplatten. Werden Aufzeichnungssysteme mit vergleichbarer Ausstattung betrachtet, kommt es auch auf die Art der Aufzeichnung (SD- oder HD-Kameras, permanent oder ereignisgesteuert) an und ob die Aufzeichnung fortgesetzt oder unterbrochen wird. Eine Testreihe von Dallmeier electronic mit vergleichbaren IPS-Systemen bei Vollauslastung zeigte sowohl für RAID 5 als auch für RAID 6 Systeme eine Rebuild-Dauer von ca. 2 Stunden pro TByte. Für die Praxis kaum relevant, lässt sich dennoch eine etwas längere Rebuild-Dauer bei RAID 6 Systemen erkennen. Als Faustregel kann mit 25% bis 35% gerechnet werden. Trotz etwas längerer Rebuild-Dauer hat RAID 6 den entscheidenden Vorteil der Toleranz von zwei Festplattenausfallen. Die Gefahr des Verlusts aller Daten während eines längeren Rebuilds ist also wesentlich geringer als bei RAID 5.

3.4.3 Fazit

RAID 6 ist eine sicherere Speichertechnologie, die eine redundante Datenhaltung bei immer noch relativ geringen Kosten ermöglicht. Die Gefahr des Datenverlusts ist im Vergleich zu einem RAID 5-System relativ gering. Insgesamt kann RAID 6 als das überlegene Speichersystem betrachtet werden.

EMPFEHLUNGEN

Planung

1. Beachten Sie, dass JBOD-Systeme kostengünstig sind, aber keinen Ausfallschutz für einzelne Festplatten bieten.

2. RAID 1-Systeme sind einfach und relativ robust, verursachen aber hohe Speicherkosten.

3. RAID 5-Systeme verursachen geringere Speicherkosten als RAID 1-Systeme und tolerieren den Ausfall einer Festplatte.

4. Beachten Sie, dass ein RAID 6-System die gleichen Speicherkosten wie ein RAID 5-System mit Spare-HDD verursacht, aber den Ausfall von zwei Festplatten toleriert.

5. RAID 6 ist das derzeit überlegene Speichersystem und bietet höchstmögliche Sicherheit bei vertretbaren Kosten.

6. Beachten Sie, dass kein RAID-System die gleiche Sicherheit wie ein Backup wichtiger Aufzeichnungen bietet.

Wartung

1. Planen Sie mit einer mindestens drei mal höheren Ausfallrate als von den Festplattenherstellern spezifiziert.

2. Berücksichtigen Sie die von den Festplattenherstellern spezifizierte Lebensdauer der Festplatten und tauschen Sie auch funktionierende Festplatten frühzeitig aus.

3. Berücksichtigen Sie bereits ab der Mitte des Produktlebenszyklus der Festplatten eine linear ansteigende Ausfallrate.

4. Berücksichtigen Sie vermehrte Ausfälle in den ersten Betriebsmonaten und gegen Ende des Produktlebenszyklus der Festplatten.

5. Beachten Sie, dass SMART-Daten nicht zur Prognose des Ausfalls einzelner Festplatten geeignet sind.

Backup

1. Kein RAID-System biete die gleiche Sicherheit wie ein Backup (Datensicherung). Wichtige Aufzeichnungen sollten immer durch ein Backup gesichert werden.

2. Backups können bequem mit der Software SMAVIA Viewing Client durchgeführt werden. Manche Aufzeichnungssysteme bieten die Möglichkeit alle Festplatten zu entnehmen. Sie können an einem sicheren Ort gelagert und später erneut angeschlossen werden.

3. Nur Backups bieten eine wirksame Sicherung der Aufzeichnungen in Fällen wie:
   - Defekte Dateien aufgrund von Speicherfehlern
   - Versehentliches Löschen von Aufzeichnungen
   - Diebstahl des Aufzeichnungssystems
   - Katastrophen wie Feuer, Wasserschaden, etc.
   - Systemstörungen durch defekte Komponenten oder Ausfall des RAID-Controllers

REFERENZEN

[1] Verschiedene Autoren, Hard disk drive, in http://en.wikipedia.org/wiki/Hard_disk_drive (2012.09.03)

[2] Verschiedene Autoren, RAID, in http://en.wikipedia.org/wiki/RAID (2012.09.03)

[3] Verschiedene Autoren, RAID, Data backup in http://en.wikipedia.org/wiki/RAID (2012.09.03)

[4] Eduardo Pinheiro, Wolf-Dietrich Weber, Luiz André Barroso (Google Inc.), in Failure Trends in a Large Disk Drive Population (Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST'07), Februar 2007

[5] Christopher Negus, Thomas Weeks. The Mythos of RAID Backups, in Linux Troubleshooting Bible, Seite 100, Wiley Publishing Inc., 2004

[6] Bianca Schroeder, Garth A. Gibson (Computer Science Department, Carnegie Mellon University), Age-dependent replacement rates, in Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you? (5th USENIX Conference on File and Storage Technologies, San Jose, CA), Februar 2007

[7] Verschiedene Autoren, Mean Time Between Failures, in http://de.wikipedia.org/wiki/Mean_Time_Between_Failures (2012.08.16)

[8] Verschiedene Autoren, Mean Time To Failures, in http://de.wikipedia.org/wiki/MTTF (2012.08.16)

[9] Gerry Cole. (Seagate Personal Storage Group), Estimating Drive Reliability in Desktop Computers, Longmont Colorado, November 2000

[10] Adrian Kingsley-Hughes, Making sense of „mean time to failure" (MTTF), in http://www.zdnet.com/blog/hardware/making-sense-of-mean-time-to-failure-mttf/310 (2012.08.20)

[11] Verschiedene Autoren, Self-Monitoring, Analysis and Reporting Technology, in http://de.wikipedia.org/wiki/Self-Monitoring,_Analysis_and_Reporting_Technology (2012.08.22)

[12] Verschiedene Autoren, RAID / JBOD, in http://de.wikipedia.org/wiki/RAID (2012.08.28)

[13] Verschiedene Autoren, RAID / RAID 1: Mirroring – Spiegelung, in http://de.wikipedia. org/wiki/RAID (2012.08.16)

[14] Verschiedene Autoren, RAID / RAID 5: Leistung + Parität, Block-Level Striping mit verteilter Paritätsinformation, in http://de.wikipedia.org/wiki/RAID (2012.08.16)

[15] Robin Harris, Why RAID 5 stops working in 2009, in http://www.zdnet.com/blog/storage/why-raid-5-stops-working-in-2009/162 (2012.08.31)

[16] Verschiedene Autoren, RAID / RAID 6: Block-Level Striping mit doppelter verteilter Paritätsinformation, in http://de.wikipedia.org/wiki/RAID (2012.08.16)

[17] Verschiedene Autoren, RAID / Statistische Fehlerrate bei großen Festplatten, in http://de.wikipedia.org/wiki/RAID (2012.08.31)

HEAD & ACCOUNTS OFFICE

Dallmeier electronic GmbH & Co.KG
Bahnhofstr. 16
93047 Regensburg
Germany

tel +49 941 8700 0
fax +49 941 8700 180
mail info@dallmeier.com

www.dallmeier.com