

Systemverwaltung (Vorl. und Blockveranstaltung) SS 2012

05.08. bis 23.08.2013

Arnold Kühnel,
Rolf Dietze



Systemverwaltung

Organisation

- Veranstalter
A. Kühnel, R. Dietze
- 05.08.-23.08.2013 (2x3h täglich)

1 Einleitung

- 1. Einleitung (Arnold)
 - RAID
- 2. Shells und mehr
 - Bash, sed, awk, vi, scripting, Powershell (Rolf)
- 3. Verschiedenes (Rolf)
 - IBM i5, AIX



1 Einleitung

- Überblick
- Teilnahmebestätigung
- Teilnehmerliste
- Scheinanforderungen
- Übersicht und Unterlagen

<https://www.mi.fu-berlin.de/w/Tec/ArtLehreSystemverwaltung2013>

2. Einführung

2.1 Entwicklungen

- Zur ausgewogenen/optimalen Leistungssteigerung unter Kostengesichtspunkten müssen alle „hinreichend“ performant sein.
- Berühmte „Gesetze“ der Prognostizierung
 - Gordon Bell (1984):

Ein-Chip-Computer haben ihre Leistung innerhalb von 10 Jahren ((1974-1984) um 40% jährlich gesteigert. Das ist die doppelte Rate verglichen mit den Minicomputer.
 - Bill Joys Gesetz (1985):

Ein noch stärkeres Wachstum für die Ausführungsgeschwindigkeit prognostiziert Bill Joys in der Größenordnung

$$MIPS = 2^{\text{Jahr}-1984}$$

Prognostiziert. Die Komplexität integrierter Schaltkreise mit minimalen Komponenten-kosten verdoppelt sich etwa alle zwei Jahre.

– Gene Amdahls Gesetz (1967):

- Beschleunigung von Programmen durch parallele Ausführung
- Der Beschleunigungsfaktor hängt vom parallel ausführbaren Anteil ab

$$S = \frac{1}{(1-p) + pk} \text{ Speedup}$$

wobei

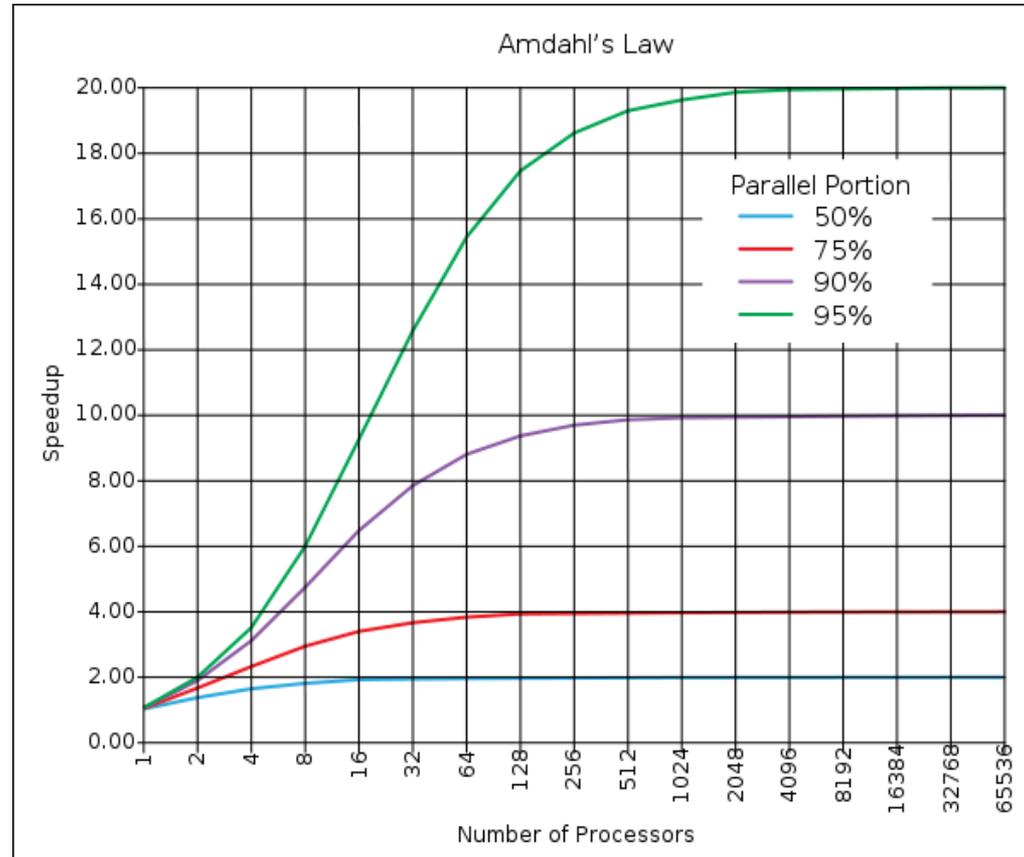
$p =$ Anteil parallel
ausführbaren Codes,

$k =$ # CPUs

- Pro CPU-Instruktion/s wird 1 Byte Hauptspeicher benötigt.

– Gordon Moores Gesetz (1965):

- $\#Transistoren = 2^{\text{Jahr}-1964}$

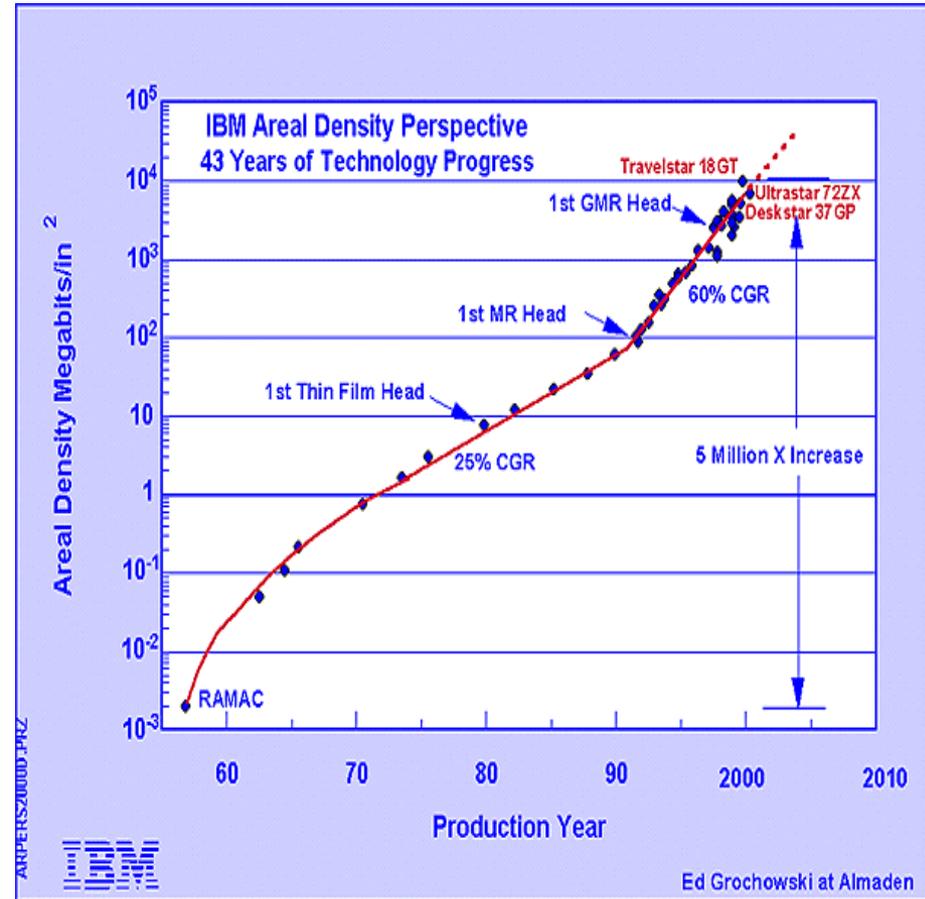


- Berühmte „Gesetze“ der Prognostizierung (Forts.)
 - Mark Kryders Gesetz (1967):
 - Informationsdichte von 2000 bits (1956) auf 100 Gigabits (2005), entspricht einer Steigerung auf das 50-Millionen-fache
 - Einschätzung: Die Informationsdichte einer Festplatte ist ausschlaggebender für neue Anwendungen als Fortschritte in der Halbleitertechnologie.
 - Die Informationsdichte verdoppelt sich jährlich.
 - ❖ Niklaus Wirths Gesetz (1995):
 - SW wird in kürzeren Abständen langsamer als die HW schneller
Schnellere HW bewirkt keine schnellere Abarbeitung von Aufgaben, da gleichzeitig die durch gestiegenen Anforderungen erhöhte Komplexität der Software deutlich mehr Leistung benötigt wird.

- **Folgerungen aus dem Amdahlschen Gesetz**
 - Quantitativer Aspekt: Große Verbesserungen im Mikroprozessorbereich verbessern die Gesamtperformanz nur, wenn gleichzeitig der Sekundärspeicher verbessert wird.
 - Qualitativer Aspekt: Schnelle Mikroprozessoren ermöglichen neue Anwendungen in der Bildverarbeitung, Videos, Hypertextanwendungen, Explorationsauswertungen.
Existierende Anwendungen können mehr Daten verwenden.
- **Illustrationen zu Kryders Gesetz**

Speicherichte

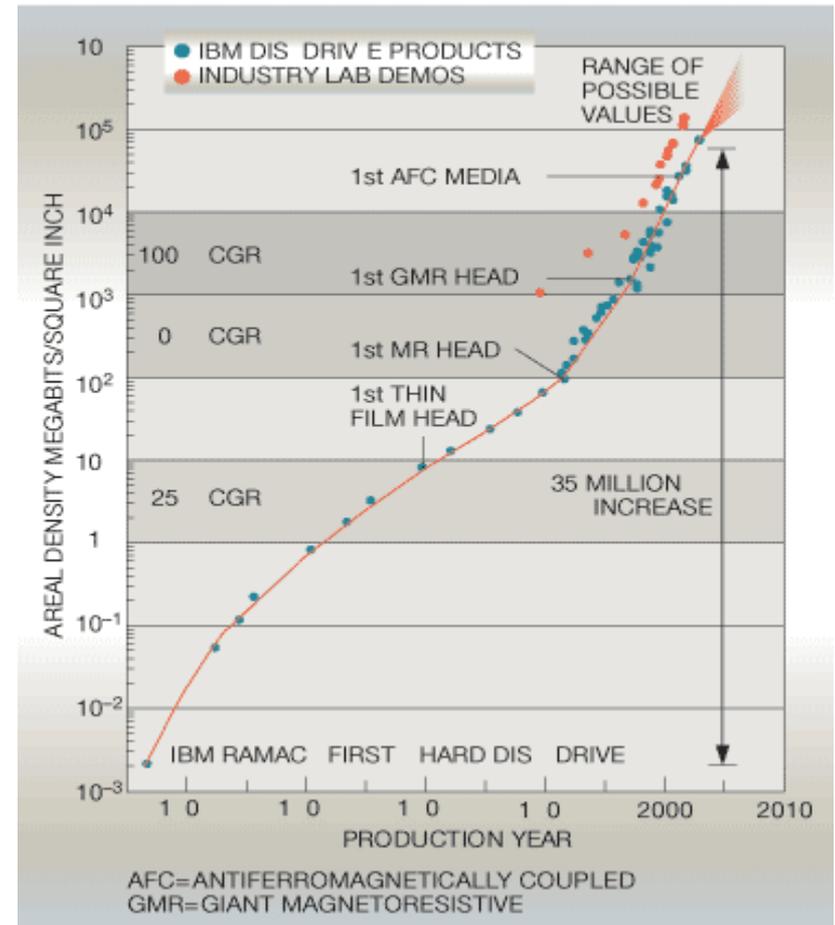
- Steigerung der Speicherdichte von
- 1991 bis 2006 um das 10.000-fache
- 1956: IBM 305 RAMAC: Erster kommerzieller Rechner mit Festplatten mit beweglichem Kopf



Plattenkapazität

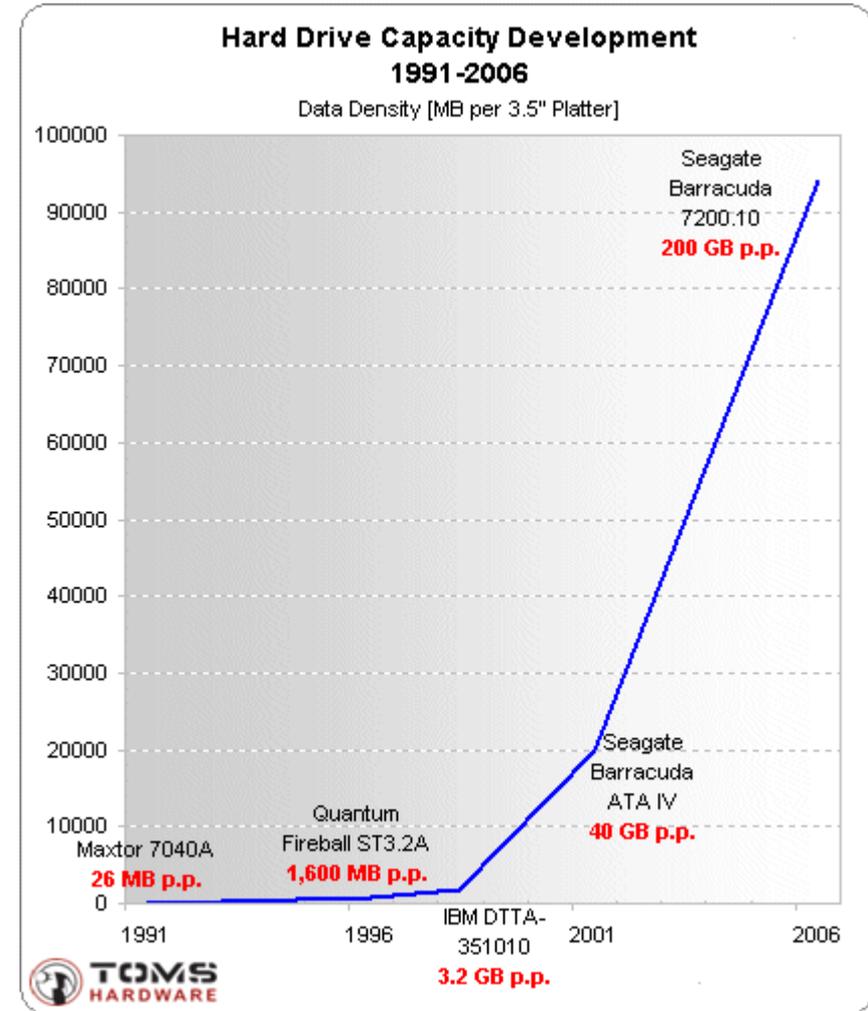
- 1991: 40 MB, 130 MB maximal
 - 2006: 750 GB pro HD
 - Steigerung auf das 5.700-fache in 15 J.
-
- Magneto-Resistive (MR) Heads
 - 2000: Giant Magneto-Resistive (GMR) Heads
 - Perpendicular Magnetic Recording
 - 2001: Hitachi Antiferromagnetically-coupled (AFC)Media

Figure 1 Hard disk drive areal density trend



Plattenkapazität

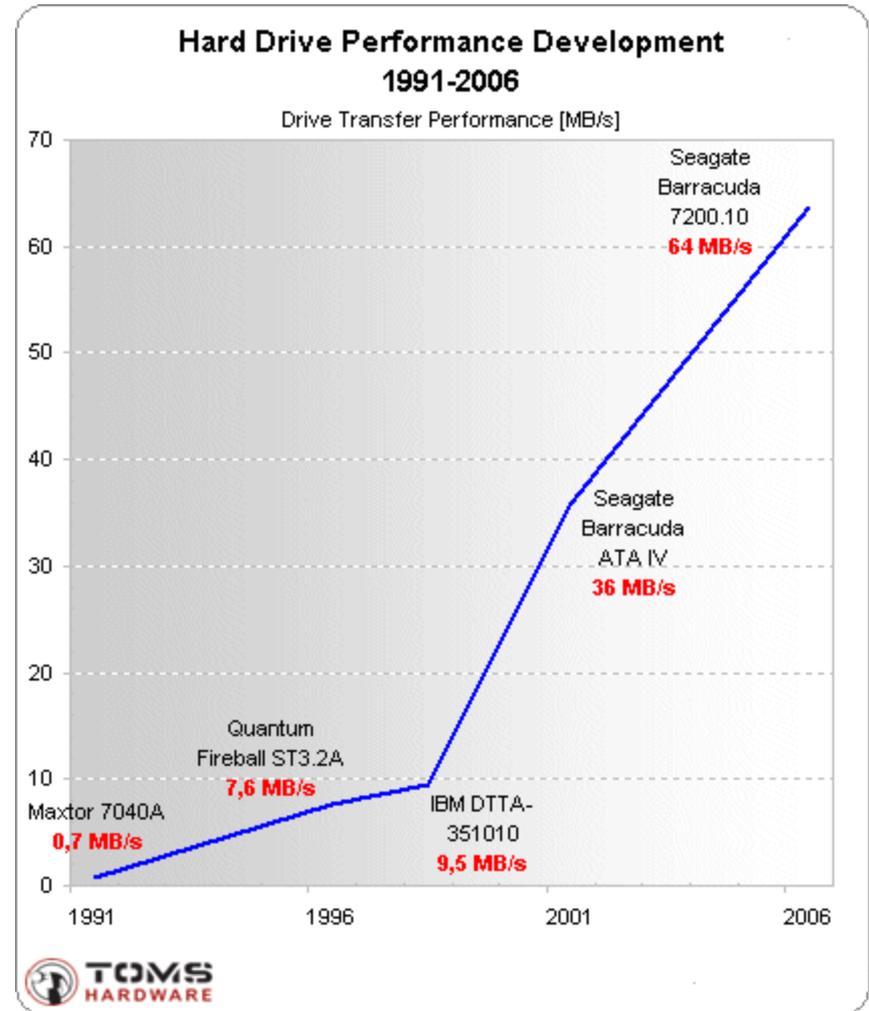
- Die Speicherdichte stieg noch stärker an.
- 2011: 2 – 3 TB am Markt
- 2012: 4 TB erschienen





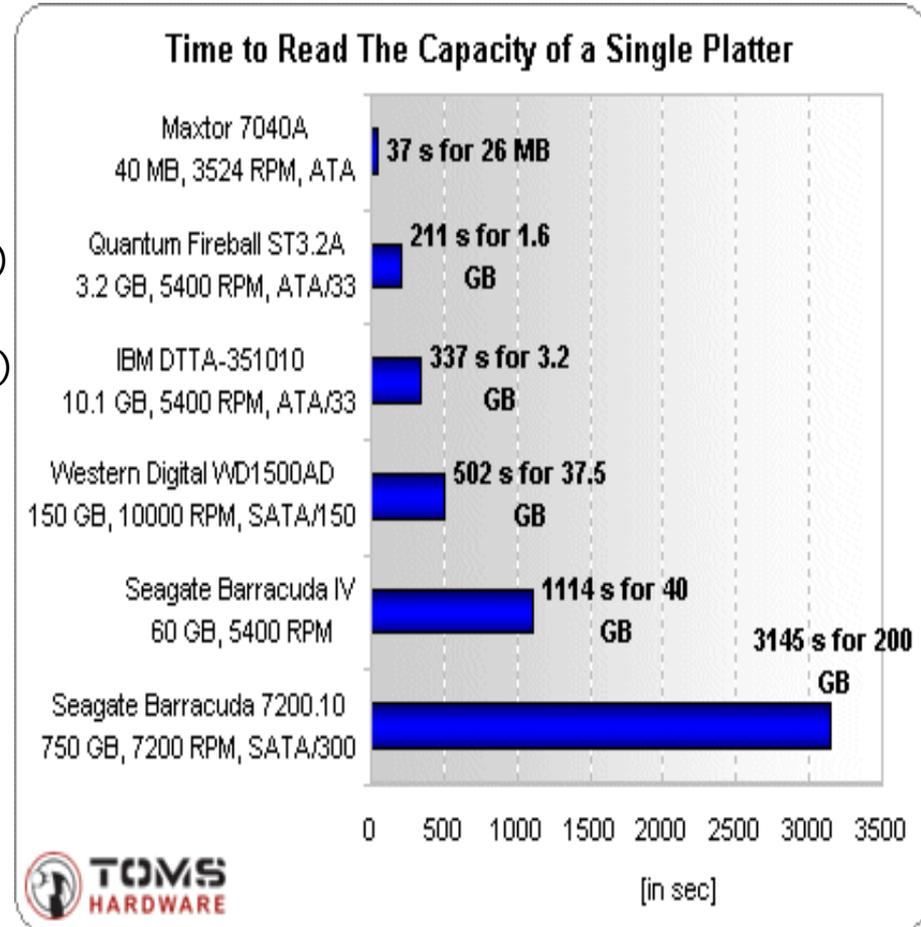
Plattendurchsatz

- 1991: 0,7 MB/s Maxtor 7040A
- 2006: 64 MB/s Barracuda 7200.10
- Steigerung um das 91-fache in 15 J.
- WD Raptor mit 85 MB/S entspricht einer 121-fachen Steigerung



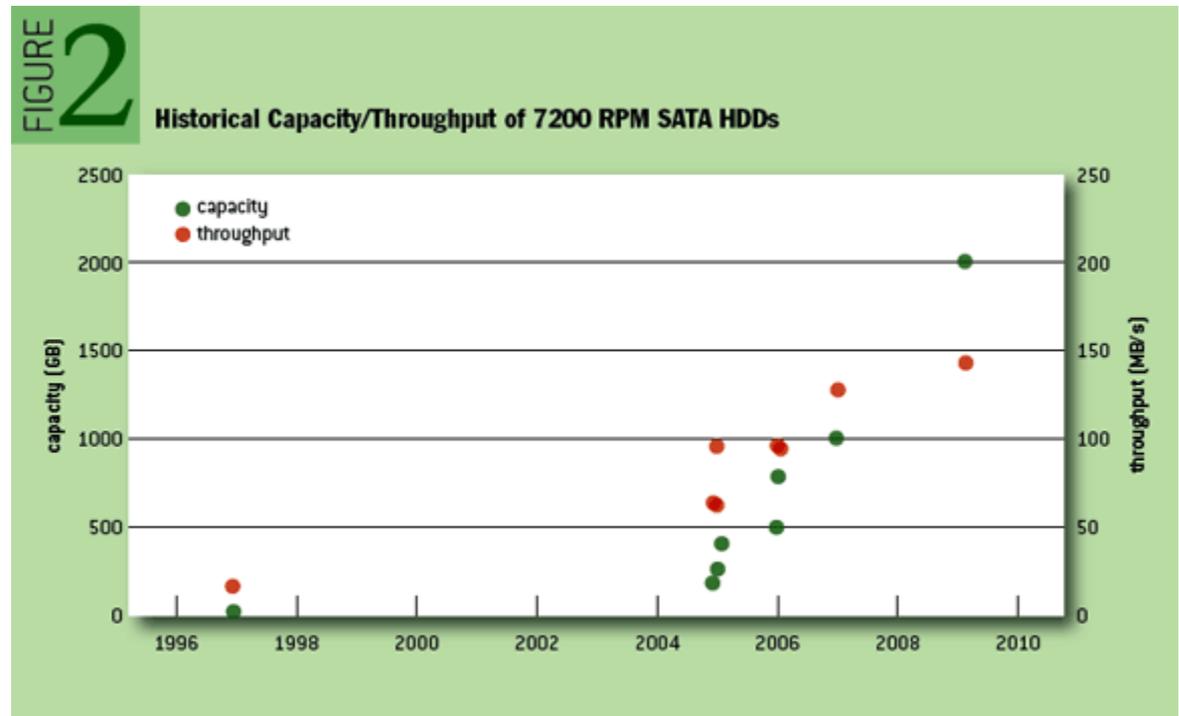
Plattendurchsatz

- 1991: 37s für 26 MB (eine von zwei Scheibe einer 40 MB HD)
- 1998: 211s für 1.6 GB (3.2 GB HD)
- 1999: 337s für 3.2 GB (10 GB HD, 3 Sch.)
- 2004: 18min 34s für 40 GB (60 GB HD, 2 Scheiben, 3 Köpfe)
- 2006: 52min für 200 GB (750 GB HD)



Kapazität/Durchsatz

- Die Kapazitäten steigen signifikant, ebenso wurden die
- Bitfehlerrate deutlich verbessert, nur der
- Durchsatz hinkt hinterher
- Vgl. [4]



Kapazität/Durchsatz

- für 10k und 15k rpm
- Vgl. [4]

FIGURE 3

Historical Capacity/Throughput of 10k RPM FC HDDs

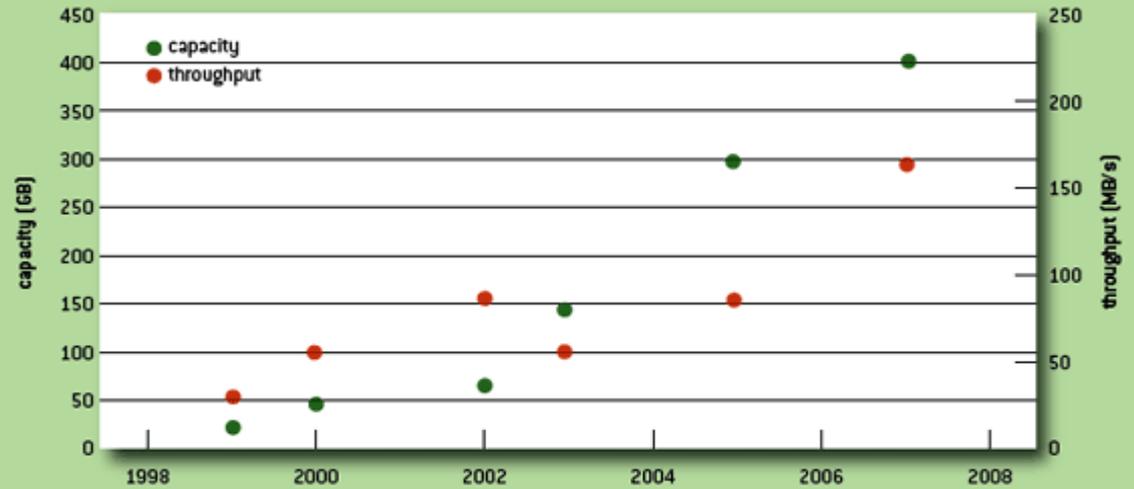
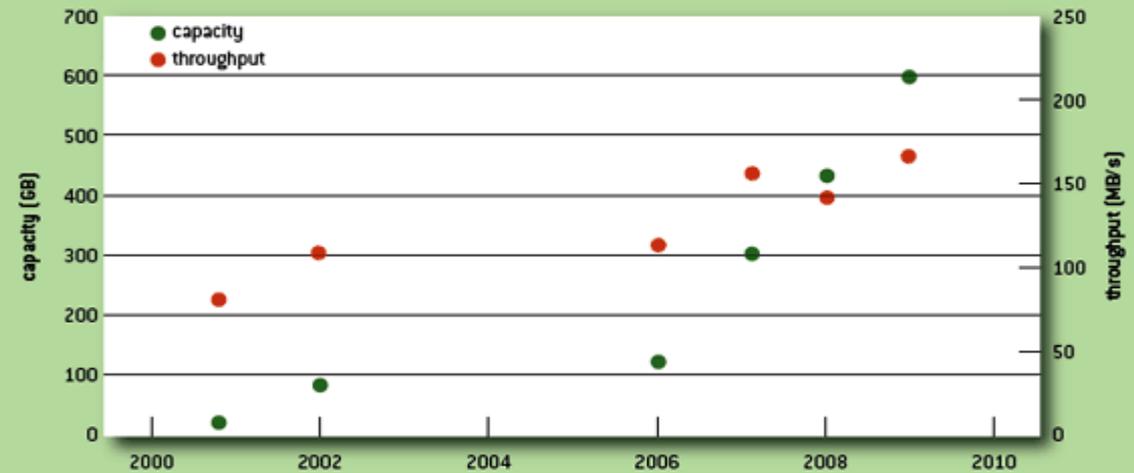


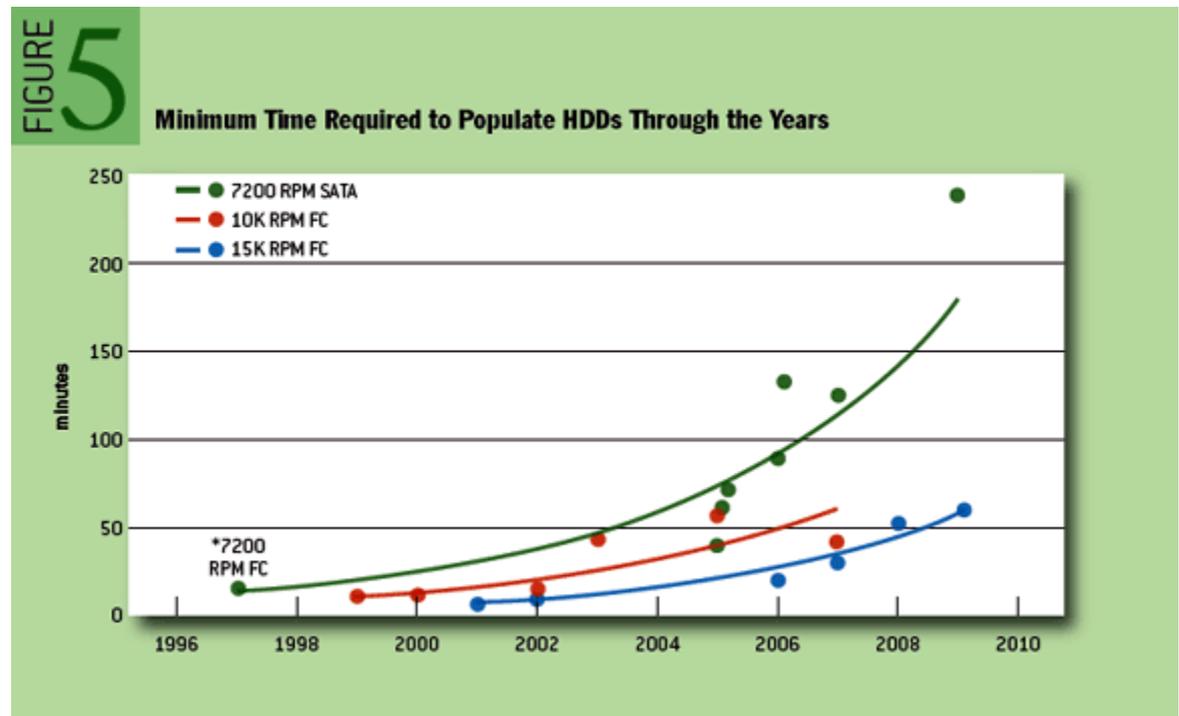
FIGURE 4

Historical Capacity/Throughput of 15K RPM FC HDs



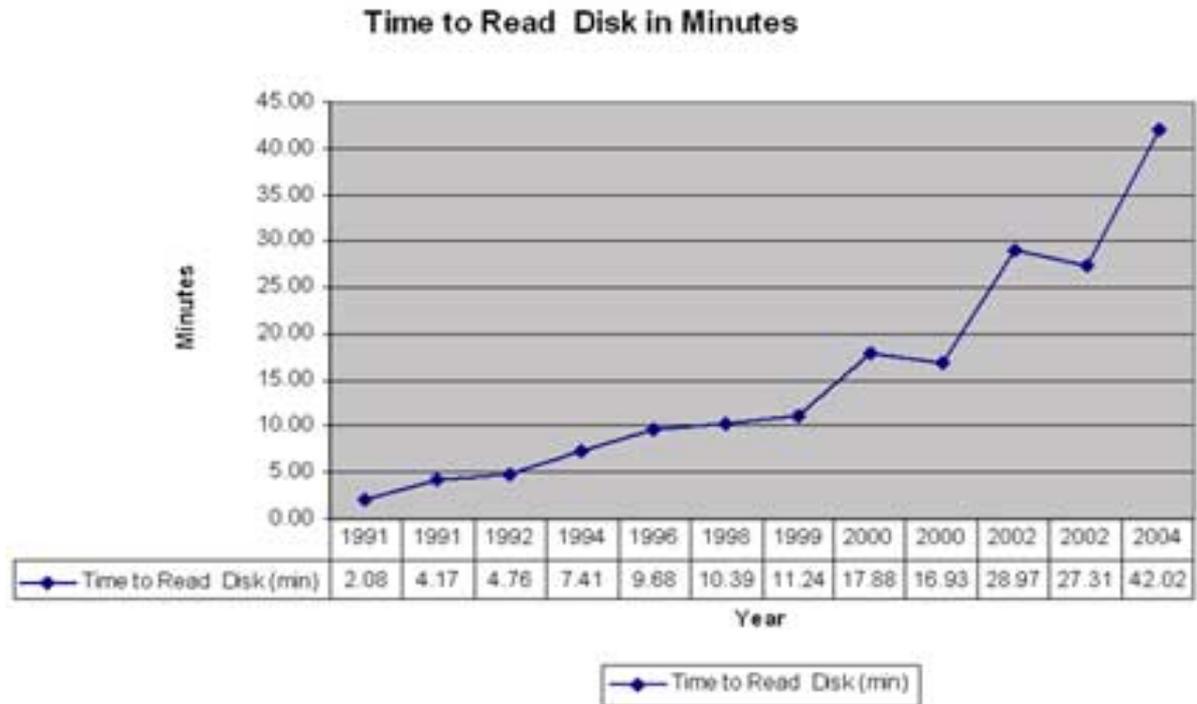
Kapazität/Durchsatz

- Kein exponentielles Wachstum beim Durchsatz
- Die Zeit zum Auslesen oder Wiederherstellen von Platten steigt
- Latente Fehler können sich einschleichen und spät bemerkbar machen (bit rotation)



Kapazität/Durchsatz

- Weiteres Beispiel



[34]

Datenspeicherung

- Schätzung
 - 90% der produzierten Information wird auf magnetischen Medien gespeichert, davon der größte Teil auf Festplatten
 - Wie sicher sind die Daten?
 - Wie kann eine Erhöhung der Sicherheit, Verfügbarkeit erreicht werden?



- Mehr Daten fallen an durch
 - Größere Konzentration/Zentralisierung,
 - wachsende Bedeutung und Abhängigkeit von remote Diensten
 - Frage nach der
 - Verlässlichkeit der Daten
 - Verfügbarkeit der Daten
 - Frage nach der Verlässlichkeit der Komponenten
 - Antworten durch Fehlerabschätzungen

Fehlerquellen

- Viele Möglichkeiten für Plattenausfälle
 - Fehlerarten
 - Fehler sind der Normalfall: Im Idealfall arbeitet die Platte perfekt oder fällt mit einem eindeutigen Fehler aus, doch meistens fällt eine Platte ohne Vorwarnung aus.
 - Latent Sector Error (LSE): Ein bestimmter Sektor kann nicht gelesen oder geschrieben werden oder es tritt ein nicht behebbarer ECC-Fehler auf. Wenn es während einer Wiederherstellung eines RAID auftritt, ist es verheerend. Die Wahrscheinlichkeit hierfür steigt mit zunehmender Zeit für die Wiederherstellung, die früher Minuten, heute Stunden benötigt.
 - Not-Ready-Condition Error: Die Platte kann im Augenblick nicht angesprochen werden oder kann ein Kommando nicht ausführen.
 - Recovered Error: Mehrere Versuche sind notwendig, um den Befehl auszuführen. Auf Plattenebene.
 - Nonrecoverable Failure Rate (AFR)

- Bei den ersten Festplatten lag eine Liste defekter Sektoren bei
- Fehler sind bei den heutigen Dichten nicht zu vermeiden. An versteckter Stelle auf der Platte wird eine Liste der Oberflächenschäden abgespeichert. Die Elektronik der Platte enthält ein Programm zur Auswertung der Liste.

- Latent Sector Error (LSE)

- Nach Angaben von Herstellern sind LSE-Raten von 1 auf 10^{14} bis 10^{15} Bits (11-113 TB) beim Lesen zu erwarten
 - Bei kleinen Platten kann dies vernachlässigt werden
 - Bei großen Platten fällt es stark ins Gewicht:
 - Verheerende Auswirkungen, wenn Metadaten betroffen sind.
- Vorgriff: Ein Plattensystem habe 4 identische Platten mit je 2 TB, eine davon sei redundant. Fällt eine Platte aus, müssen zur Wiederherstellung der ausgefallenen Platte drei Platten gleichzeitig ohne Fehler (auch ohne LSE), d.h. 3x2 TB gelesen werden.
 - Die Wahrscheinlichkeit, dabei auf einen LSE zu stoßen, liegt bei rund 40%. Für die eingesetzten Platten wird eine Rate von 10^{-14} (5% bei 10^{-15}) angenommen.
- Das Verhältnis LSE-Rate/Kapazität ist annähernd konstant. (Allerdings gibt auch Aussagen, daß die Rate mit der Kapazität wächst!)

- Ein defekter Sektor wirkt sich unterschiedlich aus.
 - Ist die Platte fast leer, besteht eine Chance, daß der Fehler einen freien Block trifft.
 - Ist die Platte ziemlich voll, ist die Wahrscheinlichkeit von Datenkorruption groß. Je größer ein File ist, umso eher kann er betroffen sein.
 - Enthält der korrumpierte Block Metadaten, können sehr viele Daten verloren sein.
- Abschätzungen der Risiken je nach Platte und Typ
 - MTTF, MTBF, (s. Kapitel 2.3)

Recovery

- Recovery Mechanismen
Durch geeignete Heuristiken zur Fehlervermeidung sollen Fehler sich erst gar nicht bemerkbar machen.
- SMART (Self-Monitoring, Analysis and Reporting Technology).
Standardisierung Ende 2000
Wird im Linux-Bereich eingesetzt. Windows-Versionen sind verfügbar.
Kann u.a. defekte Sektoren erkennen, sperren und durch Reservesektoren ersetzen.

– SMART



– s. Jon Elerath, Hard Disk Drives: ...
[1]

Einige von SMART erfaßte Fehlersituationen

	S.M.A.R.T.-Attribut	Bedeutung	Problem
	Power On Hours Count	Zahl der Betriebsstunden	Abnutzung
	Power Cycle Count	Zahl der Einschaltvorgänge	
*	Raw Read Error Rate	nicht korrigierbare Lesefehler	Plattenoberfläche
*	Write Error Rate	Schreibfehlerrate	
*	Seek Error Rate	Fehlerrate beim Positionieren	Positionierung
*	Spin-Up Time	Anlaufzeit des Spindelmotors	Motor oder Lager
	Start/Stop Count	Anzahl Start/Stop-Vorgänge	Abnutzung
!	Reallocated Sector Count	Verbrauchte Reservesektoren	Plattenoberfläche
*	Spin Retry Count	Fehlstarts	Motor
	Power Cycle Count	Anzahl Einschaltvorgänge	Abnutzung
*	ECC Recovered	(korrigierte) Bitfehler	Plattenoberfläche
	Ultra DMA CRC Error Count	Übertragungsfehler zum PC	schlechte Kabel oder Kontakte
!	Scan Error Rate	nicht korrigierbare Fehler	Plattenoberfläche
	Load/Unload Cycle Count	Parkvorgänge	Abnutzung
*	Wichtig für die Lebensdauer		
!	Wert ist kritisch für den Zustand der Festplatte		
	Aussagekraft der Werte (s. Untersuchung von Google)		

- Wichtig für die Lebensdauer

2.3 Qualität technischer Systeme

- Bewertung der Zuverlässigkeit u. Sicherheit -
Metriken zur Fehlerabschätzung

Ausfall (failure)

Ein Produkt kann seine Funktion nicht mehr erfüllen.

Dabei kann der Ausfall plötzlich (Sprungausfall) oder absehbar (Driftausfall) erfolgen.

Fehler können sich unterschiedlich auswirken: Von der Einschränkung der Funktion/Leistung (minor failure) und bis kompletten Ausfall (major failure) gehen.

Ein Fehler kann die Ursache für den Ausfall sein.

Das System kann unterschiedlich heftig auf Fehler reagieren. Wenn es nicht gleich auf „kleinere Fehler“ mit einem Ausfall reagiert, gilt es als *fehlertolerant*, andernfalls als *fehlerintolerant*.

Die Fehlertoleranz kann durch zusätzlichen Aufwand (Redundanz) vergrößert werden.

Fehler können behebbbar sein/das System reparierbar sei oder sie führen zum Totalausfall/das System ist dann nicht reparierbar.

Ausfall (Forts.)

Was kann in welcher Situation als Ausfall gezählt werden? (Vgl. [27], [28])

Angaben zur MTBF gibt es seit über 60 Jahren. Viele verschiedene Methoden und Verfahren zur Vorhersage der Nutzungsdauer wurden entwickelt. Wenn angegeben wird, was unter einem Ausfall verstanden wird, sind die Angaben schwer nachvollziehbar: „Sie hängen in der Luft!“

Da Produkte aus vielen verschiedenen Komponenten bestehen, ist die Definition für Ausfall wichtig (vgl. [27]):

- 1) Das Produkt insgesamt kann die gewünschte Funktion nicht mehr ausführen.
- 2) Eine einzelne Komponente kann die gewünschte Funktion nicht mehr ausführen, das Produkt insgesamt jedoch wohl.

Fällt an einem Rechner eine LED-Lampe aus, der Rechner aber uneingeschränkt in Betrieb bleibt, gilt das als Ausfall des Rechners?

Führt die falsche Handhabung des Rechners durch den Benutzer zum Absturz, ist das als Ausfall des Rechners einzustufen?



Zuverlässigkeit (reliability)

Unter Zuverlässigkeit wird der Teil der Eigenschaften eines Produkts verstanden, die von ihm erwarteten Leistungen während eines vorgegebenen Zeitraumes zu erbringen. Sie ist charakteristisch für das Produkt.

Zuverlässigkeit:

Ist die Fähigkeit eines Systems oder einer Komponente, die gewünschte Funktion unter festgelegten Bedingungen für einen bestimmten Zeitraum auszuführen [IEEE 90].

Verfügbarkeit

Die Verfügbarkeit versteht man die Wahrscheinlichkeit, daß sich ein System in einem Zustand befindet, der es erlaubt, die gewünschte Funktion durchzuführen. Sie wird durch die Zuverlässigkeit und durch die Wiederherstellungszeit bestimmt.

Verfügbarkeit:

Ist der Grad, in dem ein System oder eine Komponente einsatzbereit oder zugänglich ist, wenn es bzw. sie verwendet werden soll [IEEE 90].

Lebensdauer

Die Lebensdauer eines Produkts ist die Zeit bis zum (endgültigen) Ausfall eines Produkts. Sie kann in konkreten Zeiteinheiten (Stunden, Tage, Monate, ...) oder in Lastwechseln, Schaltvorgängen, ... gemessen werden.

Bestimmung der Zuverlässigkeit (reliability)

Unter Zuverlässigkeit wird der Teil der Eigenschaften eines Produkts verstanden, die von ihm erwarteten Leistungen während eines vorgegebenen Zeitraumes zu erbringen. Sie ist charakteristisch für das Produkt und wird als Zufallsfunktion R definiert.

Wir unterscheiden theoretische und praktische Modelle für reparierbare und nichtreparierbare Produkte.

Definitionen:

Die **Zuverlässigkeit** $R(t)$ ist die Wahrscheinlichkeit, daß das Produkt zum Zeitpunkt t seine Funktion erfüllt, vorausgesetzt es funktioniert zur Zeit $t_0 = 0$, d.h. $R(t_0) = 1$.

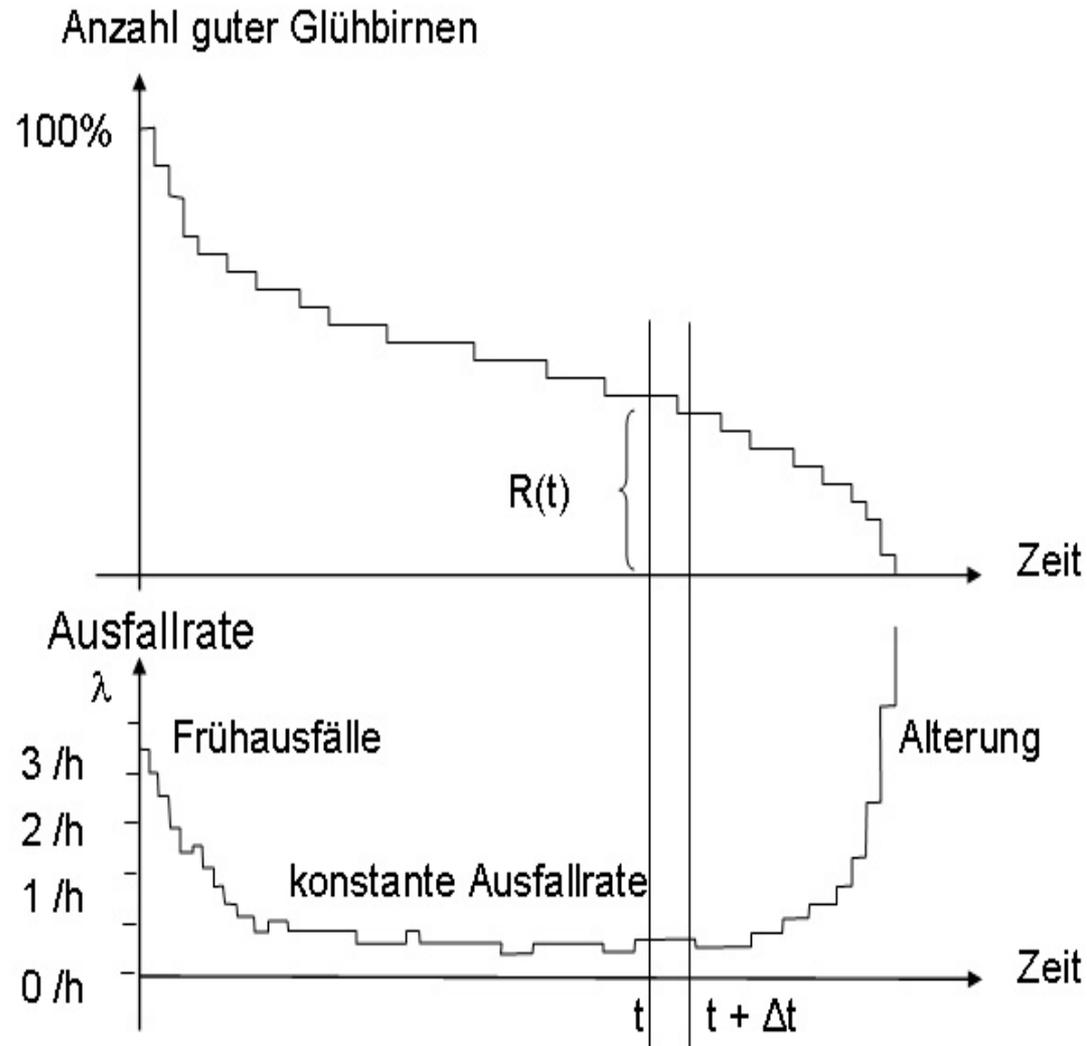
Das Komplement der Wahrscheinlichkeit für den Ausfall des Produkts im Zeitraum $[0,t]$ ist als **Unzuverlässigkeit** $F(t) = 1 - R(t)$ definiert.

Die Zuverlässigkeit ist eine monoton fallende Funktion, wie das folgende Bild zeigt.

Experimentelle Zuverlässigkeit

Über die Zuverlässigkeit eines einzelnen Produkts kann man erst etwas aussagen, wenn es ausgefallen ist. Insofern ist dieser Zugang wenig hilfreich. Umgekehrt kann man aus der Beobachtung einer größeren Anzahl des gleichen Produkts dagegen eher etwas über die Zuverlässigkeit eines einzelnen aussagen. Ausgehend von einer festen Anzahl gleicher Produkte wird das Verhalten über einen längeren Zeitraum beobachtet.

Beispiel: Einige Bezeichnungen im Zusammenhang mit Angaben zu diesem Komplex sollen an einem Beispiel erläutert werden. Die Qualität von Glühlampen soll kontrolliert werden. Dazu wird die Qualität einer größeren festen Menge über einen längeren Zeitraum beobachtet. Die ausgefallenen Glühlampen werden erfaßt. Stellt man die Zahl funktionierender Glühlampen graphisch dar, so ergibt sich eine Kurve wie auf der nächsten Seite.



Experimentelle Zuverlässigkeit

Sei λ die Ausfallrate. Dann gilt

$$\lambda = \frac{\text{Anzahl ausgefallener Lampen}}{\text{Anzahl heiler Lampen zu Beginn des betrachteten Zeitraums}}$$

$$\frac{R_{i+1} - R_i}{R_{i+1}} = -\lambda_i \Delta t$$

$$\frac{R_{i+1} - R_i}{t_{i+1} - t_i} = \frac{\Delta R}{\Delta t} = -\lambda_i R_i$$

$$\frac{dR(t)}{dt} = -\lambda(t)R(t)$$

$$R(t) = e^{-\int_0^t \lambda(\tau) d\tau}$$

Allgemein setzen wir

$$\lambda = \frac{\# \text{Ausfälle innerhalb eines Zeitraums}}{\# \text{Einheiten zu Beginn des Zeitraums}}$$

mit passend gewählter Einheit für den Zeitraum als Minute, Stunde, Tag, Jahr usw.

Experimentelle Fehlerrate

Ist die Anzahl der Produkte, die in die Bestimmung eingehen, relativ groß und Teil einer größeren Serie gleicher Produkte (Stichprobe), so kann mit statischen Mitteln auf das Verhalten der Gesamtmenge geschlossen werden. Das verfolgen wir aber nicht weiter. Es gibt einige Kataloge für elektronische und mechanische Komponenten, denen wir die Fehlerrate entnehmen können wie z.B.:

Military Reliability Handbook, MIL-HDBK-217D

MIL-HDBK-718

Siemens SN29500

RAC

RDP

Bellcore

FIDES 2009

FIT

Die Ausfallraten sind normalerweise sehr klein, die Dezimalzahl hat viele Nachkommastellen. Um ganze Zahlen verwenden zu können, wird eine andere Einheit verwendet. Die Fehlerrate wird als „Fehler in einer Milliarde Stunden ($10^9 h$)“ oder „FIT“ (Failure In Time) angegeben. Die Einheit 1 FIT entspricht einem Fehler in 114.000 Jahren. Zur Bestimmung von der Größe in FIT können 1000 Geräte jeweils 1 Millionen Stunden oder 1 Millionen Geräte jeweils 1000 Stunden eingesetzt werden. (Andere Varianten sind auch möglich.) Die Fehleranzahl ergibt den Wert für FIT.

Bestimmung der Zuverlässigkeit (Forts.)

Viele Werte können bereits Tabellen entnommen werden. Die folgenden Fehlerraten für einige elektronische Komponenten sind aus Wikipedia:

Bauelement	λ /FIT
Lötstelle	1
Widerstand	1,5
Silizium-Diode	3
Silizium-Transistor	5
Keramikkondensator	6
Folienkondensator	10
IC-Sockel (je Kontakt)	10
Steckkontakt	10
Aluminium-Elektrolytkondensator	10 bis 50 (baugrößenabhängig)
Tantal-Elektrolytkondensator	40
Silizium-Leistungsdiode	50
Silizium-Leistungstransistor	60
Integrierte Schaltung (SSI)	100
Integrierte Schaltung (MSI/LSI)	200
Netztrafo, Relais	200
Potentiometer	200

Bestimmung der Zuverlässigkeit (Forts.)

Beispiel: Ein Keramikkondensator mit 6 FIT oder MTBF von 19.000 Jahren soll 10 Jahre lang eingesetzt werden. Wie groß ist die Wahrscheinlichkeit, diese Zeit fehlerfrei zu laufen?

$$\begin{aligned} t &= 10 \text{ Jahre} \\ e^{-\frac{t}{MTBF}} &= e^{-\frac{10 \text{ Jahre}}{19.000 \text{ Jahre}}} \\ &= e^{-0,0005263} \\ &= 0,999.473 \end{aligned}$$

Die Wahrscheinlichkeit für den fehlerfreien Betrieb ist 99,95%.



Bestimmung der Zuverlässigkeit (Forts.)

Je nach praktischem Einsatz und Fragestellung können *andere Einheiten* sinnvoller sein, in dem die Ausfallrate anhand der

- Leistungsaufnahme
- Umgebungstemperatur Schaltfrequenz
- Fehler pro Jahr
- Start-/Stoppsyklen

bestimmt wird.

Einflußgrößen auf die Ausfallrate

Sie hängt von der *Umgebung* ab. Hier sind die

- Umgebungstemperatur (Arrhenius-Gesetz),
- Erschütterungen,
- Strahlung und
- Feuchtigkeit

zu nennen.

Sie hängt vom *Alter* des Objekts ab.

Bestimmung der Zuverlässigkeit (Forts.)

Theoretische Zuverlässigkeit

Die experimentelle Bestimmung der Zuverlässigkeit ist zeitaufwendig.

Vorhandener Werte für einzelne Komponenten eines Produkts können nicht genutzt werden. Unter entsprechenden Annahmen können theoretische Modelle aufgestellt werden – die wieder an der Praxis gemessen und angepaßt werden.

Schaut man sich das Beispiel mit den Glühlampen einmal an, so ist im *mittleren Bereich* die Ausfallrate $\lambda(t)$ annähernd konstant. Wird λ als *konstant* angenommen, so erhalten wir

$$R(t) = e^{-\lambda t}$$

und für die Fläche unter $R(t)$

$$MTTF = \int_0^{\infty} R(t) dt = 1/\lambda$$

(MTTF (Mean-Time-To-Failure)) Die Funktion

$$F(t) = 1 - R(t) = 1 - e^{-\lambda t}$$

ist die Verteilungsfunktion der Exponentialverteilung mit der Dichtefunktion $f(t) = \lambda e^{-\lambda t}$.



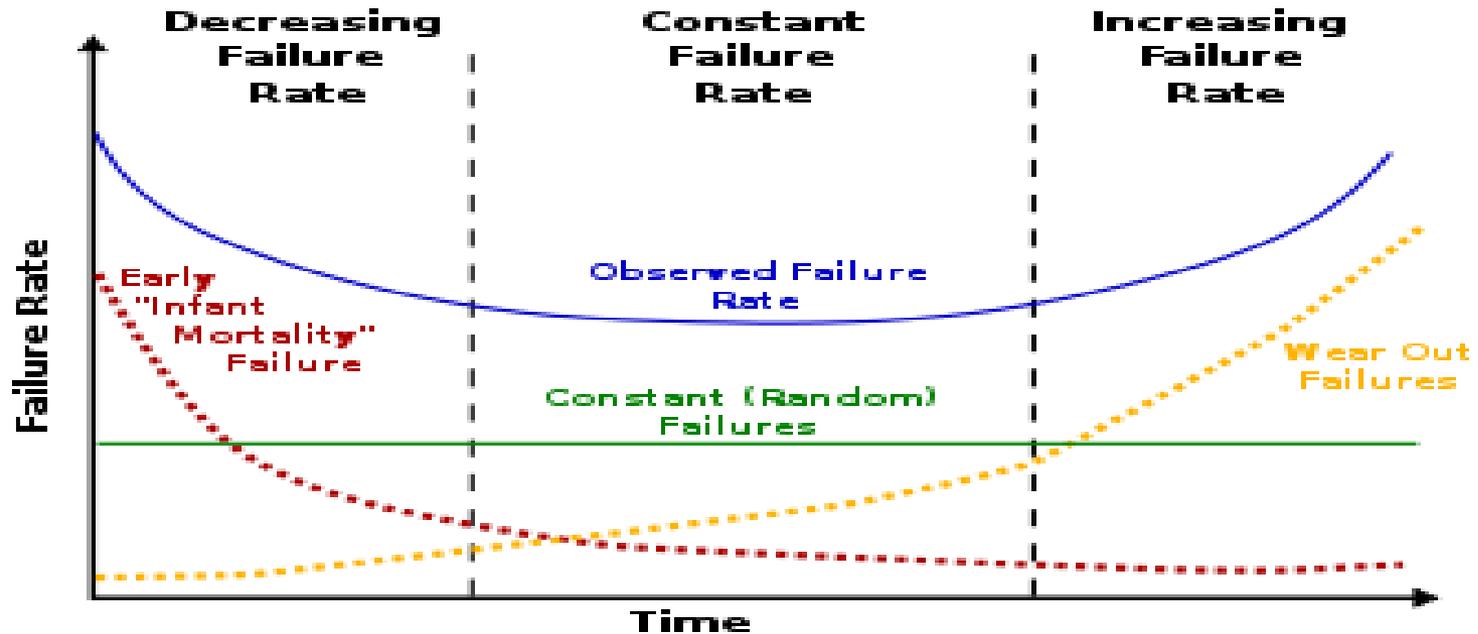
Theoretische Zuverlässigkeit (Forts.)

Mit der Exponentialverteilung läßt sich nur die mittlere Phase beim Einsatz eines Produkts untersuchen, wenn die Fehlerrate nahezu konstant ist. Sie ist aber nicht konstant über die gesamte Einsatzzeit eines Produkts. Grob lassen sich drei Phasen unterscheiden:

- Zuerst ist sie relativ hoch, aber fällt mit der Zeit („Kindersterblichkeit“).
- In der zweiten Phase ist sie relativ niedrig und annähernd konstant („Reife“).
- Dann steigt sie kontinuierlich an („Verschleißphase/Alterung“).

Werden die Fehlerraten gegen die Zeit angegeben, so ergeben sich drei fallende, konstante bzw. steigende Kurven, deren Summe als „typische Badewannenkurve“ bekannt ist:

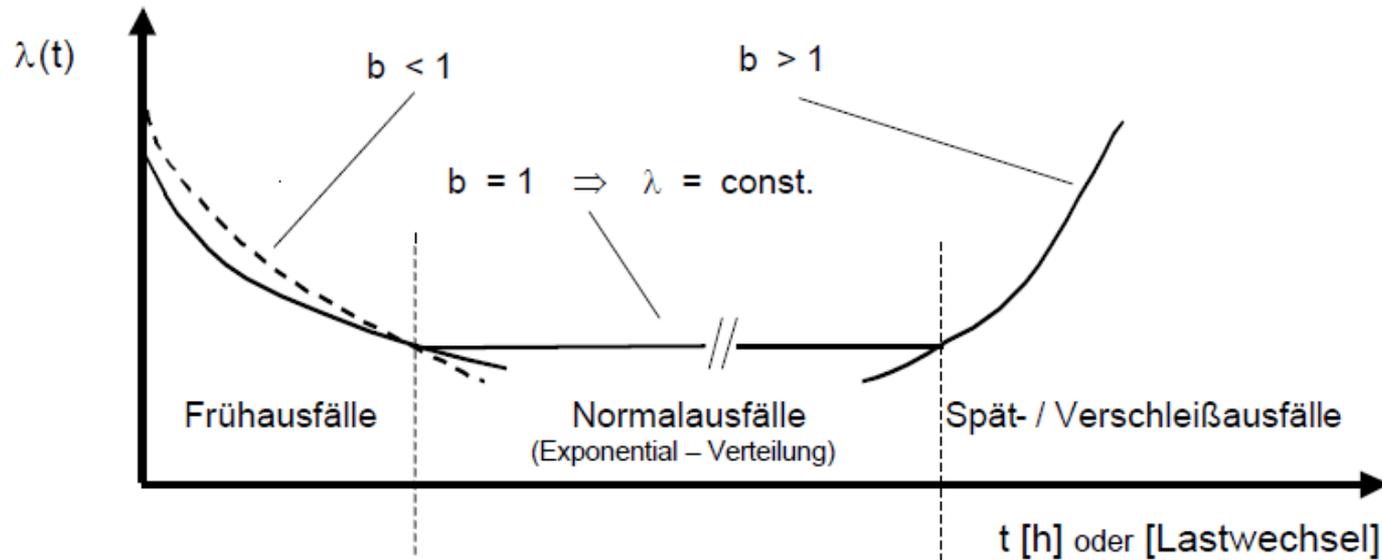
Theoretische Zuverlässigkeit (Forts.)



- FR (Failure Rate) = Fehleranzahl / Komponentenanzahl pro Zeitintervall
- Infant Mortality Failure: Messung bei elektronischen Komponenten über 200 h

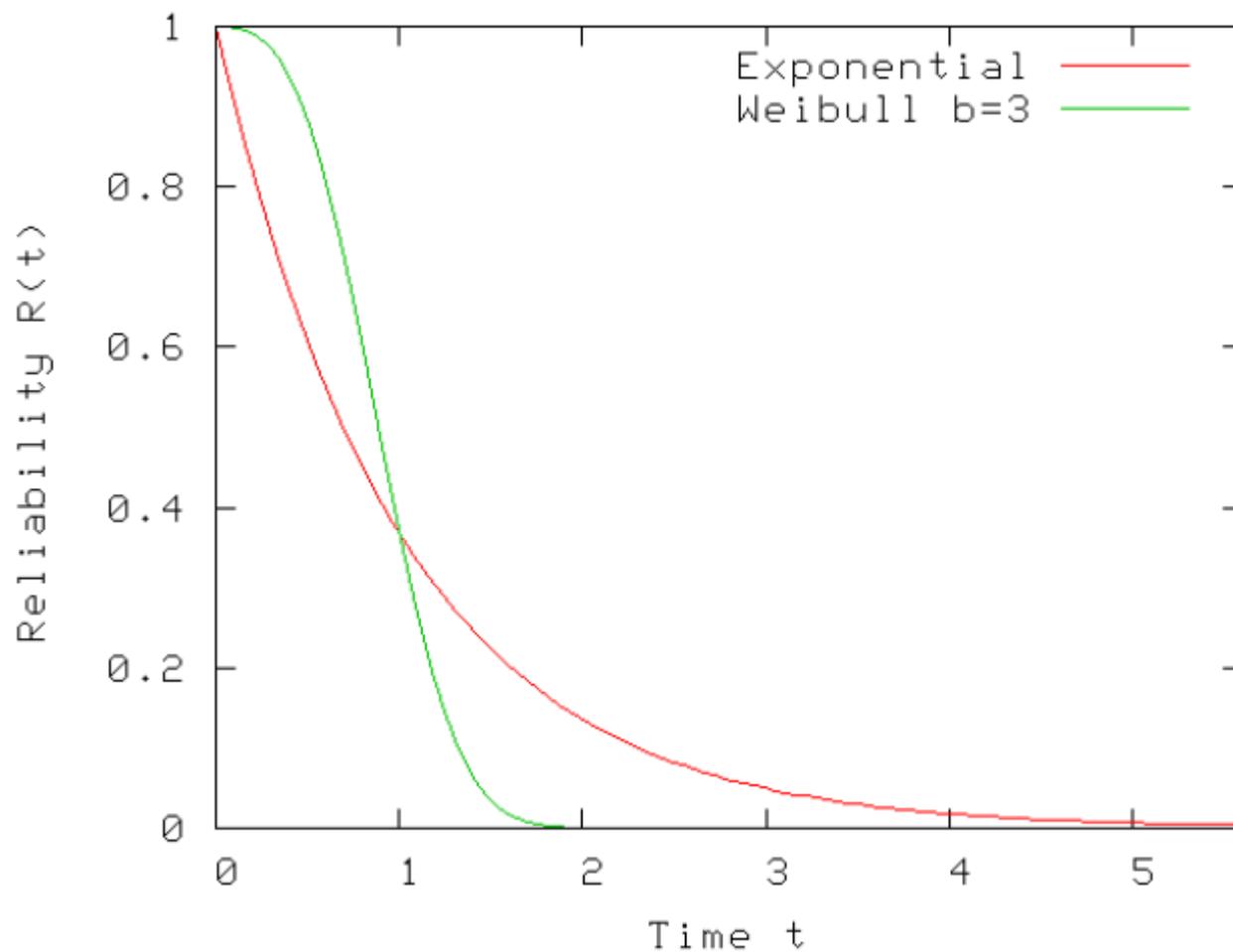
Theoretische Zuverlässigkeit (Forts.)

Die Badewannenkurve läßt sich nicht komplett mit einer Exponentialverteilung beschreiben. Dafür gibt es die Weibullverteilung mit einem Formparameter b und Lageparameter T . λ ist wieder die Fehlerrate.

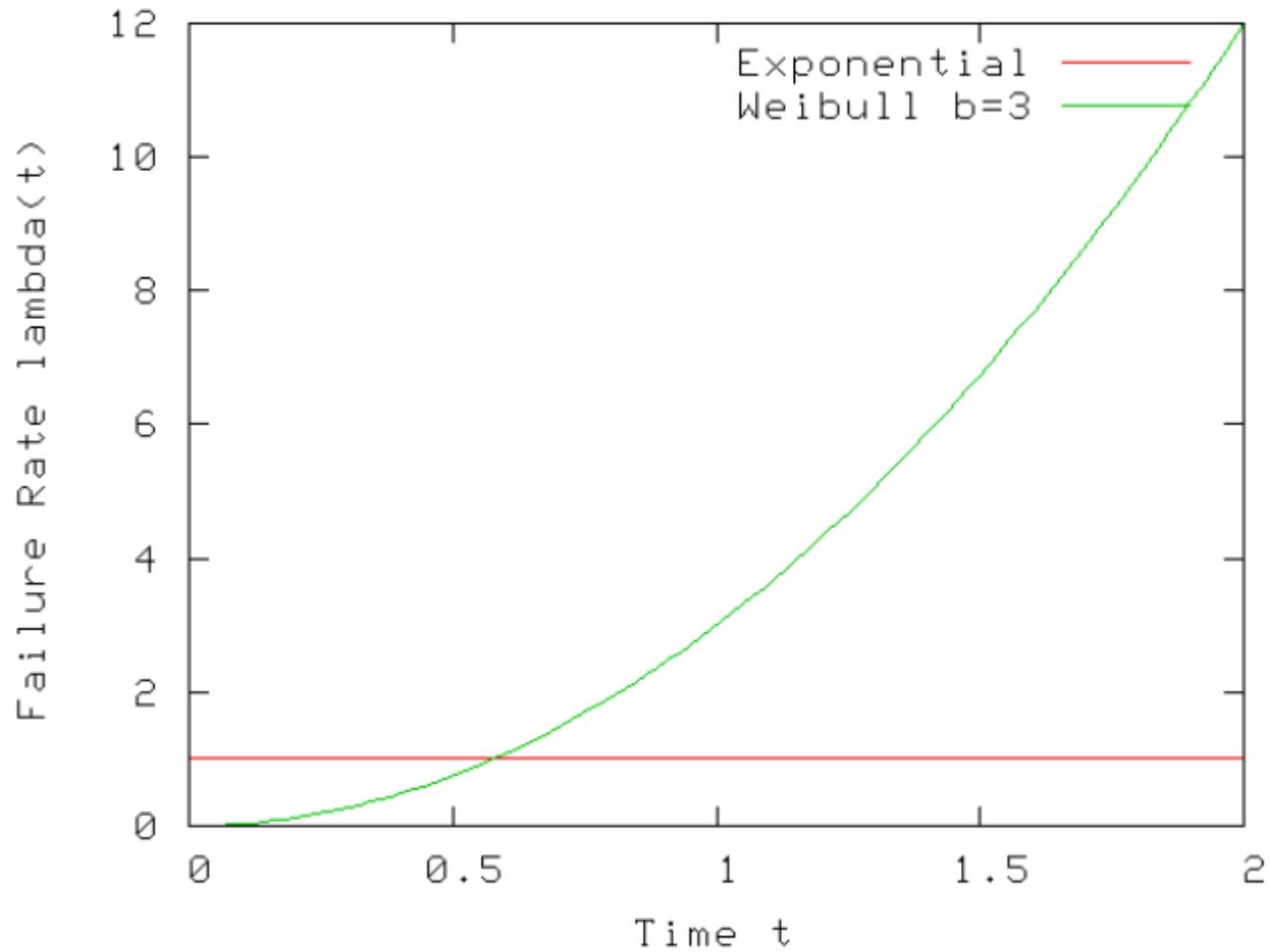


Dichtefunktion $f(t) = \frac{b}{T} \cdot \left(\frac{t}{T}\right)^{b-1} \cdot e^{-\left(\frac{t}{T}\right)^b}$ für $t, T, b \geq 0$

Verteilungsfunktion $F(t) = 1 - e^{-\left(\frac{t}{T}\right)^b}$ für $t, T, b \geq 0$



[29]

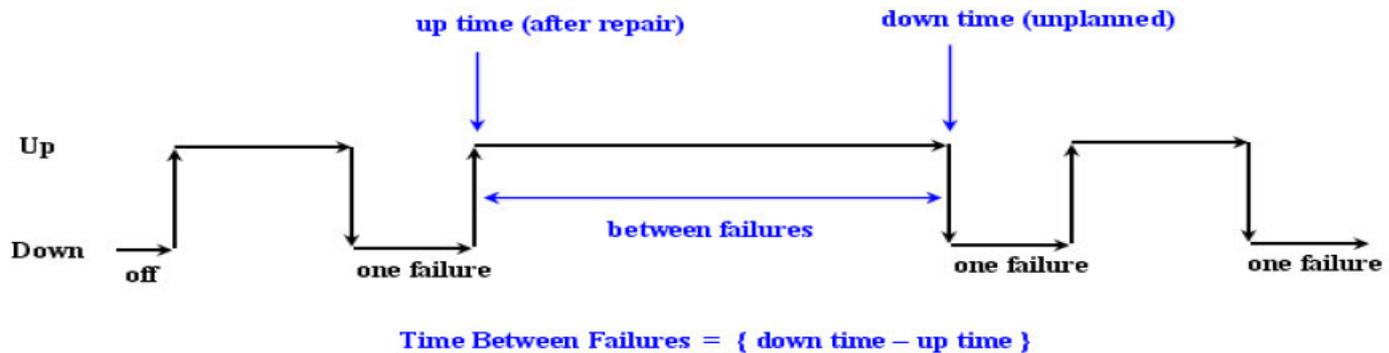


[29]

2.2 Qualität technischer Systeme

- Fehlerfunktionen (Zusammenstellung)
 - MTBF (Mean-Time-Between-Failure [mittlerweile auch Mean-Time-Before-Failure])
Mittlere Betriebsdauer/Erwartungswert der Betriebsdauer zwischen zwei aufeinander folgenden Ausfällen. Wenn up-time der Einschalt- und down-time der nächste Ausfallzeitpunkt sowie n die Zahl der Ausfälle während des Beobachtungszeitraums ist:

$$MTBF = \frac{\Sigma(\text{down-time} - \text{up-time})}{n}$$



- Fehlerfunktionen (Zusammenstellung)

- MTTR (Mean-Time-To-Repair)

Mittlere Reparaturzeit ist die Zeit, die zur Instandsetzung des defektes Gerätes erforderlich ist. Häufig müssen defekte Hardwareteile ausgewechselt werden.

Um die Reparaturzeiten und damit die Ausfallzeiten und Kosten möglichst gering zu halten, ist eine Ersatzteilbevorratung im Einzelfall nötig und genügend Personal für die Reparatur ist erforderlich.

Eine andere Möglichkeit, die Reparaturzeiten gering zu halten, ist, Systeme mit genügend „Reserve“ redundant zu betreiben. Die redundante Erweiterung kann gleichzeitig mit dem System betrieben werden oder bei Ausfall zugeschaltet werden, um die Funktion des ausgefallenen Systems so lange übernehmen, bis die Reparatur erfolgt und das reparierte System wieder einsatzbereit ist.

- Fehlerfunktionen

- AFR (Annual Failure Rate)

Sie kann aus einer Anzahl eingesetzter Geräte (Grundmenge) durch die Praxis bestimmt werden. Wird angenommen, daß alle Geräte 24 Stunden pro Tag und 365 Tage im Jahr eingesetzt werden, so gilt

$$AFR = \frac{\text{Ausfälle im Probenzeitraum} \times (52 \text{ Wochen pro Jahr} / \text{Anzahl der Wochen im Probenzeitraum})}{\text{Anzahl der Geräte in der Grundmenge}}$$

- AFR (Annual Failure Rate - Variante)

Sind die Geräte nicht rund um die Uhr im Einsatz, so ist eine Berechnung nach

$$AFR = \frac{\text{Ausfälle im Probenzeitraum} \times (52 \text{ Wochen pro Jahr} / \text{Anzahl der Wochen im Probenzeitraum})}{\text{Kumulativer Einsatz in Jahren}}$$

- $AFR = 1 - \exp(-8760/MTBF) \approx 8760/MTBF$, da $MTBF \gg 8760$.

- POH (Power-On Hours)

$$AFR = 1/MTBF \cdot POH$$

- Reparaturrate $\mu = \mu(t) = \text{const}$

$$\mu = 1 / MTTR \quad \text{Mittlere Reparaturrate}$$

- Fehlerfunktionen

- Beispiel:

- AFR (Annual Failure Rate)

Die Grundmenge umfaßt 10.000 Auto, die über einen Zeitraum von 2 Monate eingesetzt werden. Im Durchschnitt wird jedes Auto 400 Stunden eingesetzt. Im Testzeitraum fallen 10 Autos aus. Dann ist

$$AFR = \frac{10 \text{ Ausfälle} \times (52 \text{ Wochen pro Jahr} / 8 \text{ Wochen im Testzeitraum})}{10.000}$$

was $AFR = 0,0065$ oder $0,65\%$ ergibt.

- AFR (Annual Failure Rate - Variante)

Unter Berücksichtigung der tatsächlichen Einsatzstunden erhält man mit der zweiten Formel als realistischere Angabe

$$AFR = \frac{\text{Ausfälle im Probenzeitraum} \times (52 \text{ Wochen pro Jahr} / \text{Anzahl der Wochen im Testzeitraum})}{\text{Kumulativer Einsatz in Jahren}}$$

$$= \frac{10 \text{ Ausfälle} \times (52 \text{ Wochen pro Jahr} / 8 \text{ Wochen im Testzeitraum})}{10.000 \times 400 \text{ Stunden}}$$

Im Nenner stehen rund 457 kumulative Autojahre, so daß $AFR = 0,14$ oder 14% ist.

- Fehlerfunktionen

- MTTF (Mean-Time-To-Failure)

- Gilt für nichtreparierbare oder reparierbare Einheiten nur unter der Annahme, daß die Betrachtungseinheit nach der Reparatur neuwertig sind.

$$MTTF = POH / AFR$$

- Die Aussagekraft ist vorsichtig zu nehmen.

Ein dreißig Jahre alter Mensch hat eine MTTF von 900 Jahre [6]. Das heißt nicht, daß er 900 Jahre alt wird, sondern die Wahrscheinlichkeit, im nächsten Jahr zu sterben, beträgt 1:900.

Annahme, eine Platte hat eine MTTF von 20 Jahren. 100 Platten werden 6 Monate eingesetzt. Die Einsatzdauer als Summe beträgt $100 \cdot 0,5 = 50$ Plattenjahre. Wir würden $50 / 20 = 2,5$ Ausfälle erwarten.

- Fehler- oder Ausfallrate $\lambda = \lambda(t) = \text{const}$

$$\lambda = 1 / MTTF \quad \text{Mittlere Zeit bis zum Ausfall}$$

$$MTBF = 10^9 h / FIT \approx 114.000 \text{ Jahre}/FIT$$



- Fehlerfunktionen

- MTDDL (Mean-Time-To-Data-Loss)

Wenn wir eine Platte ein Jahr ohne Backup einsetzen und sie fällt dann aus, dann ist

MTTF = 1 Jahr und MTDDL = MTTF. (Schlecht!)

Normalerweise sollte MTDDL >> MTTF sein.

- Availability = $MTBF / (MTBF + MTTR)$

Wenn der Fehler sehr schnell behoben wird, so ist die Verfügbarkeit fast 1.

- Nicht enthalten sind geplante Abschaltzeiten (Wartung), kleine Probleme des Betriebsalltags, Bedienungs- und SW-Fehler

– Wie verhalten sich MTTF und MTBF zu einander?

- MTBF (mean time between failures) gilt nicht einer einzelnen Komponente. Sie stellt die mittlere Zeit zwischen zwei auf einander folgender Zufallsausfälle und kann insofern nur sinnvoll auf reparierbare Systeme angewendet werden. Diese können mehrfach ausfallen. Ein Austausch wird als Reparatur angesehen. Die MTBF bezieht sich immer auf die Phase mit konstanter Ausfallrate.
- MTTF (mean time to failure) ist die entsprechende Größe für nicht-reparierbare Systeme. Wird MTTF auf reparierbare Systeme angewendet, so wird damit die mittlere Zeit bis zum ersten Fehler (MTTFF) oder die mean uptime (MUT) innerhalb eines Fehler-Reparaturzyklus beschrieben.

Ist $MTTR \ll MTBF$, dann entspricht $MTBF \approx MTTF$

Wird durch die Reparatur ein Zustand „wie neu“ erreicht, kann MTTF durch MTBF ersetzt werden.

- Wie verhalten sich MTTF und MTBF zu einander?
 - Wenn bei RAIDs von Reparatur einer Platte gesprochen wird, ist es genau genommen ein Austausch der defekten gegen eine lauffähige Platte ist. Wenn Plattenhersteller MTBF-Angaben für ihre Platten veröffentlichen, ist es eigentlich die MTTF-Angabe.
 - Nach dem Austausch einer Platte in einem RAID sind weitere Arbeiten für die Wiederherstellung des RAID-Levels (Rebuild) erforderlich. Diese Zeiten zusammen mit der Zeit für den Plattenaustausch (Besorgen der Platte und Einbau, eventuelles manuelles Einschalten einer Reserveplatte,...) ergeben die MTTR.

- Fehlerfunktionen

Verfügbarkeit/Downtime –Tabelle

	Betriebszeiten				
	Verfügbarkeit 7 x 24 Stunden	Downtime [h]	Downtime [min]	Verfügbarkeit 7 x 12 Stunden	Verfügbarkeit 5x 12 Stunden
Betriebsstunden pro Jahr	8760			4380	3132
99,000000	8.672,4000	87,6000	5.256,0000	4.336,2000	3.100,6800
99,900000	8.751,2400	8,7600	525,6000	4.375,6200	3.128,8680
99,990000	8.759,1240	0,8760	52,5600	4.379,5620	3.131,6868
99,999000	8.759,9124	0,0876	5,2560	4.379,9562	3.131,9687
99,999900	8.759,9912	0,0088	0,5256	4.379,9956	3.131,9969
99,999990	8.759,9991	0,0009	0,0526	4.379,9996	3.131,9997

- Rate of Latent Sector Errors

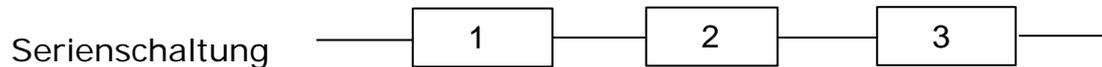
- Weiteres

- Annahme: Fehler treten unabhängig von einander auf(!)
- Idealfall: Die Platte arbeitet perfekt oder fällt mit einem eindeutigen Fehler aus

Fehlerabschätzungen bei zusammengesetzten Systemen

Welche Fehlerabschätzungen kann man für ein Produkt angeben, wenn sie für die einzelnen Komponenten bekannt sind? (Quellen wurden bereits genannt. Die Angaben können abweichen.)

Jede Anordnung kann in die Grundstrukturen Serien- und Parallelschaltung mit und ohne (heiße oder kalte) Redundanz zerlegt werden, woraus die Werte für das Gesamtsystem berechnet werden können.



Die einzelnen Komponenten werden als unabhängig von einander hinsichtlich ihres Fehlerverhaltens mit Fehlerrate λ_i ($i=1,2,3$) betrachtet. Die Zuverlässigkeit der Serienschaltung ist kleiner als die jeder Komponente:

$$R_s(t) = R_1(t) \times R_2(t) \times R_3(t) \times \dots$$

Für die Exponentialverteilung gilt

$$R_s(t) = \exp\{ - (\lambda_1 + \lambda_2 + \lambda_3 + \dots) \times t \}$$

$$= \exp\{ -\lambda_s \times t \}$$

mit Fehlerrate λ_s fürs Gesamtsystem

$$\lambda_s = \lambda_1 + \lambda_2 + \lambda_3 + \dots$$

Fehlerabschätzungen bei zusammengesetzten Systemen

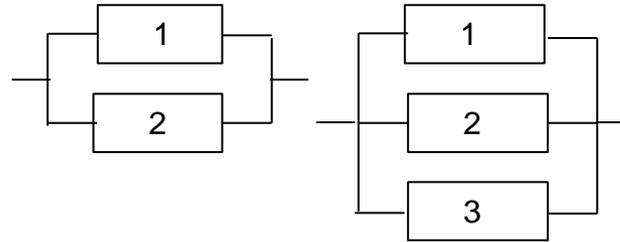
MTBF bei zwei seriell zusammengesetzten Komponenten a und b beträgt , wenn $MTBF_a = MTBF_b$ ist,

$$A_{\text{seriell}} = A_a \times A_b$$

$$MTBF_{\text{seriell}} = \frac{1}{\frac{1}{MTBF_a} + \frac{1}{MTBF_b} + \frac{MTTR}{MTBF_a * MTBF_b}}$$

$$\lambda_{\text{seriell}} = \lambda_a + \lambda_b + MTTR \times \lambda_a * \lambda_b$$

Parallelschaltung



Wir unterscheiden mehrere Fälle:

1. Aktive Redundanz ohne Reparatur

Alle Komponenten sind im Einsatz, wobei nur eine für den erfolgreichen Einsatz reicht. Ausgefallene Komponenten werden nicht ersetzt.

2. Aktive Redundanz mit Reparatur

Alle Komponenten sind im Einsatz. Für den erfolgreichen Einsatz reicht eine. Ausgefallene Komponenten werden sofort repariert oder ersetzt, während das System weiterläuft.

Weitere Unterscheidungen können hinsichtlich der Durchführung der Reparatur getroffen werden. Wir nehmen an, daß beliebig viele Reparaturen gleichzeitig ausgeführt werden, wenn mehr als eine Komponente ausgefallen ist. Es muß also nicht erst mit der Reparatur der zweiten ausgefallenen Komponente gewartet werden, bis der erste Ausfall behoben ist.

1. Redundanz, nicht reparierbar

Die einzelnen nichtreparierbaren Komponenten werden als unabhängig von einander in ihrem Fehlverhalten betrachtet. Die Zuverlässigkeit der Parallelschaltung ist größer und die Ausfallwahrscheinlichkeit umgekehrt kleiner als die jeder Komponente.

Betrachten wir den linken Fall, so gilt für das Systemverhalten

$$\begin{aligned}R_s(t) &= R_1(t) \times (1 - R_2(t)) + (1 - R_1(t)) \times R_2 + R_1(t) \times R_2(t) \\ &= R_1(t) + R_2(t) - R_1(t) \times R_2(t) \\ &= 2R(t) - R^2(t) \quad \text{für } R(t) = R_i(t)\end{aligned}$$

$$F_s(t) = F_1(t) \times F_2(t) \times F_3(t) \times \dots$$

Mit $F(t) = 1 - R(t)$ folgt für den allgemeinen Fall

$$R_s(t) = 1 - \{[1 - R_1(t)] \times [1 - R_2(t)] \times [1 - R_3(t)] \times \dots\}$$

Im Falle einer Exponentialverteilung mit gleicher Ausfallrate λ gilt bei zwei parallel geschalteten nichtreparierbaren Komponenten

$$R_s(t) = R_1(t) + R_2(t) - R_1(t) \times R_2(t)$$

$$R_s(t) = 2 \cdot \exp(-\lambda t) - \exp(-2 \lambda t)$$

$$MTTF_s = \frac{2}{\lambda} - \frac{1}{2\lambda} = \frac{1,5}{\lambda} = 1,5 \cdot MTTF_{\text{einer Komponente}}$$

Es gibt viele Ursachen für den Ausfall von Produkten. Bauteilen können ausfallen, die Konstruktion oder die Fertigung weisen Mängel auf. Weiter kann fehlerhafter Software die Ursache sein (vgl. [25]).

2. Redundant, reparierbar

- a) Wir nehmen an, daß für einzelne Komponenten und das Redundanzsystem die gleiche Wiederherstellungszeit MTTR benötigt wird. Für zwei parallele Komponenten a und b gilt:

$$A_{parallel} = 1 - (1 - A_a) * (1 - A_b)$$

Setzt man hier die Definitionen für Verfügbarkeit für die Einzelkomponenten und das Gesamtsystem ein, so erhält man unter Berücksichtigung der Annahme für MTTR : $MTTR(\text{System}) = MTTR(\text{Komponente } a) = MTTR(\text{Komponente } b)$ durch Auflösen nach MTBF(parallel)

$$MTBF_{parallel} = \frac{MTBF_a * MTBF_b}{MTTR} + MTBF_a + MTBF_b$$

Unter Verwendung von gleichen Komponenten a und b mit Ausfallrate λ und Reparaturrate μ erhalten wir

$$\lambda_{parallel} = \frac{\mu + 2\lambda}{\lambda^2}$$

2. Redundant, reparierbar

b) Wir nehmen jetzt an, daß während der Reparatur der ersten Komponente die zweite auch ausfällt - das System ist nicht mehr einsatzbereit. Im Durchschnitt wird die zweiten Komponente zur Halbzeit der ersten Reparatur ausfallen, d.h. bei gleichen Komponenten a und b gilt $MTTR(\text{System}) = MTTR(\text{Komponente a}) / 2 = MTTR(\text{Komponente b}) / 2$. Dann bekommen wir

$$\lambda_{parallel} = \frac{2\mu}{\lambda^2} + \frac{1}{\lambda}$$

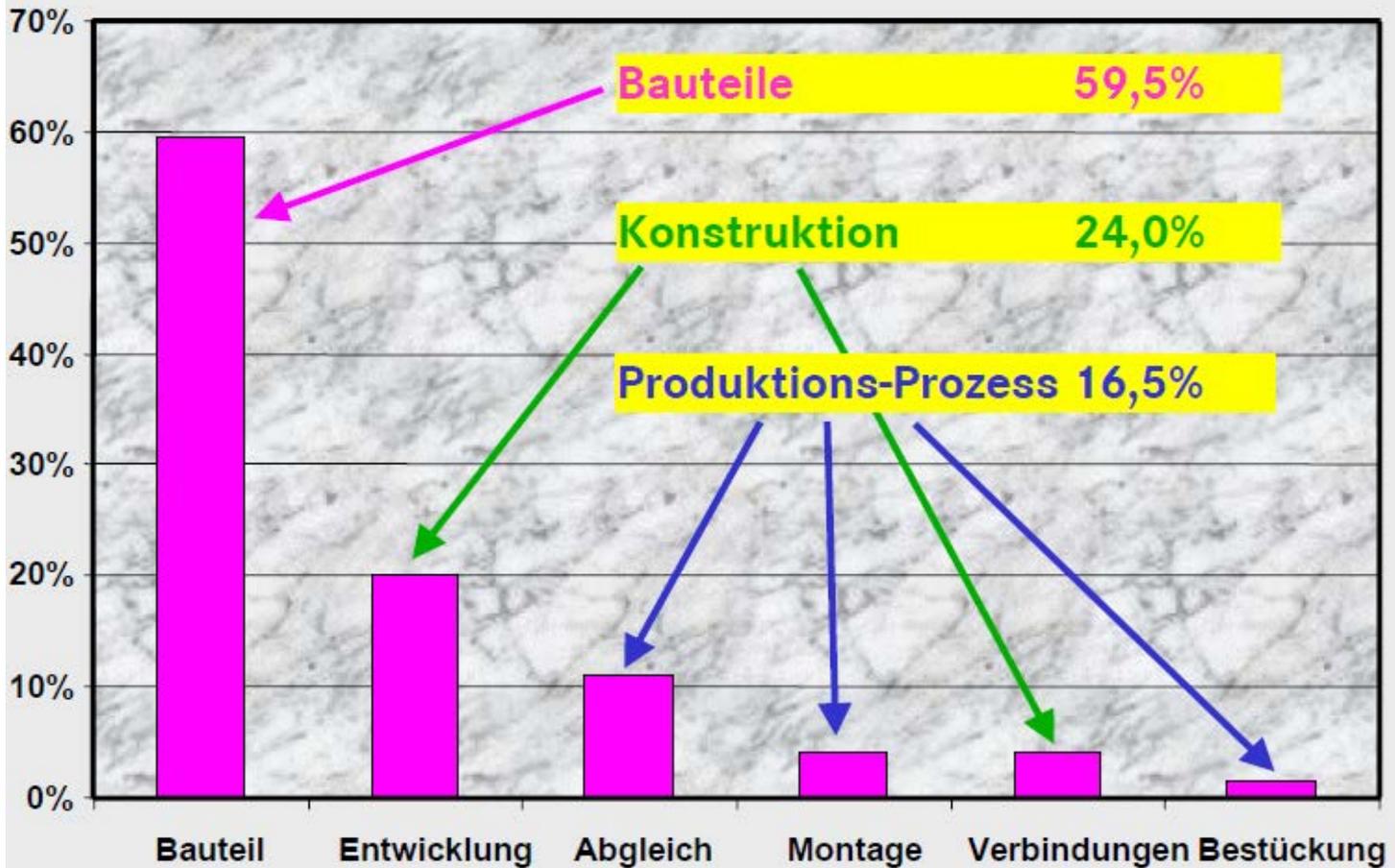
c) Die Fälle a und b enthalten für $\mu = 0$ nicht den Fall der einfachen Redundanz ohne Reparatur. Mit einem Ansatz als Markoffsysteem erhält man mit einigem Aufwand die andere Abschätzung

$$\lambda_{parallel} = \frac{2\mu}{\lambda^2} + \frac{3}{2\lambda}$$

die nun als Spezialfall mit $\mu = 0$ wieder die Abschätzung für die einfache Redundanz ohne Reparatur enthält. In der Praxis arbeitet man je nach Genauigkeitsanforderung bei Bedarf mit verschiedenen Annahmen.

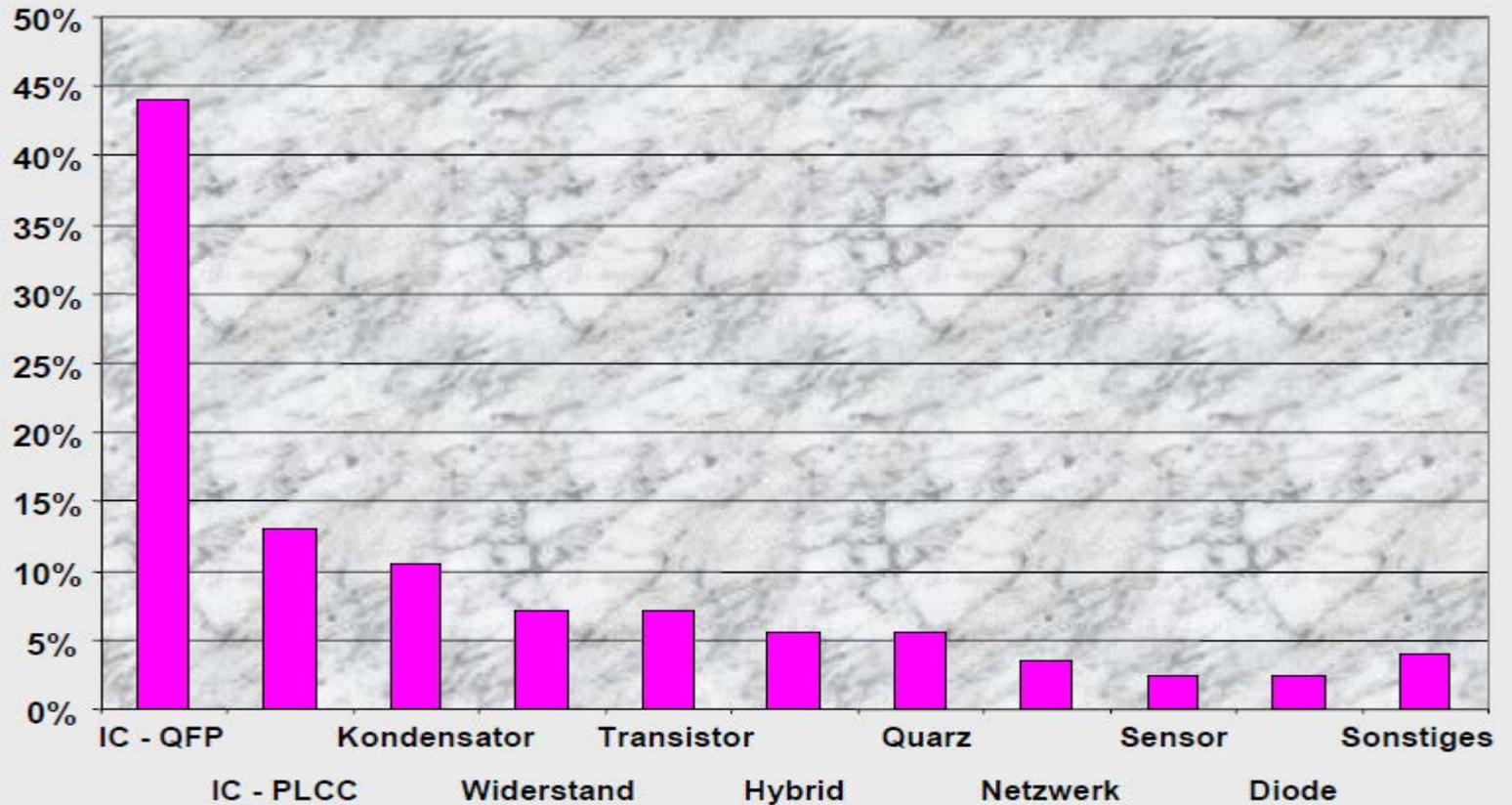
In der Praxis lassen sich verschiedene Fehlerquellen für den Ausfall einer Komponente ausmachen, wie auf den beiden folgenden Folien zu sehen ist.

- Anteil an Ausfällen



[25]

• Fehleranteil bei Bauelementen



[25]

2.3 Qualität magnetischer Speichermedien

- Wie ermitteln die Hersteller die Kenngrößen?

- Theoretische und operationale MTBF-Kenngrößen bei Produkteinführung

Theoretische MTBF: Basiert auf der Analyse bisheriger Modelle mit ähnlichen Eigenschaften. Die Zahlen gehen in ein statisches Modell ein.

Operationale MTBF: Nach der Markteinführung werden die Daten der ausgefallener Platten analysiert und die MTBF korrigiert. Allerdings dringen solche Zahlen nie oder nur selten nach außen.

Die theoretischen Angaben sind meistens zu optimistisch. Sie können aber auch Faktoren wie Fehlbehandlungen nicht berücksichtigen. Wichtig ist sich zu merken, daß es sich um *statistische Durchschnittsangaben* handelt. Man kann daran aber ablesen, wie weit der Hersteller selber seinem Produkt „traut“. Wenn ein Modell 3 Jahre Garantie hat und die Servicezeit mit 5 Jahren angegeben wird und ein anderes Modell 2 Jahre Garantie bzw. 3 Jahre Servicezeit, so ist dem ersten eher als dem zweiten zu vertrauen, selbst wenn dessen MTBF höher als die vom ersten ist.

- Wie ermitteln die Hersteller die Kenngrößen?

- Seagate [17]

500 Platten werden bei 42° C Umgebungstemperatur für 672 Stunden (28 Tage) intensiv mit maximaler Anzahl an Such-, Lese- und Schreibvorgängen) getestet . Die Ergebnisse werden auf 25° C Umgebungstemperatur und einer Betriebsdauer von 2400 Stunden (Economy) bzw. 8760 Stunden für Serverplatten (POH/year) umgerechnet. Daraus ergeben sich die MTBF sowie andere Daten fürs erste Jahr. Weitere Modelle und „Erfahrungswerte“ werden für die Abschätzungen der weiteren Jahre genommen.

– Differenziertere Ergebnisse von NetApp und Google

NetApp: In „*An Analysis of Latent Sector Errors in Disk Drives*“ untersuchten Lakshmi N. Bairavasundaram u.a. [21] die Fehlerlogs von 50.000 RAID-Systemen mit 1,53 Mill. Platten aus dem Enterprise- und Consumerbereich über einen Zeitraum von 32 Monaten.

- Plattenaustausch 2-4x häufiger als nach Herstellerangaben zu erwarten wäre
- Häufig führten andere Fehler, die irrtümlicherweise als Plattenfehler diagnostiziert wurden, zu unnötigem Plattenaustausch
- Die Studie untersuchte Hinweise auf Silent Storage Corruption:
 - 3,45% aller Platten hatte einen LSE in dem Zeitraum von 32 Monaten.
 - Enterprise-Platten haben ungefähr nur 1/10 der LSE-Rate von Consumer-Platten (0,86% von 358.000 nearline SATA und 0,065% von 1,17 Mill. Enterprise-Class Fibre-Channel-HD).
 - Die LSE-Rate steigt mit der Kapazität.
 - Für die meisten Plattenmodelle haben 80 der Platten weniger als 50 LSE.
 - LSE sind nicht unabhängig von einander.
 - Scrubbing ist sehr nützlich zur proaktiven Fehlerbehandlung. 60% aller LSEs wurden dadurch entdeckt. Unbehandelt hätten sie zu Datenverlust geführt.

Google: In einer großflächigen Untersuchung von 100.000 Platten aus dem Consumer- und Enterprise-Bereich kommt Google zu anderen Zahlen als die Hersteller, deren MTBF-Angaben zu optimistisch sind (s. [23], [33]).

- Die Aussage der Hersteller zu MTBF-Angaben sind zu optimistisch.
 - Die SMART-Statistiken zum Erkennen von Scan-Fehlern können sinnvoll genutzt werden, auch wenn es unterschiedliche Versionen gibt.
 - Höhere Belastungen von Platten führen früher – auch noch in der Garantiezeit - zum Ausfall.
 - Einen signifikanten Zusammenhang zwischen Lebensalter und Ausfälle konnten auf Grund der Daten nicht ermittelt werden.
 - Auch bei höheren Temperaturen ist kein Zusammenhang mit vorzeitigem Ausfall zu verzeichnen. Erst bei Extremtemperaturen tritt dies ein.
- **Fazit:**
 - Die Software ist oft fehlerhaft und führt zu Datenverlust oder Fehlhandlungen vom Personal führt zu Datenverlust
 - Die Angaben der Hersteller sind oft zu optimistisch wie Untersuchungen in der Praxis zeigen. U.a. haben die großflächige Untersuchungen von Google und NetApp dies gezeigt.
 - Spreadsheet für MTTDL-Berechnung für verschiedene Platten und Fehlertypen, s. Zetta_MTTDL_June10_2009.xls in [9]

- Die Unabhängigkeit von Bitfehlern ist fraglich bei gering steigendem Budget aber stark wachsenden Datenbeständen: (s. [20])
 - Je mehr Kopien, desto sicherer. Mit den Datenbeständen wachsen auch die Kosten für die Datensicherung pro Kopie. Dadurch können nur weniger Backup-Kopien finanziert werden und damit wächst die Abhängigkeit.
 - Je unabhängiger die Kopien sind, um so sicherer. Mit wachsenden Datenbeständen stehen weniger Storagetechniken zur Verfügung, die man sich leisten kann. Dadurch sinkt auch der Grad der Unabhängigkeit.
 - Je häufiger Kopien gezogen werden, um so sicherer. Mit wachsenden Datenbeständen steigen auch die Zeiten und Kosten, um Fehler zu entdecken und reparieren. Dadurch sinkt die Häufigkeit für Backups.
- Beschränkte Aussagekraft von MTDDL
 - Für das ST5800 Honeycomb-Storagesystem der Fa. Sun wurde eine MTDDL von $2,4 \times 10^6$ Jahren angegeben. Das macht scheinbar jedes Backup überflüssig. Dem würde aber keiner folgen. Angenommen wir haben 10 ST5800 im Einsatz, von denen n Systeme, $1 \leq n \leq 5$, bereits nach einem Jahr ausfallen, $10 - 2n$ Systeme nach $2,4 \times 10^6$ Jahren und n Systeme nach $4,8 \times 10^6$ Jahren. Trotzdem wäre die MTDDL der ST5800 $2,4 \times 10^6$ Jahre.



3. RAID

- Redundant Array of Inexpensive Disks
 - Performanz und Ausfallsicherheit
 - Besseres Preis-/Leistungsverhältnis als bei SLED
 - Viele Anwendungen brauchen Platz größer als eine Platte
 - Hat sich in den letzten 20 Jahren bis in den Heimbereich durchgesetzt
 - Der Begriff wurde von David Patterson, Garth Gibson und Randy Katz 1988 geprägt (s. [10]). Sie prägten die Levels 1,2,3,4 und 5.
 - Später erweiterten sie mit Peter Chen und Edward Lee um die Levels RAID 0 und RAID 6.
 - RAID 0 ist kein RAID im eigentlichen Sinn. Es hat keine Redundanz. Jeder Ausfall einer Platte führt zum Datenverlust.

- Redundant Array of Inexpensive Disks

- RAID 6 kann den Ausfall zweier Platten verkraften. Der Ausfall einer dritten Platte führt zum Datenverlust.
- RAID 1, 3 bis 5 haben eine einfache Redundanz und können den Ausfall einer Platte verkraften. Sie sind dann im Zustand einer RAID 0: Der Ausfall einer weiteren Platte führt zum Datenverlust.
- Mittlerweile wird der Begriff als „Redundant Array of Independent Disks“ gelesen. Der Preisvorteil gilt nicht mehr.

SLED



SLED: Laufwerksmodul einer IBM 3380, die dem Vergleich mit kleineren Laufwerken standhalten musste



Eine Fujitsu M2351 Eagle; ein ähnliches Modell wurde von Patterson, Gibson und Katz verglichen

[2]

- Datenabbildung auf Plattenarray

- Streifeneinheit (stripe unit): Datenmenge, die auf einer Platte gespeichert wird, bevor mit der nächsten fortgefahren wird.
- Paritätsstreifen (parity stripe): Minimale Zusammenfassung von Streifeneinheiten für die eine Parität berechnet wird.
- Zeile (row): Minimale Anzahl von Platten über die sich ein Paritätsstreifen erstreckt.
- Spalte (column): In unserem Fall eines eindimensionalen Arrays eine Platte
- Paritätsanordnung: Verteilung der Parität auf das Plattenarray. Verschiedene Varianten sind möglich.

• Paritätsanordnung

1. RAID level 4
2. Right-Assymmetric
3. Left-Assymmetric
4. Left-Symmetric
5. Right-Symmetric
6. Extended-Left-Symmetric
7. Flat-Left-Symmetric

Ausführliche Diskussion in [37]:
Die Wahl ist zwischen 4, 6 und 7 zu treffen.

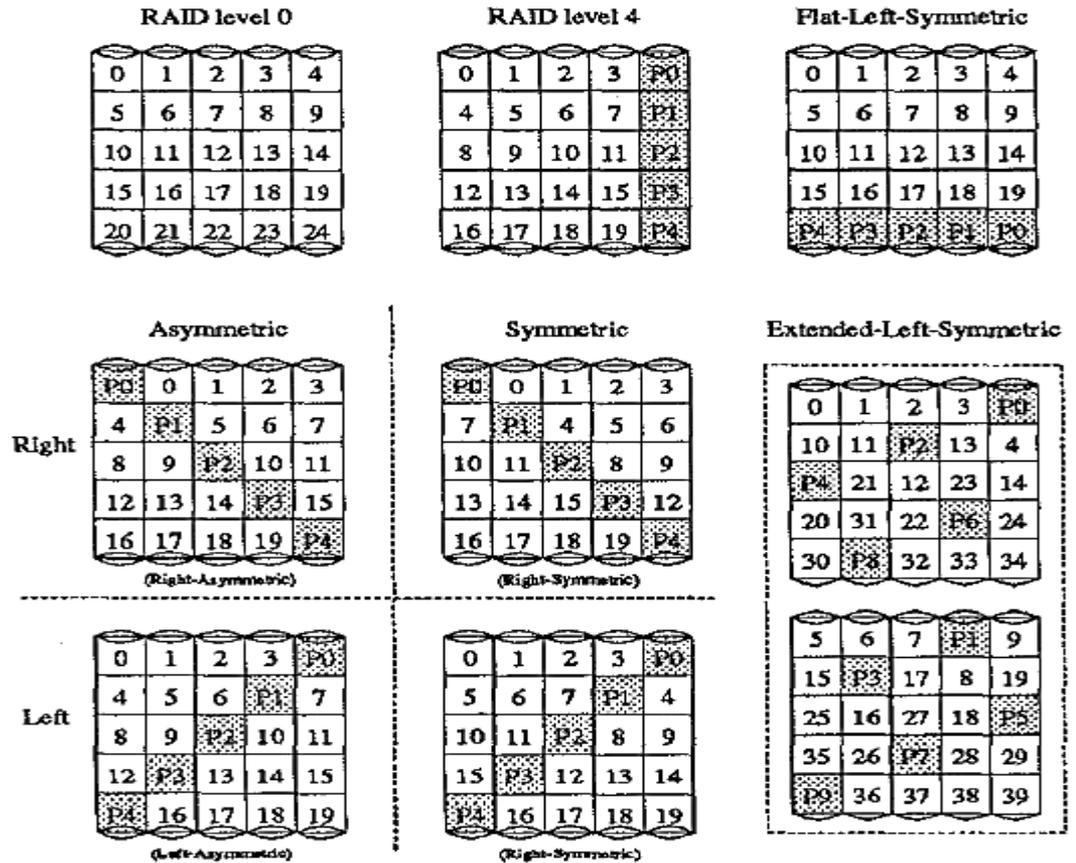


Figure 3: Parity Placements. Each square corresponds to a stripe unit. Each column of squares corresponds to a disk. Each matrix corresponds to a single row of disks. In all placements, P_0 computes parity over stripe units 0, 1, 2 and 3; P_1 computes parity over stripe units 4, 5, 6 and 7; etc. In all cases except the RAID level 0 and RAID level 4 placements, the minimum repeating pattern is shown.

[37]

3. RAID - DDF

- Disk Data Format (DDF)

Notwendige Angaben zu den unterschiedlichen RAID-Level werden hierdurch auf der Platte angegeben.

- RAID-Aufbau

Abhängig von Produkten wie z.B. Kontroller oder herstellerabhängig?
Oberhalb von RAID 5 verschiedene z.T. herstellerabhängige Formate.

- Storage Networking Industry Association (SNIA)

Eine Non-Profit-Handelsvereinigung von Produzenten und Verbrauchern zum Zwecke einer Standardisierung von Speichernetzwerken des Datenformats für RAID-Systeme (DDF)

<http://www.snia.org/>

http://www.snia.org/tech_activities/standards/curr_standards/ddf

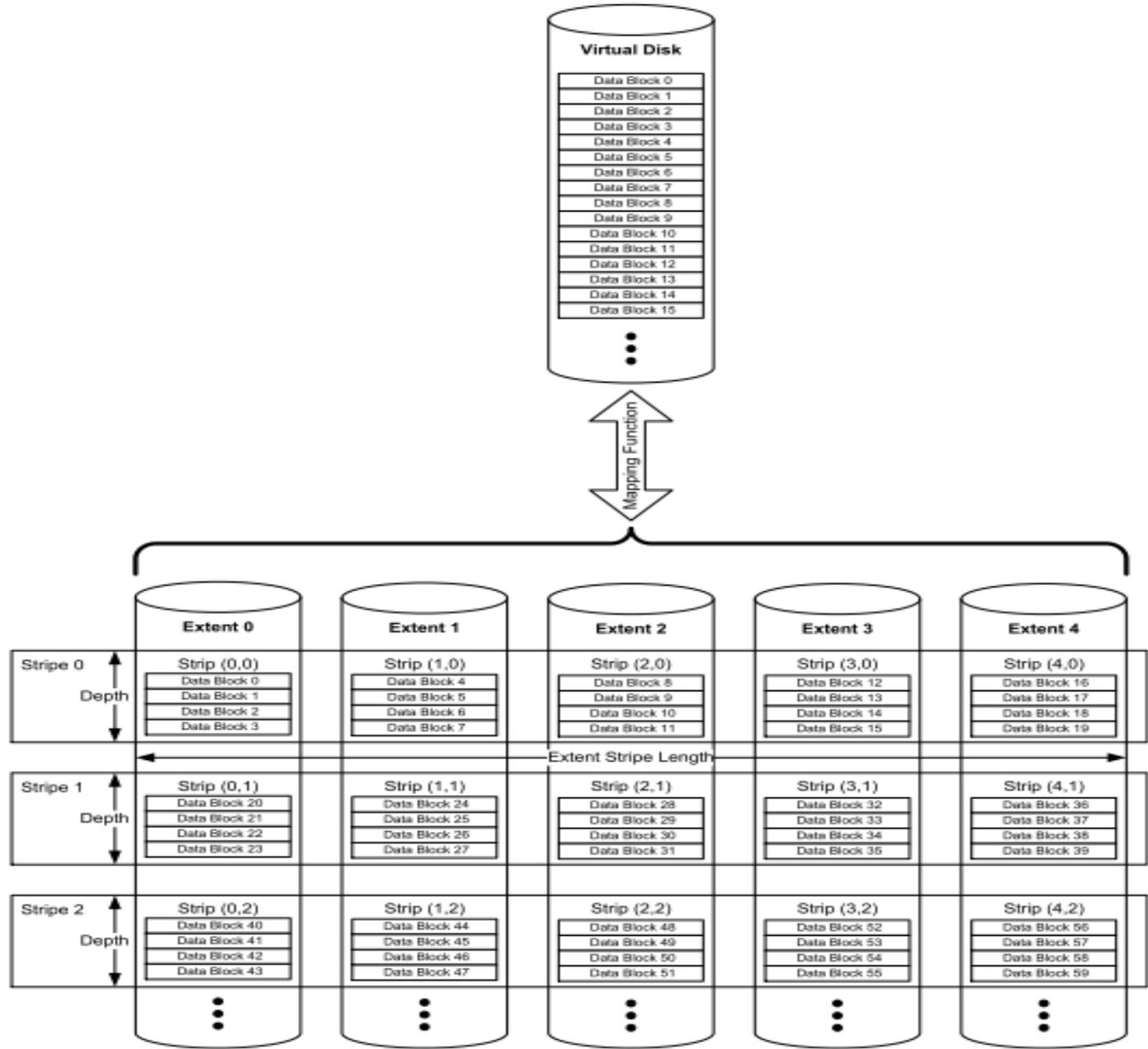
3. RAID - DDF

- Disk Data Format (DDF)
 - Beschreibt Datenformatierung über mehrere Platten im RAID
 - Ermöglicht ein Mindestmaß an Interoperabilität zwischen verschiedenen Herstellern von RAID Technologie
 - Übernahme von Userdaten in RAID Level 1
 - Informationen zur RAID-Struktur auf Daten- und Paritätsplatten, Hotspare-Platten und NVRAM des Controllers
 - Partitionierung, RAID Level und Cache-Parameter für jede virtuelle Platte
 - Größe der Streifen (stripes; abhängig von der Plattenanzahl)
 - Tiefe der Streifen (depth; #Cluster pro Platte und Streifen)
 - Unterschiedliche Lage der Parität bei Rotation
 - L5: linke oder rechte Diagonale
 - L3: erste oder letzte Platte
 - L1E:

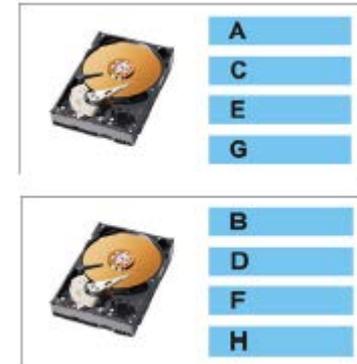
3. RAID - DDF

- Disk Data Format (DDF)
 - Varianten mit Hot-Spare-Bereiche
 - Zusätzliche Platte
 - Reservierung auf Daten- und Paritätsplatten

Von SNIA
 Extent (Platte)
 Stripe (Streifen)
 Depth, Chunk (Tiefe)



- RAID 0
 - Striping
 - Kein echtes RAID-Verfahren
 - Daten parallel auf alle Festplatten verteilt
 - Bei Plattenausfall sind die Daten des kompletten Raids verloren.



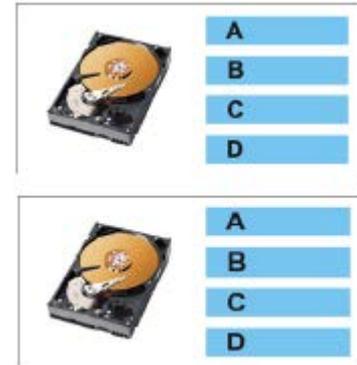
Mindestanzahl n benötigter Festplatten: $n=2$

Maximal verfügbare Nutzkapazität k : $k=n$

Ausfallsicherheit s (in #Festplatten): $s=0$

- RAID 1

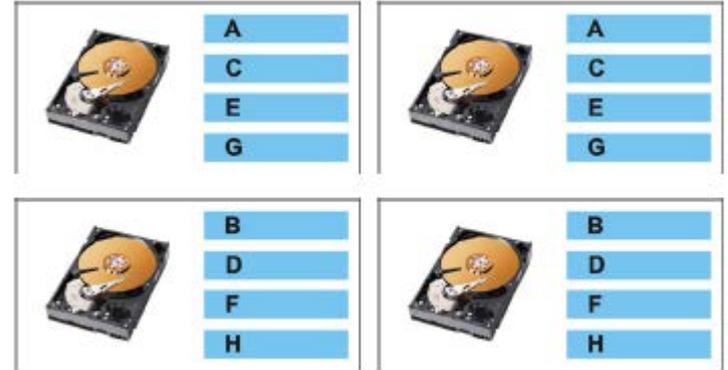
- Mirroring/Spiegelung
- 100% Redundanz
- Daten parallel auf zwei Festplatten gespeichert
- Bei Plattenausfall werden Daten vom Spiegel gelesen.



Mindestanzahl n benötigter Festplatten: n=2
Maximal verfügbare Nutzkapazität k: k=1
Ausfallsicherheit s (in #Festplatten): s=1

- RAID 01

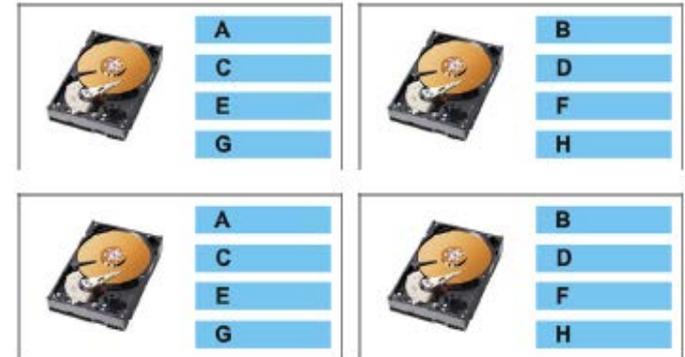
- Weitere Bezeichnungen RAID 0+1, RAID 0,1 oder RAID 0/1
- Kombiniert ein RAID 0 mit einem RAID 1 (ein RAID 0 wird gespiegelt), um die Transferrate zu erhöhen bei gleichzeitiger 100%iger Redundanz.



Mindestanzahl n benötigter Festplatten: n=4
Maximal verfügbare Nutzkapazität k: k=2
Ausfallsicherheit s (in #Festplatten): s=1

- RAID 10

- Kombiniert ein RAID 1 (Spiegelung) mit einem RAID 0 (Striping), um die Transferrate zu erhöhen bei gleichzeitiger 100%iger Redundanz.



- Der Unterschied zwischen RAID 10 und RAID 01 zeigt sich im Fehlerfall:

- Fällt bei RAID 01 eine Festplatte aus, so ist das betroffene RAID 0 unbrauchbar. Die Daten sind aber über das zweite RAID 0 (den Spiegel) verfügbar. Fällt eine weitere Platte aus dem zweiten Verbund aus, so sind die Daten verloren.
- Fällt beim RAID 10 eine Platte aus, so ist bei einem weiteren Ausfall einer Platte nur dann die Datenintegrität gefährdet, wenn sie die gleiche Position wie der erste Ausfall hat.
- Die Datenrekonstruktion geht bei RAID 01 im schlimmsten Fall über $n/2$ Platten, während bei RAID 10 sie nur über eine Platte geht. Die defekte Platte wird ersetzt und die Daten vom zugeordneten Spiegel werden kopieren. Ab $n \geq 6$ macht es sich bemerkbar.

Mindestanzahl n benötigter Festplatten: $n=4$
 Maximal verfügbare Nutzkapazität k : $k=2$
 Ausfallsicherheit s (in #Festplatten): $s=1$

• RAID 2

- Stammt aus Großrechnerbereich als Platten keine eigenen ECC-Mechanismen zum Schutz der Daten hatten.
- Die einzelnen Bits werden auf 8 Datenplatten geschrieben und zusätzlich 2 Bits nach dem Hamming-Algorithmus berechnet und auf Platte 9 und 10 geschrieben.
- Bei Bitfehlern kann auch die genaue Position bestimmt werden.
- Die Transferrate beim Lesen erhöht sich durch parallelen Zugriff auf das achtfache. Beim Schreiben liegt die unter Transferrate der unter einer einzelnen Platte.
- Parallele I/O-Operationen sind nicht möglich. Sind die Platten nicht synchronisiert, ist der Datentransfer geringer.



Mindestanzahl n benötigter Festplatten: $n=10$

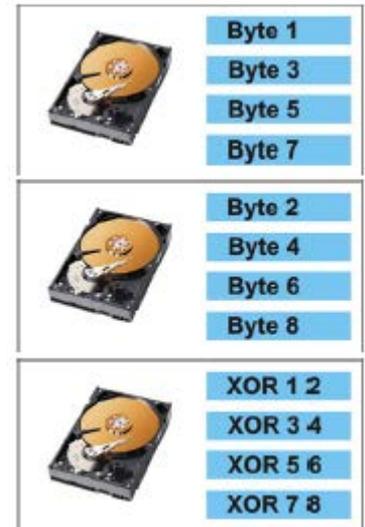
Maximal verfügbare Nutzkapazität k: $k=n-2$

Ausfallsicherheit s (in #Festplatten): $s=2$

• RAID 3

- Vorgänger von RAID 4 und 5
- Verfügt über Redundanz
- Daten werden byteweise auf mindestens zwei Festplatten gespeichert.
- Über Exklusives-ODER (XOR – entspricht einer Addition modulo 2) werden Paritätsbits aus den Daten berechnet und byteweise auf einer Paritätsplatte gespeichert.
- Die Paritätsplatte ist der ‚Flaschenhals‘.
- Synchronisiert die Kopfbewegungen der Laufwerke von Daten- und Paritätsplatte beim Schreiben zur Erhöhung des Durchsatzes und Minimierung der Kopfbewegungen.
- Parallele I/O nicht möglich. Beim Lesen und Schreiben müssen immer alle Platten angesprochen werden.

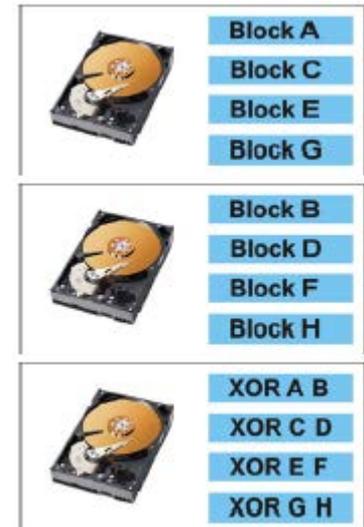
A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0



Mindestanzahl n benötigter Festplatten: $n=3$ (bei $n=2$ entspricht es RAID 1)
 Maximal verfügbare Nutzkapazität k: $k=n-1$
 Ausfallsicherheit s (in Anzahl an Festplatten): $s=1$

- RAID 4

- Vorgänger RAID 3
- Verfügt ebenfalls über Redundanz
- Daten werden blockweise auf mindestens zwei Festplatten gespeichert.
- Über XOR werden aus den Daten Paritätsbits und blockweise auf eine dedizierte Paritätsplatte gespeichert.
- Die Paritätsplatte ist der ‚Flaschenhals‘.
- Die Kopfbewegungen der Laufwerke werden beim Lesen nicht synchronisiert. Dies hat Vorteile beim Lesen, aber das Schreiben dauert dadurch länger, weil die passende Stelle auf der Paritätsplatte gefunden werden muß.



Mindestanzahl n benötigter Festplatten:

n=3 (bei n=2 entspricht es RAID 1)

Maximal verfügbare Nutzkapazität k:

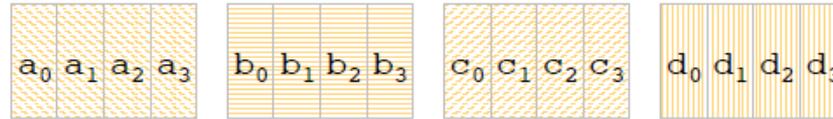
k=n-1

Ausfallsicherheit s (in #Festplatten):

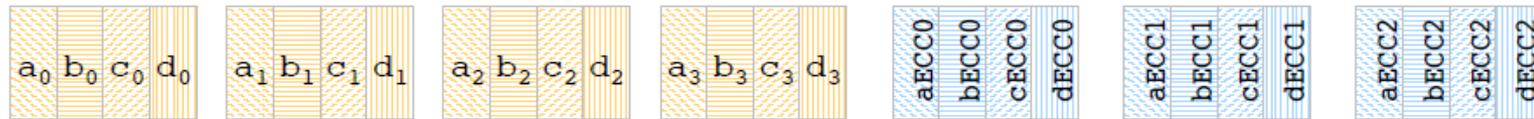
s=1

Vergleich der RAID-Leves 2, 3 und 4

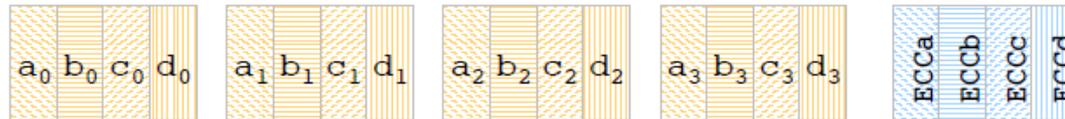
**4 Dateneinheiten
a, b, c, d sind zu
speichern**



RAID-2

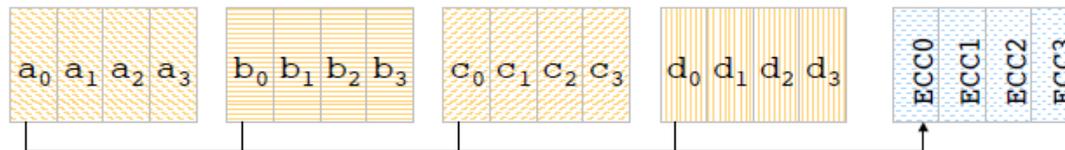


RAID-3



**Parität wird über
jeder Dateneinheit
berechnet**

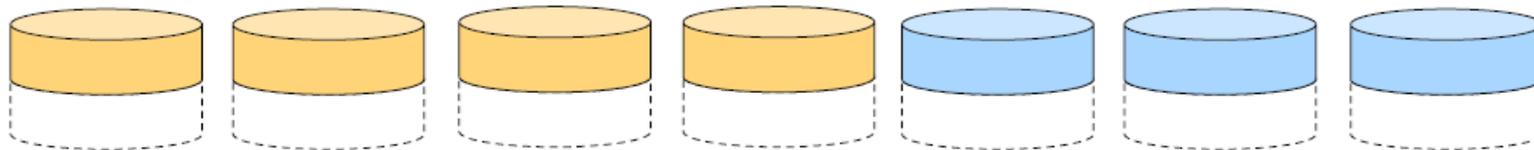
RAID-4



**Parität wird über
einem Teil jeder
Dateneinheit
berechnet**

Data disks

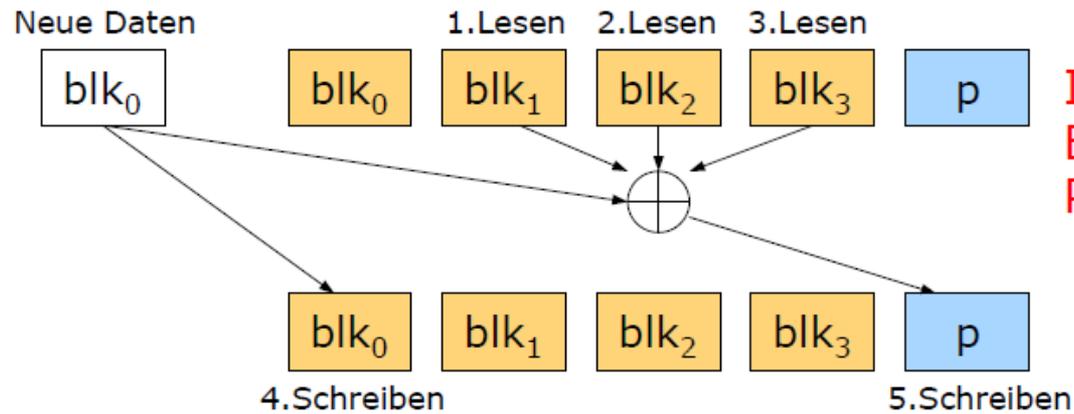
Check disk(s)



[32]

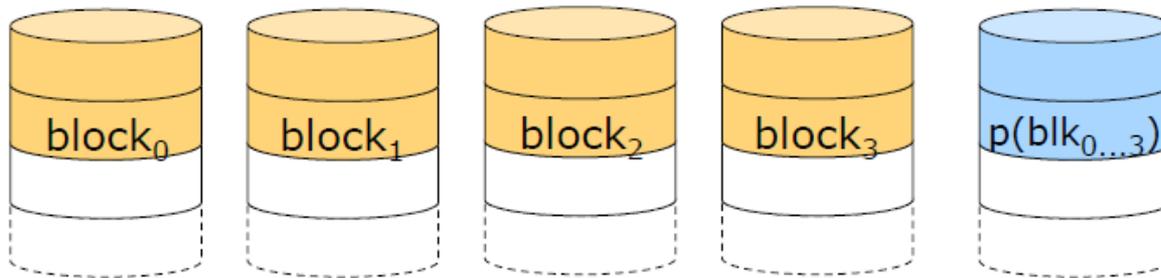
RAID 4 • Merkmale

- Verteilung (striping): sector interleaving
- Paritätsinformation auf zusätzlicher Platte



Data disks

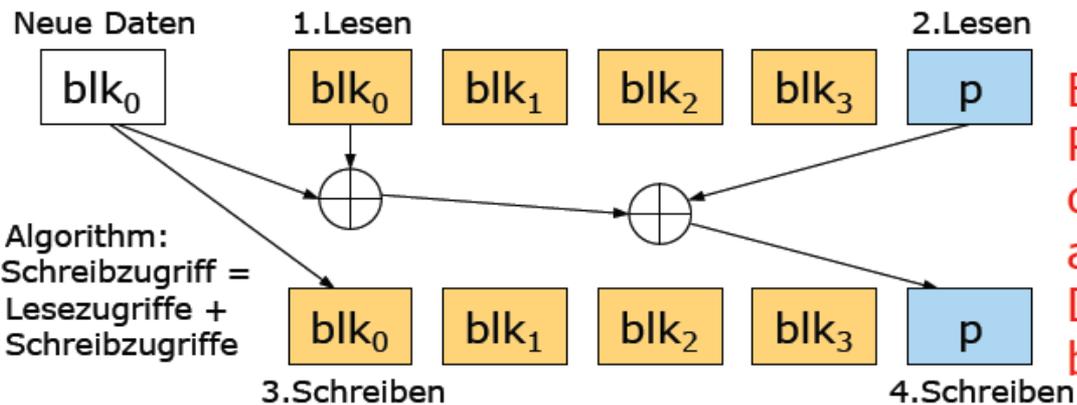
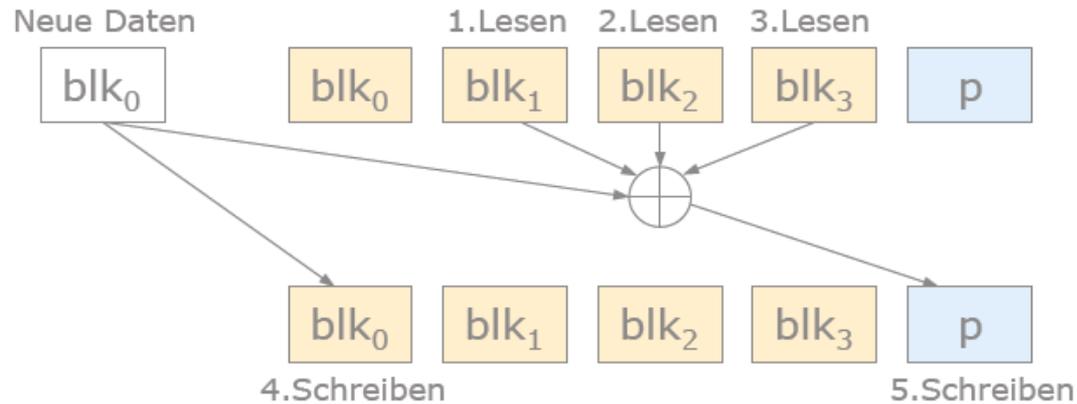
Check disk



[32]

RAID 4 • Merkmale

- Verteilung (striping): sector interleaving
- Paritätsinformation auf zusätzlicher Platte



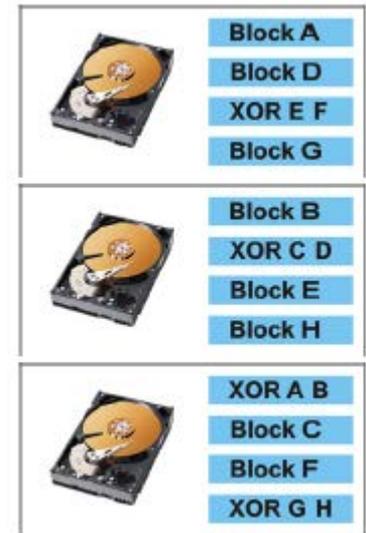
Besser: Neue Parität wird aus der Differenz des alten und neuen Datenblocks berechnet

Small-Write Algorithm:
 1 Logischer Schreibzugriff =
 2 physische Lesezugriffe +
 2 physische Schreibzugriffe

[32]

• RAID 5

- Vorgänger sind RAID 3 und 4
- Verfügt ebenfalls über Redundanz
- Die Daten werden blockweise verteilt. Die dedizierte Paritätsplatte entfällt. Der Flaschenhals entfällt. Die ECC-Daten werden gleichmäßig über alle Laufwerke verteilt.
- Beim Lesen Transferraten wie RAID 4. Der Schreibzugriff wird durch die gleichmäßige Verteilung der Paritätsblöcke schneller als RAID 4, aber natürlich noch deutlich langsamer als bei RAID 0.



Mindestanzahl n benötigter Festplatten:

$n=3$ (bei $n=2$ entspricht es RAID 1)

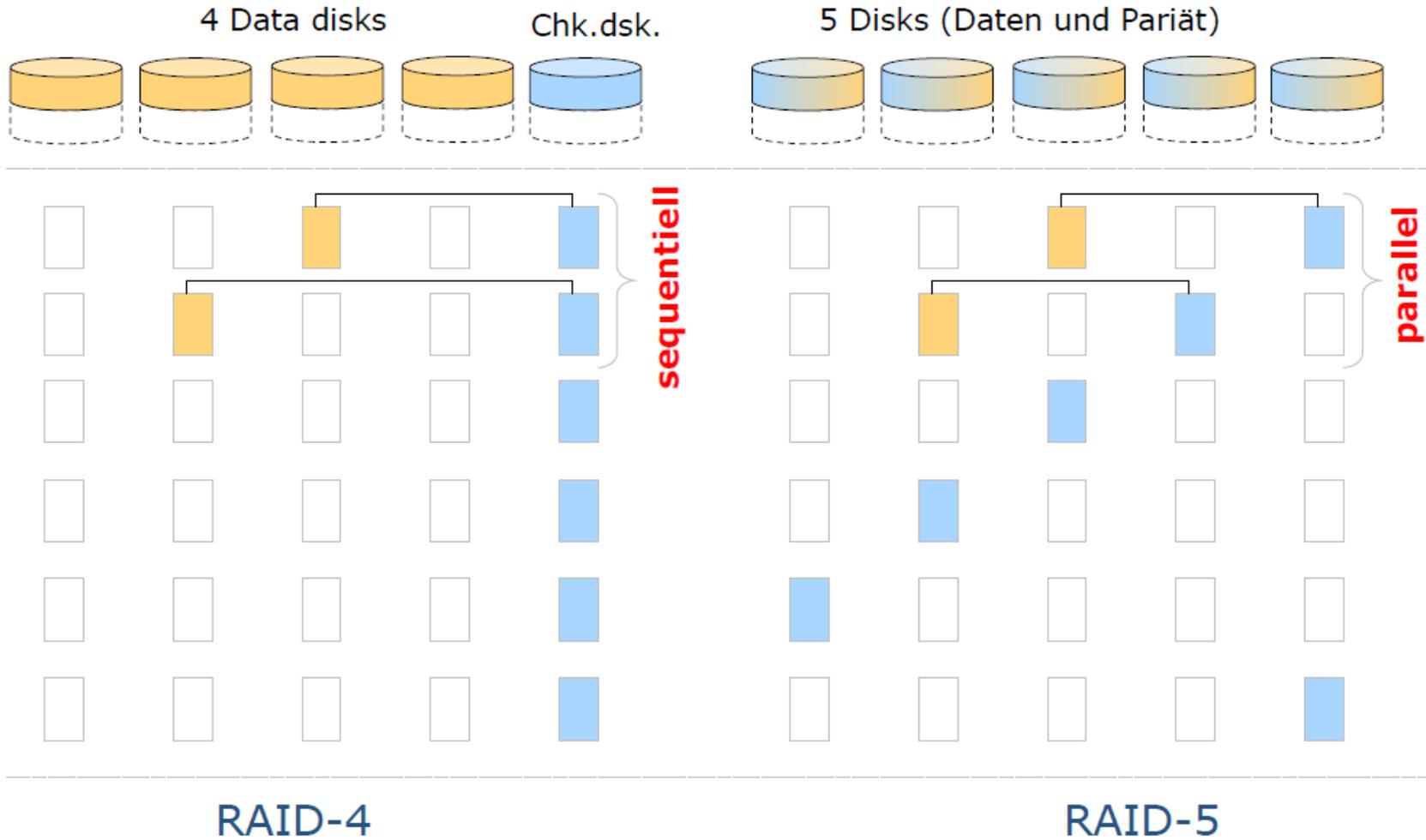
Maximal verfügbare Nutzkapazität k :

$k=n-1$

Ausfallsicherheit s (in #Festplatten):

$s=1$

Vergleich von RAID-4 und RAID-5 (Schreiben von Daten)



[32]

- Weitere RAID-Verfahren

Es reicht in der Praxis nicht aus, nur den Ausfall einer Platte aufzufangen. Gerade nach dem Ausfall einer Platte kann es bei der Wiederherstellung des RAIDs zu weiteren Ausfällen kommen.

Je größer die Systeme werden, um so mehr Zeit wird fürs Rebuild benötigt und um so größer ist die Wahrscheinlichkeit weiterer Plattenausfälle. Treten Bitfehler beim Rebuild auf, so führt dies ebenfalls zu Datenverlust. Deshalb wurden weitere RAID-Verfahren entwickelt, die den Ausfall von maximal 2 Platten kompensieren können.

So muß bei einem Array mit 60 Laufwerken unter normalen Betriebsbedingungen und gleichen Fehlerwahrscheinlichkeiten bei einem RAID 5 nach 6 Monaten durch zweifachen Totalausfall oder einem Totalausfall und einem zusätzlichen Bad Block Failure gerechnet werden.

- Weitere RAID-Verfahren

Auch wenn die Anzahl 60 etwas „übertrieben“ erscheint, RAID-Systeme haben die Tendenz, stark zu wachsen. Und nicht immer bleibt es bei einem Ausfall einer Platte. Häufig sind weitere in Mitleidenschaft gezogen, was sich erst beim Rebuild zeigt. Dann war die Annahme der Unabhängigkeit von Fehlern eventuell zu optimistisch.

Die wachsende Plattenanzahl erhöht das Risiko von Datenverlust (MTTDL).

Es ist kein Ausweg, die Platten durch größere dafür aber weniger Platten bei gleicher Gesamtkapazität zu ersetzen. Allenthalben wird der Zeitpunkt kurzfristig hinausgeschoben. Das zeigt die folgende Folie.

Mit heutigen Plattengrößen, MTTF-Angaben zu neueren Platten und steigender Plattenanzahl ergeben vergleichbare Situationen.

Die Grafik gibt zum Vergleich das RAID-6-Verfahren an.



Eine Prüfsumme ist nicht genug:

Alle Platten haben die gleiche MTTF von 200.000h.

Jeder Ausfall senkt die MTTF der verbleibenden Platten um den Faktor 10.

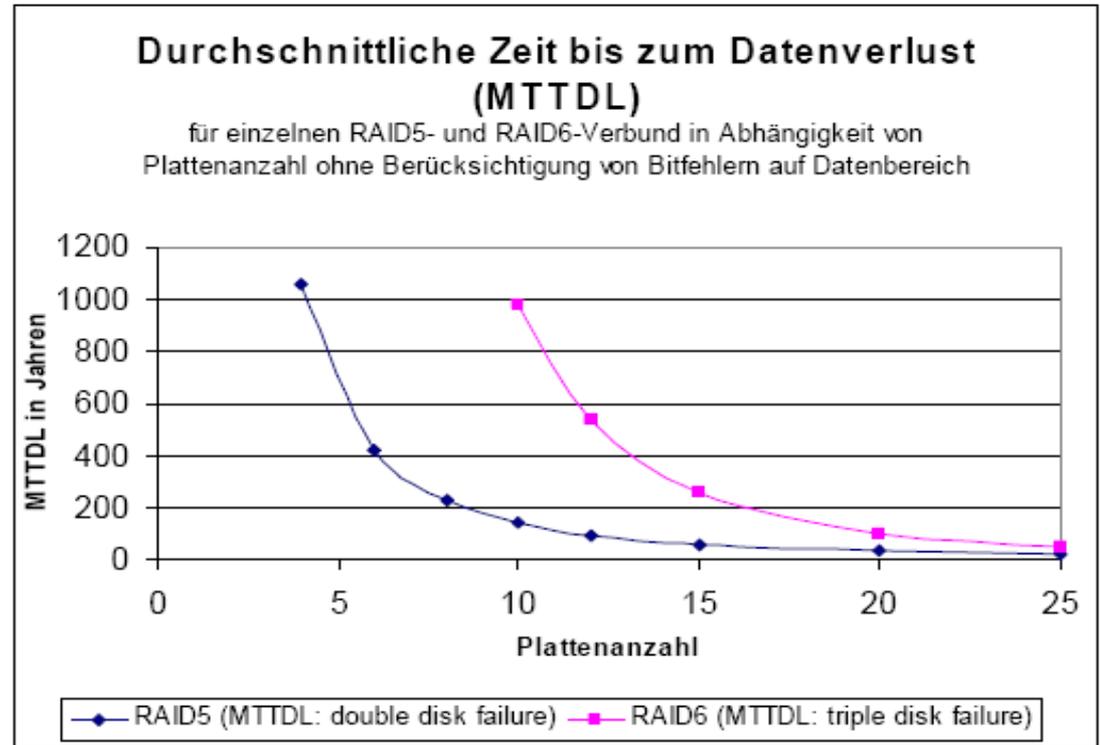
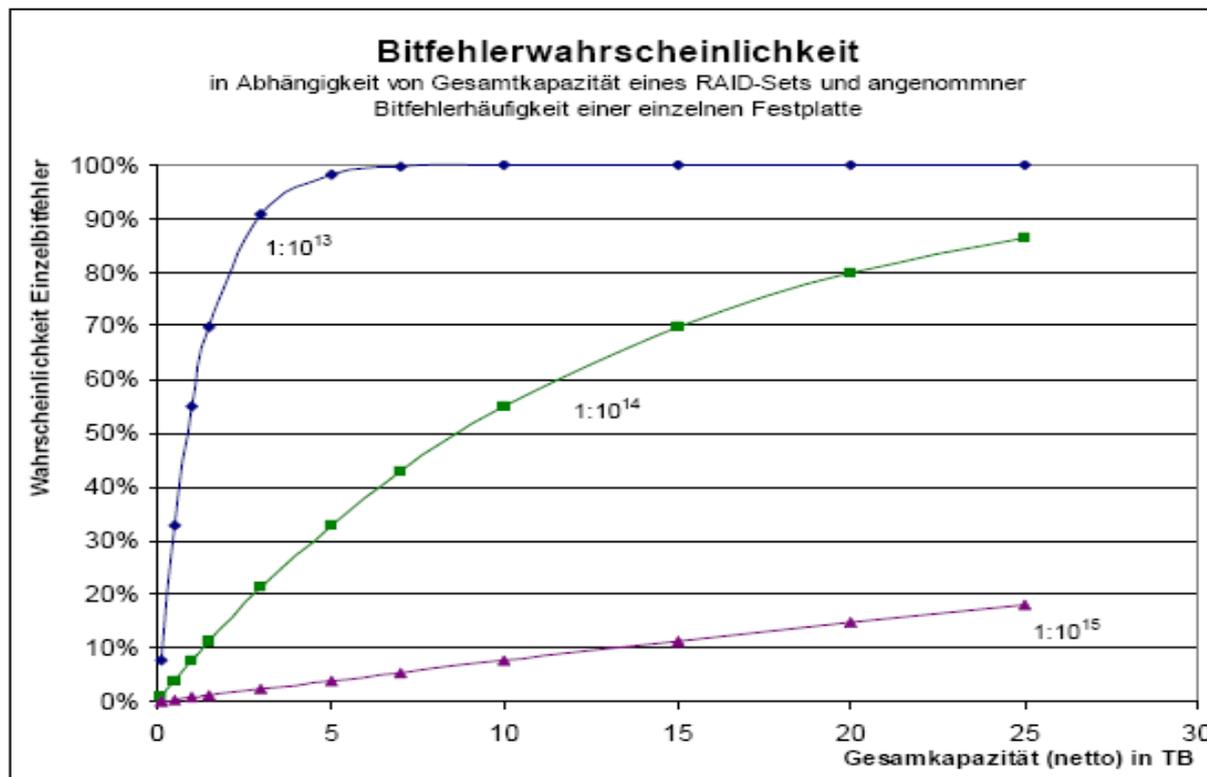


Abbildung 1: Berechnungsgrundlage $MTTF(\text{disk1})=200000\text{h}$, $MTTF(\text{disk2})=20000\text{h}$, $MTTF(\text{disk3})=2000\text{h}$, $MTTR=36\text{h}$

[19]



Eine Prüfsumme ist nicht genug:



[19]

Abbildung 2: Bitfehlerwahrscheinlichkeiten für unterschiedliche Bitfehlerraten

Aussagekraft von MTDDL beschränkt.
Eine MTDDL von 100 Jahren ergibt
ein Restrisiko von 5% bei 5 Jahren
Betriebsdauer

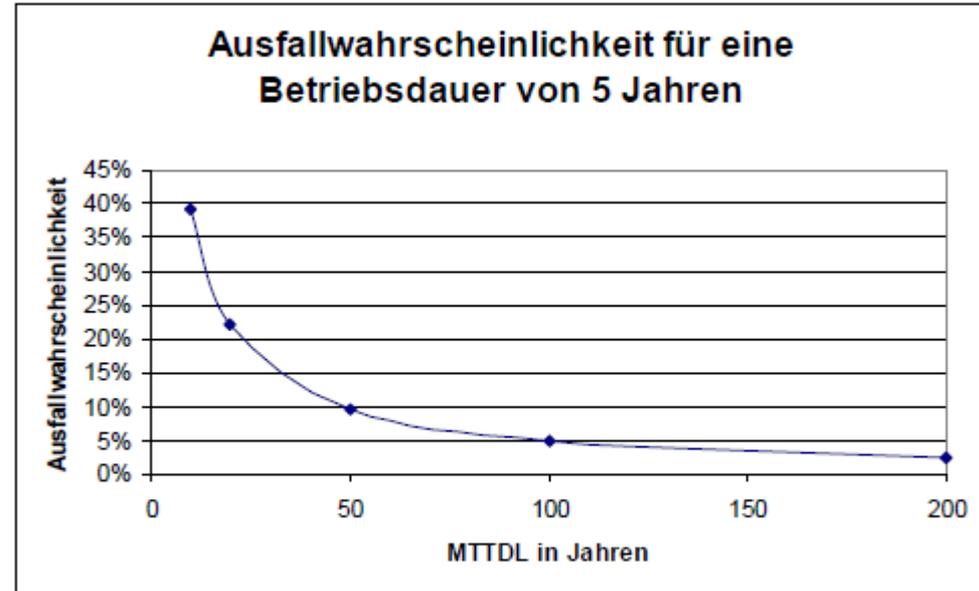


Abbildung 3: MTDDL und Ausfallwahrscheinlichkeit

[19]

- RAID 6

RAID 6 arbeitet mit zwei Paritätsplatten. Da die Paritäten unabhängig von einander sind, stehen zwei unabhängige Werte zur Fehlerkorrektur zur Verfügung. Maximal zwei Fehler können damit korrigiert werden.

Beide Paritäten werden gleichmäßig über alle Platten verteilt. Die erste Parität wird wieder mit XOR ermittelt und erfordert nicht viel Aufwand. Die zweite wird mittels Reed-Solomon-Codes berechnet. Der Rechenaufwand zum Erstellen der Korrekturwerte ist erheblich höher.

- RAID 6

Horizontale und diagonale Prüfsummen P und Q mittels Reed-Solomon-Code (bei 2^8 Platten => P durch XOR).

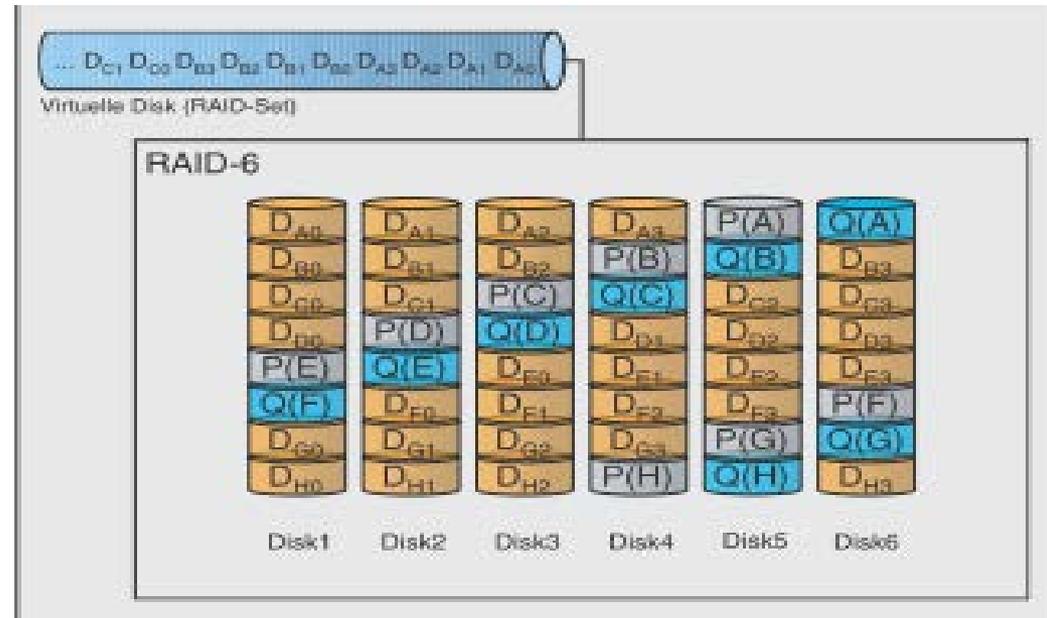
Die Berechnung mit endlichen Körpern ist aufwendig. Durch HW-Unterstützung wird ein Teil des Aufwands aufgefangen. Bei

gleicher Plattenzahl ist die Lesegeschwindigkeit etwas und die Schreibgeschwindigkeit deutlich geringer (30% werden genannt) als bei RAID 5. Weitere RAID-Verfahren mit doppelter Redundanz werden mit dem RAID-6-Verfahren verglichen.

Mindestanzahl n benötigter Festplatten: $n \geq 4$

Maximal verfügbare Nutzkapazität k: $k = n - 2$

Ausfallsicherheit s (in #Festplatten): $s = 2$



- Zweifaches XOR im RAID-4 Stil - RAID DP

- Network Appliance

Weiterentwicklung von RAID 4: Zwei Paritäten P und Q, wobei die erste wie bisher horizontal, die zweite unabhängig von der ersten Parität „diagonal“ unter ihrer Einbeziehung gebildet wird. In der Regel bestehen RAID-DP-Sets aus 14+2 Platten und haben ein ähnliches Brutto/Netto-Verhältnis wie RAID 4 und 5. Zur Vermeidung „typischer“ RAID-4-Probleme wird RAID DP zusammen mit dem Dateisystem WAFL eingesetzt.

$$P_1 = XOR(A_1, B_1, C_1)$$

$$P_2 = XOR(A_2, B_2, C_2)$$

$$P_3 = XOR(A_3, B_3, C_3)$$

$$Q_1 = XOR(A_1, B_2, C_3, 0)$$

$$Q_2 = XOR(0, B_1, C_2, P_3)$$

$$Q_3 = XOR(A_3, 0, C_1, P_2)$$



Mindestanzahl n benötigter Festplatten: n=5

Maximal verfügbare Nutzkapazität k: k=n-2

Ausfallsicherheit s (in #Festplatten): s=2

- Beispiel

Wir zeigen jetzt die Vorgehensweise für fünf Daten- und zwei Paritätsplatten. Die erste Parität P wird wie bisher durch Exklusives-Oder bestimmt:

$$P_1 = XOR(A_1, B_1, C_1, D_1)$$

$$P_2 = XOR(A_2, B_2, C_2, D_2)$$

$$P_3 = XOR(A_3, B_3, C_3, D_3)$$

$$P_4 = XOR(A_4, B_4, C_4, D_4)$$

A	B	C	D	P	DP	
1	1	0	1	1	1	= XOR(1,1,1,0)
1	1	0	1	1	0	= XOR(1,0,0,1)
0	1	1	0	0	0	= XOR(0,1,0,1)
1	1	1	0	1	1	= XOR(1,1,0,1)

Die zweite Parität DP wird über die

Diagonalen, die durch gleichfarbige Blöcke hervorgehoben sind, durch XOR gebildet und in die gleichfarbigen Stelle in DP gespeichert. Die Paritätsplatte P wird dabei einbezogen. Die weißen Stellen werden ausgelassen:

$$DP_1 = XOR(A_1, B_2, C_3, D_4, 0)$$

$$DP_2 = XOR(0, B_1, C_2, D_3, P_4)$$

$$DP_3 = XOR(A_4, 0, C_1, D_2, P_3)$$

$$DP_4 = XOR(A_3, B_4, 0, D_1, P_2)$$

[35]

- Wiederherstellung nach Plattenausfall
Angenommen, Platte A und B fallen aus und werden ausgetauscht.

A	B	C	D	P	DP
		0	1	1	1
		0	1	1	0
		1	0	0	0
		1	0	1	1

Pro Zeile fehlen zwei Daten. Die erste Diagonale mit nur einem fehlenden Datum (B_1) ist die blaue Diagonale. Es kann mit dem „blauen“ Datum von Platte DP berechnet werden:

A	B	C	D	P	DP
	1	0	1	1	1
		0	1	1	0
		1	0	0	0
		1	0	1	1

1a) $XOR(0,0,0,1)=1$

Im zweiten Schritt wird mit der horizontalen Parität P_1 das Datum das fehlende Datum A_1 berechnet. Die erste Zeile ist jetzt wiederhergestellt:

A	B	C	D	P	DP
1	1	0	1	1	1
		0	1	1	0
		1	0	0	0
		1	0	1	1

1b) $XOR(1,1,0,1)=1$

A_1 gehört zur roten Diagonalen, in der jetzt nur noch B_2 als Datum fehlt. Wir berechnen B_2 wie im ersten Schritt mit Diagonalwerten, nur diesmal von der roten Diagonalen:

A	B	C	D	P	DP
1	1	0	1	1	1
	1	0	1	1	0
		1	0	0	0
		1	0	1	1

1)
2a) $XOR(1,1,1,0)=1$

Danach erhält man A_2 mit der horizontalen Parität P_2 usw.

- Wiederherstellung nach Plattenausfall

Das Verfahren kann auch schön mit ganzen Zahlen illustriert werden. XOR entspricht einer Addition modulo 2. („Zufällig“ erhält damit das erste Beispiel modulo 2.)

A	B	C	D	P	DP	
3	1	2	3	9	7	= 3+1+1+2
1	1	2	1	5	12	= 1+2+2+7
2	3	1	2	8	12	= 2+1+8+1
1	1	3	2	7	11	= 3+5+2+1

Angenommen, Platte A und B fallen aus und werden ausgetauscht.

A	B	C	D	P	DP
		2	3	9	7
		2	1	5	12
		1	2	8	12
		3	2	7	11

Wir beginnen wieder mit dem fehlenden Datum B_1 in der blauen Diagonale, das mit dem „blauen“ Datum von Platte DP berechnet wird. Danach wird die Parität P_1 für A_1 genutzt:

A	B	C	D	P	DP
	1	2	3	9	7
		2	1	5	12
		1	2	8	12
		3	2	7	11

1a) $12-7-2-2=1$

A	B	C	D	P	DP
3	1	2	3	9	7
	1	2	1	5	12
		1	2	8	12
		3	2	7	11

1)

2a) $7-2-1-3=1$

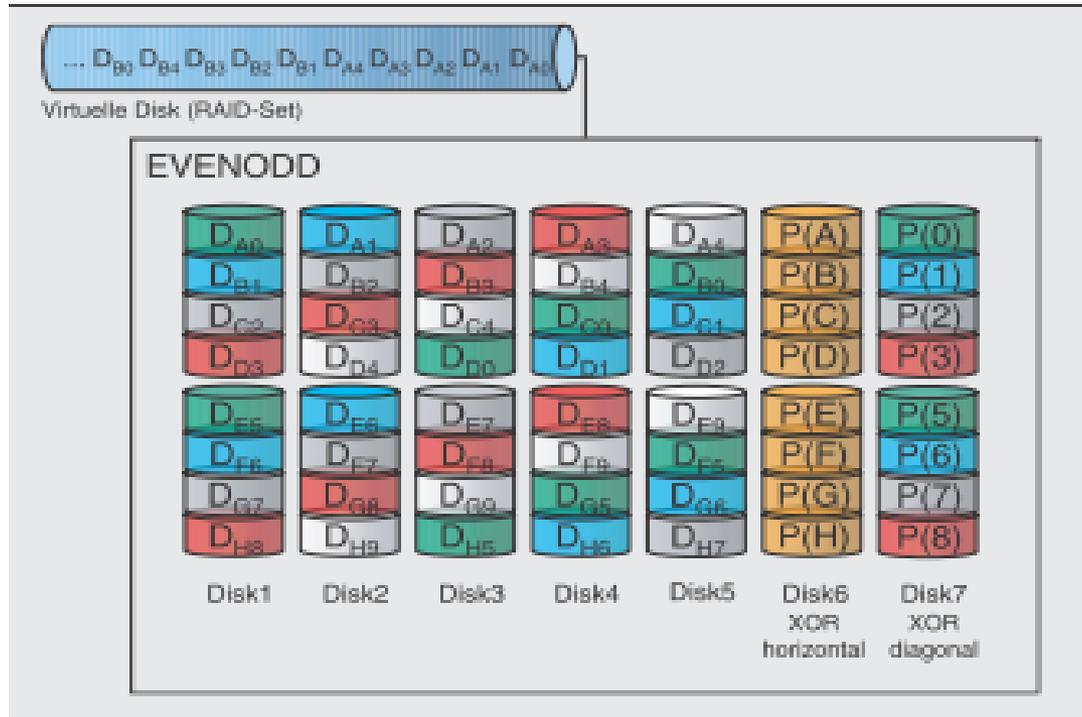
A	B	C	D	P	DP
3	1	2	3	9	7
		2	1	5	12
		1	2	8	12
		3	2	7	11

1b) $9-3-2-1=3$

Und so weiter

- Zweifaches XOR im RAID-4 Stil
 - RAID 5DP (RAID 5 Double Parity von HP)
 - RAID ADP von HP
- EVENODD
 - 1994
 - Setzt ausschließlich auf Paritäten als Redundanz
 - Zwei Plattenausfälle werden toleriert
 - Anzahl der Datenplatten muß eine Primzahl m sein, was keine wesentliche Einschränkung ist, indem fehlende Platten als „Nullplatten“ angesetzt werden.
 - Anzahl der Knoten $m + 2$ davon m Daten- und 2 Paritätsplatten
 - Paritäten können auch auf alle Platten verteilt werden wie bei RAID-6
 - Funktioniert auf Bit-, Byte-, Blockebene usw. Für die Darstellung werden deshalb in der Literatur häufig Symbole verwendet

- EVENODD



- EVENODD

- Definition von Blöcken als Matrix mit Dimension $(m-1) \times (m+1)$ von Elementen a_{ij}
- Element a_{ij} mit $0 < i < m-2$ and $0 < j < m$ ist ein Symbol j auf Platte i
- Platten m and $m+1$ enthalten die redundante Information
- Gedankliche Nullreihe $a_{m-1,j}$ als letzte Zeile zur Vereinfachung

- Berechnung der horizontalen Parität:
$$a_{l,m} = \bigoplus_{t=0}^{m-1} a_{l,t} \quad (1)$$

- Berechnung der Diagonalparitäten
$$a_{l,m+1} = S \oplus \left(\bigoplus_{t=0}^{m-1} a_{\langle l-t \rangle_m, t} \right) \quad (2)$$

 Schreibweise $\langle n \rangle_m = j \Leftrightarrow j \equiv n \pmod{m}$ und $0 \leq j \leq m-1$.

Bsp.: $\langle 7 \rangle_5 = 2$ und $\langle -2 \rangle_5 = 3$

- Diagonalparität S
$$S = \bigoplus_{t=1}^{m-1} a_{m-1-t,t} \quad (3)$$

- Beispiel

Wir berechnen jetzt für das Beispiel

D_1	D_2	D_3	D_4	D_5	P	$DP+S$	S	DP
1	0	1	1	0	1	0	1	1
0	1	1	0	0	0	0		1
1	1	0	0	0	0	1		0
0	1	0	1	1	1	0		1

mit fünf Datenplatten D_1 bis D_5 beide Paritätsplatten P und $DP+S$ nach den Gleichungen (1) bis (3) aus. Die horizontale Parität P wird über die Datenplatten wie gewohnt berechnet. Zur Bestimmung der zweiten Parität merken wir uns als Zwischenergebnis noch S und das Ergebnis über die Summe in (2).

- Wiederherstellung nach Plattenausfall

Wir nehmen an, daß die Platte D_1 und D_3 ausgefallen ist und zeigen deren Wiederherstellung.

- Wiederherstellung nach Plattenausfall

Angenommen, Platte D_1 und D_3 fallen und werden ausgetauscht.

D_1	D_2	D_3	D_4	D_5	P	DP+S	S
	0		1	0	1	0	1
	1		0	0	0	0	
	1		0	0	0	1	
	1		1	1	1	0	

Pro Zeile fehlen zwei Daten. Bei der blauen und grauen Diagonale fehlt nur ein Datum. Angenommen, wir kennen die Diagonalparität S . Damit lassen sich nach einander die fehlenden Werte berechnen. Wir starten mit der grauen Diagonalen:

D_1	D_2	D_3	D_4	D_5	P	DP+S	S
	0		1	0	1	0	1
	1		0	0	0	0	
	1	0	0	0	0	1	
	1		1	1	1	0	

1a) $S = 1$

Im zweiten Schritt wird mit der horizontalen Parität P_3 das Datum das fehlende Datum D_{13} berechnet. Die dritte Zeile ist jetzt wiederhergestellt:

D_1	D_2	D_3	D_4	D_5	P	DP+S	S
1	0	1	1	0	1	0	1
0	1	1	0	0	0	0	
1	1	0	0	0	0	1	
1	1	1	1	1	1	0	

1b) $P = 0$

In der gelben Diagonalen fehlt jetzt nur noch der Wert D_{31} . Wir berechnen ihn wie im ersten Schritt nur diesmal von der gelben Diagonalenwerten:

D_1	D_2	D_3	D_4	D_5	P	DP+S	S
1	0	1	1	0	1	0	1
0	1	1	0	0	0	0	
1	1	0	0	0	0	1	
1	1	1	1	1	1	0	

2a) $DP+S = 0$
1)

Danach erhält man D_{11} mit der horizontalen Parität P_1 usw.

Danach erhält man D_{11} mit der horizontalen Parität P_1 , dann D_{32} usw.

D_1	D_2	D_3	D_4	D_5	P	$DP+S$	S	
1	0	1	1	0	1	0	1	2b) $P = 1$
0	1	0	0	0	0	0		
1	1	0	0	0	0	1		1)
0	1	0	1	1	1	0		
D_1	D_2	D_3	D_4	D_5	P	$DP+S$	S	
1	0	1	1	0	1	0	1	2)
0	1	0	0	0	0	0		3a) $DP+S=0$
1	1	0	0	0	0	1		1)
0	1	0	1	1	1	0		

Und so weiter

- Wiederherstellung:

Die Wiederherstellung hängt noch in der Luft, weil wir die Kenntnis der Diagonalparität S vorausgesetzt hatten. Sie läßt sich aber leicht aus

$$S = \left(\bigoplus_{l=0}^{m-2} a_{l,m} \right) \oplus \left(\bigoplus_{l=0}^{m-2} a_{l,m+1} \right)$$

berechnen. Die Teilsummen werden aus den Elementen der ersten bzw. zweiten Parität gebildet. Nach Umsortieren sieht man, daß alle Datenelemente doppelt und die Diagonalelemente S $(m-2)$ -mal enthalten sind. Da m eine Primzahl ist, ist $m-2$ ungerade, d.h. die Summe ist gleich S . Die Annahme, daß wir S kennen, war keine Einschränkung.

Im Allgemeinen unterscheiden wir bei zwei ausgefallenen Platten folgende Fälle

$i = m$ und $j = m+1$:

$i < m$ und $j = m$:

$i < m$ und $j = m+1$:

$i < m$ und $j < m$: Dies ist der wichtigste Fall. Hierzu gehört auch unser Beispiel.

Die Ausführung sind unter [36] nachzulesen.

- EVENODD – Vergleich mit RAID 6

Links: Anzahl an XOR Operationen zu Kodierung von $m-1$ Bytes pro Platte bei m Datenplatten:

# of information disks	EVENODD	Reed-Solomon	improvement factor
5	312	376	1.21
7	664	954	1.44
11	1752	3250	1.86
13	2488	5112	2.05
17	4344	10624	2.45
23	8088	24442	3.02
29	12948	46648	3.59
31	14872	56250	3.78
41	26232	124000	4.73
43	28888	142002	4.92

# of information disks	EVENODD vs. Parity	RS vs. Parity
5	2.43	2.93
7	2.30	3.31
11	2.19	4.06
13	2.15	4.43
17	2.12	5.18
23	2.08	6.30
29	2.07	7.43
31	2.06	7.80
41	2.05	9.68
43	2.04	10.06

[36]

Rechts: Anzahl an XOR Operationen vergleichen mit RAID-5.

EVENODD benötigt etwa doppelt so viel Operationen wie RAID-5, kann aber dafür doppelt so viel Platten wiederherstellen wie RAID-5. EVENODD ist dahingehend optimal.

- 2-dimensionales Paritäts-RAID

Ein orthogonales Paritäts-RAID stellt sicher, dass jeder Datenblock durch zwei unterschiedliche Paritäts-Blöcke abgesichert wird. Voraussetzung ist:

- Es dürfen nie zwei Datenblöcke gemeinsam in zwei Paritäts-Blöcke einfließen
- Auch nach dem Ausfall von zwei Festplatten kann jedes Datum rekonstruiert werden

Aufbau:

- Bilde Paritäten sowohl über Zeilen als auch Spalten des Feldes
- Benötigt dreimal Lesen und dreimal Schreiben pro Schreibzugriff
- Overhead:

$m \times n$ – Feld erfordert $(m + n - 1)$ zusätzliche Paritäts-Felder

- 2-dimensionales Paritäts-RAID

Es gibt zwei Anordnungen, wobei die zweite noch eine zusätzliche „Superparitätsplatte“ S hat. Man sieht leicht, daß der gleichzeitige Ausfall von zwei Platten verkraftet werden kann.

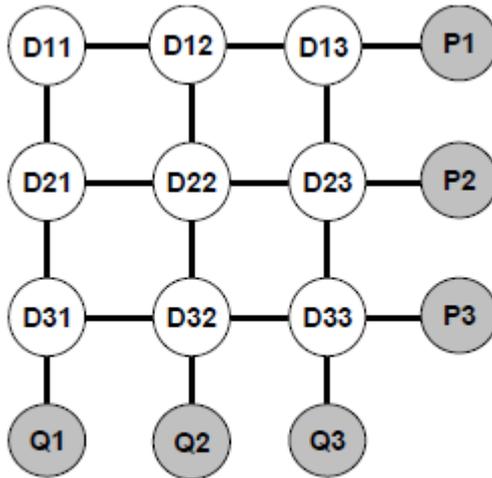


Fig. 1. A two-dimensional RAID array with 9 data and 6 parity disks.

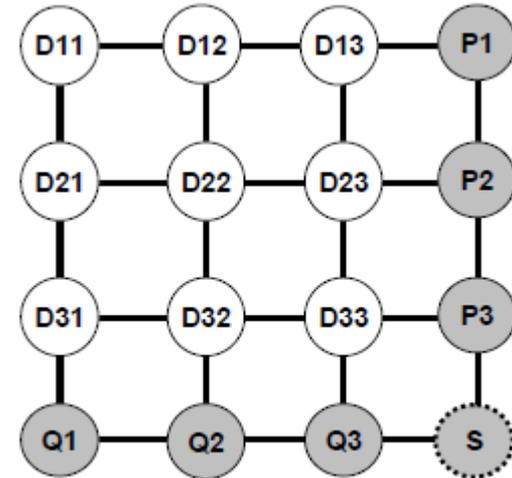


Fig. 2. The same array with a superparity disk (S).

$$S = P1 \oplus P2 \oplus P3 = Q1 \oplus Q2 \oplus Q3$$

[38]

- 2-dimensionales Paritäts-RAID

Der Ausfall von drei Platten kann bei dem einfachen 2-dimensionalen Paritäts-RAID zum Datenverlust führen. Bei der erweiterten Version fängt die Superparitätsplatte diese Fälle ab.

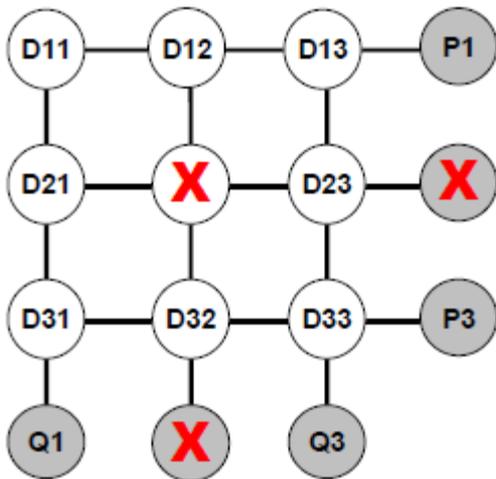


Fig. 3. A triple failure resulting in a data loss.

[38]

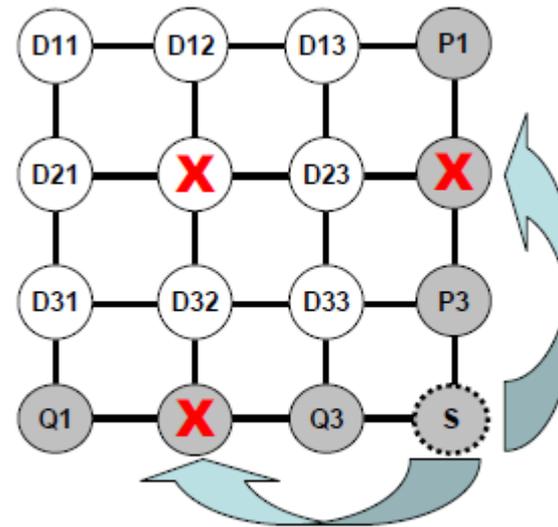
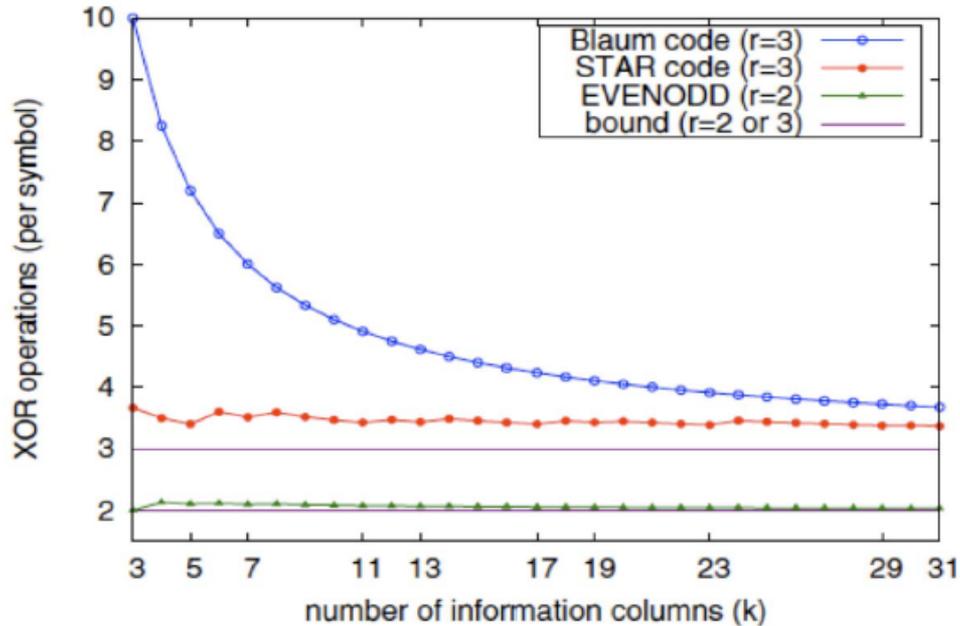


Fig. 7. Recovering from a previously fatal triple failure.

Auf der nächsten Folie werden verschiedene Verfahren hinsichtlich der der Aufwand an XOR-Operationen verglichen.

- 2-dimensionales Paritäts-RAID



# of total columns (n)	# of XORs (= rL)	
	r = 2	r = 3
$n \leq 8$	6	9
$9 < n \leq 16$	8	12
$17 < n \leq 32$	10	15
$33 < n \leq 64$	12	18

Table 1: Complexity of the RS Code (per 32 bits)

Comparison between Star, Evenodd, an additional codes from Blaum, and a purely XOR-basierten RS-Code from Blömer et al.

[39]

- Verlässlichkeit
 - Performance -> RAID 10 (1+0)
 - Hohe Kapazitäten und besten Schutz gegen Auszeiten und Datenverlust -> RAID 6
- Datenverlust beim Rebuild (Beispiel von Intel):

Bei einem Array mit 60 Laufwerken unter normalen Betriebsbedingungen und gleichen Fehlerwahrscheinlichkeiten muß mit Datenverlust gerechnet werden beim

 - RAID 5: nach 6 Monaten durch zweifachen Totalausfall oder einem Totalausfall und einem zusätzlichen Bad Block Failure
 - RAID 6: nach 25 Jahren



- RAID 7
 - Hersteller Storage Computer
 - RAID-Controller mit Echtzeitbetriebssystem und großem Puffer
 - Die Laufwerke werden vom Datenbus abgekoppelt.
 - Mehrere Paritätsinformationen werden verwendet

- RAID 5E
 - IBM
 - RAID 5 Enhanced
 - Hot-Spare nicht als separate Platte; Hot-Spare wird ans Ende aller Festplatten aufgeteilt

- RAID 5EE
 - IBM
 - RAID 5 Enhanced Extended
 - Hot-Spare nicht als separate Platte; Hot-Spare wird diagonal über alle Festplatten verteilt

• Fehlerbetrachtungen:

- Beim Rebuild müssen alle Sektoren fehlerfrei bearbeitet werden
- Probleme sind durch defekte Platten, defekte Plattensektoren, Einzelbitfehler möglich

MTTDL ist ein Maßstab für Reliability, Availability und Serviceability.

- Für nichtgeschützte Daten (Dynamisches Striping, RAID-0):

$$\text{MTTDL} = \text{MTTF} / N$$

- Für Systeme mit einfache Parität (RAID-1, RAID-5, 2-fach Spiegel)

$$\text{MTTDL} = \text{MTTF}(\text{disk1}) * \text{MTTF}(\text{disk2}) / (N * (N-1) * \text{MTTR}(\text{disk1}))$$

- Für Systeme mit zweifacher Parität (RAID-6, RAID-DP, RAID-Z2)

$$\text{MTTDL} = \text{MTTF}(\text{disk1}) * \text{MTTF}(\text{disk2}) * \text{MTTF}(\text{disk3}) / (N * (N-1) * (N-2) * \text{MTTR}^2(\text{disk1}))$$

Dabei wird zur Herleitung die Unabhängigkeit der Ausfälle angenommen. Mögliche Auswirkungen von Plattenausfällen auf andere Platten können durch verringerte MTTF berücksichtigt werden.

- Fehlerbetrachtungen bei LSE und N Platten:

$$MTTF_{Erster\ Fehler} = MTTF / N$$

Sei f_{N-1} der Fehler während des Rebuilds bei N-1 Platten. Dann gilt

$$f_{N-1} = 1 - (1 - LSE_{Rate})^{BitsPerDisk(N - 1)}$$

$$\begin{aligned} MTDDL_{RAID5_LSE} &= MTTF_{Erster\ Fehler} / f_{N-1} \\ &= MTTF_{Platte} / (N * f_{N-1}) \end{aligned}$$

Die Summe der beiden Fehlergrößen ergibt für RAID 5

$$MTDDL_{RAID5} = ((MTDDL_{RAID5_DF})^{-1} + (MTDDL_{RAID5_LSE})^{-1})^{-1}$$

Zweifacher Plattenfehler gefolgt von LSE

$$MTTDL = MTTF(disk1) * MTTF(disk2) / (N * (N-1) * MTTR(disk1) * f_{N-2})$$

- Fehlerbetrachtungen bei LSE und N Platten

Table 4. Reliability Parameters

Total User Capacity	100 disks (500 GB)
Disk Size	5 GB
Sector Size	512 bytes
Bit Error Rate (BER)	1 in 10^{14} bits 1 in $2.4 \cdot 10^{10}$ sectors
$p(\text{disk})$ The probability of reading all sectors on a disk (Derived from disk size, sector size, and BER.)	99.96%
Parity Group Size	16 disks
MTTF(disk)	200,000 hours
MTTF(disk2)	20,000 hours
MTTF(disk3)	2,000 hours
MTTR(disk)	1 hour
MTTF(sys)	1 month
MTTR(sys)	1 hour

This table lists parameters used for reliability calculations in this section.

Beispiel [11]

	MTTDL	MTTDL	Probability of Data Loss over 10 Year Period
Double Disk Failure	$\frac{MTTF(disk) \times MTTF(disk2)}{N \times (G - 1) \times MTTR(disk)}$	285 yrs.	3.4%
Sys Crash + Disk Failure	$\frac{MTTF(sys) \times MTTF(disk)}{N \times MTTR(sys)}$	154 yrs.	6.3%
Disk Failure + Bit Error	$\frac{MTTF(disk)}{N \times (1 - (p(disk))^{G-1})}$	36 yrs.	24.4%
Software RAID	(harmonic sum of above)	26 yrs.	31.6%
Hardware RAID (NVRAM)	(harmonic sum excluding sys crash+disk failure)	32 yrs.	26.8%

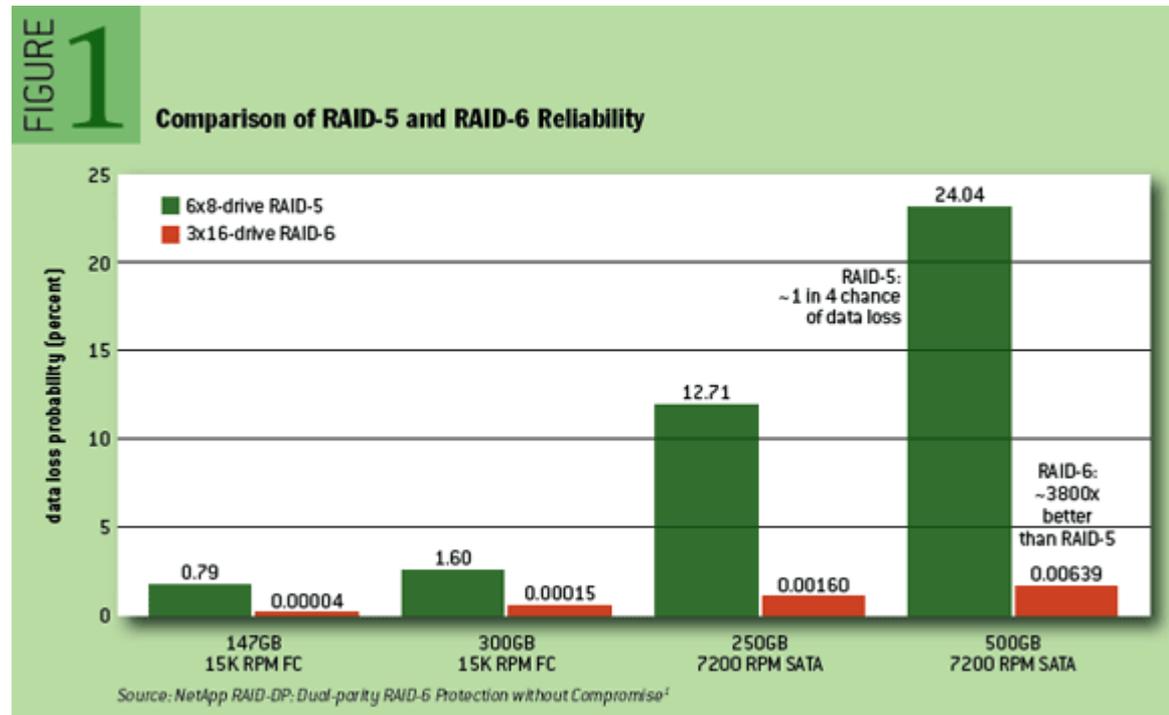
MTTDL is the mean time to data loss. The 10-year failure rate is the percent chance of data loss in a 10-year period. For numeric calculations, the parity group size, G , is equal to 16, and the user data capacity is equal to 100 data disks. Note that the total number of disks in the system, N , is equal to the number of data disks times $G/(G - 1)$.

Wir betrachten nur den Fall $N=G$ (eindimensionales Array).

	MTTDL	MTTDL	Probability of Data Loss over 10 Year Period
Triple Disk Failure	$\frac{MTTF(disk) \times MTTF(disk2) \times MTTF(disk3)}{N \times (G - 1) \times (G - 2) \times MTTR^2(disk)}$	38052 yrs.	0.03%
Sys Crash + Disk Failure	$\frac{MTTF(sys) \times MTTF(disk)}{N \times MTTR(sys)}$	144 yrs.	7.7%
Double Disk Failure + Bit Error	$\frac{MTTF(disk) \times MTTF(disk2)}{N \times (G - 1) \times (1 - (p(disk))^{G-2}) \times MTTR(disk)}$	47697 yrs.	0.02%
Software RAID	(harmonic sum of above)	143 yrs.	6.8%
Hardware RAID (NVRAM)	(harmonic sum excluding sys crash+disk failure)	21166 yrs.	0.05%

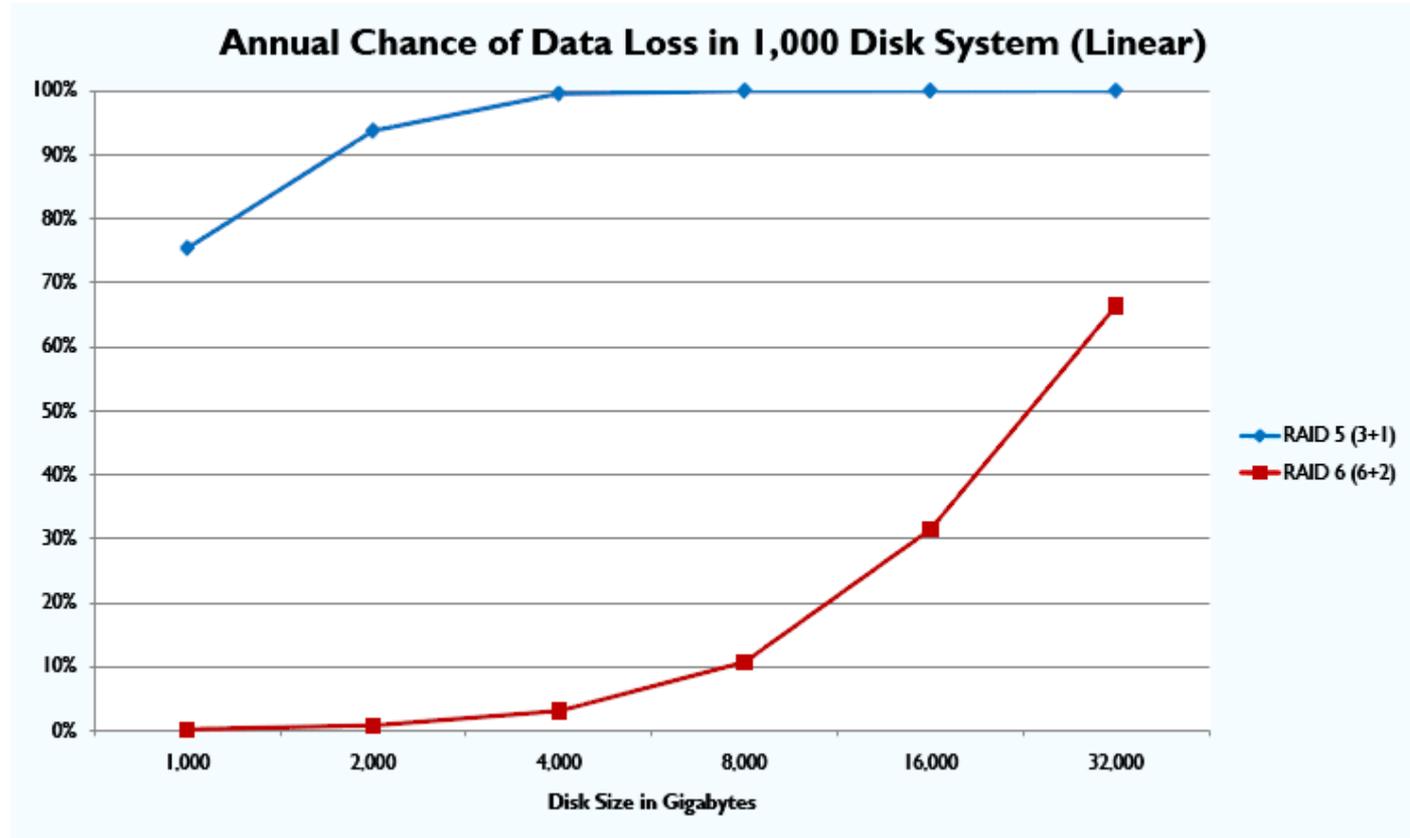
MTTDL is the mean time to data loss. The 10-year failure rate is the percent chance of data loss in a 10-year period. For numeric calculations, the parity group size, G, is equal to 16, and the user data capacity is equal to 100 data disks. Note that the total number of disks in the system, N, is equal to the number of data disks times G/(G - 2).

- Vergleich RAID-5 mit RAID-6

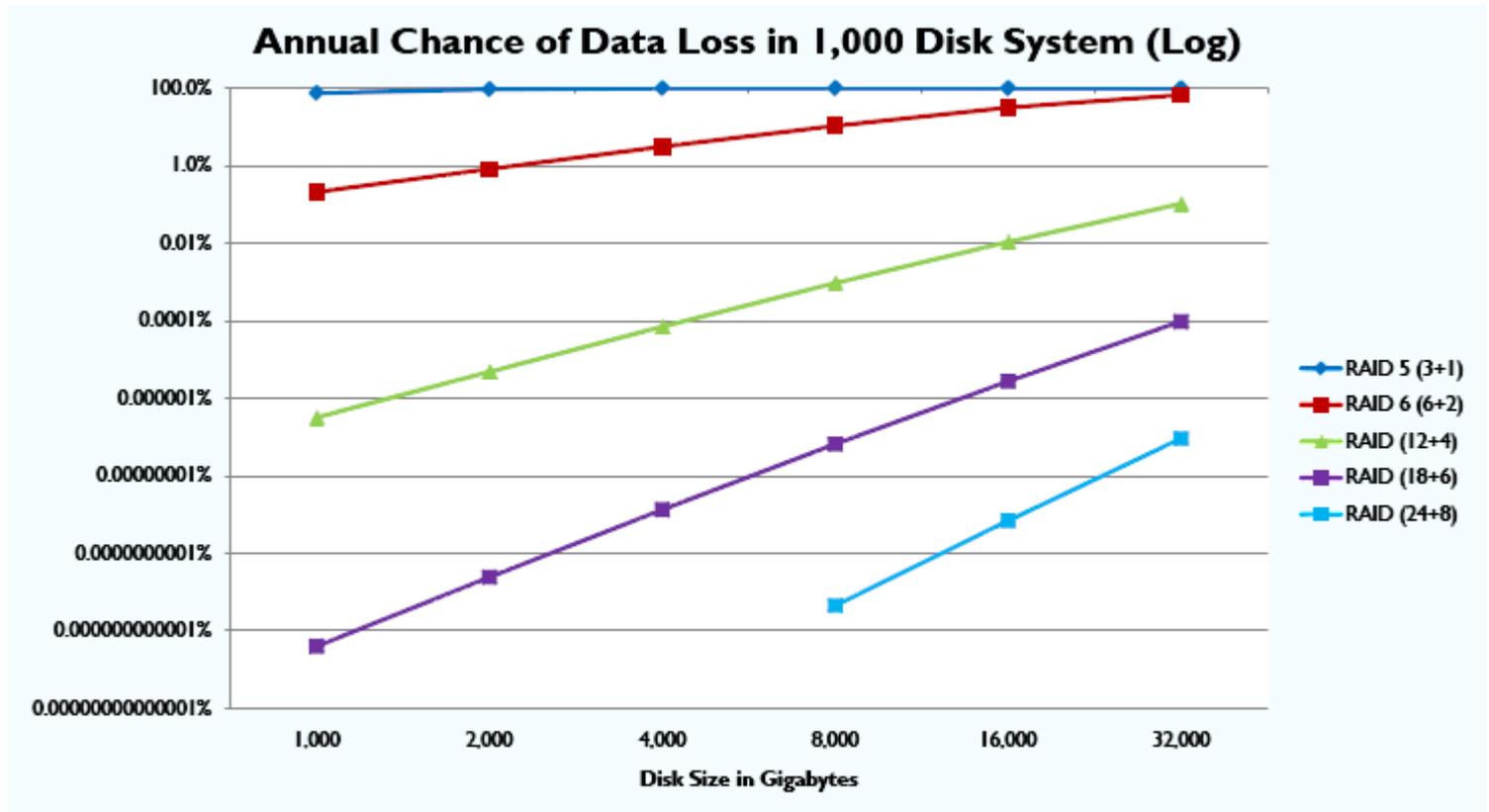


[4]

Grenzen von RAID 5 und RAID 6



[5]

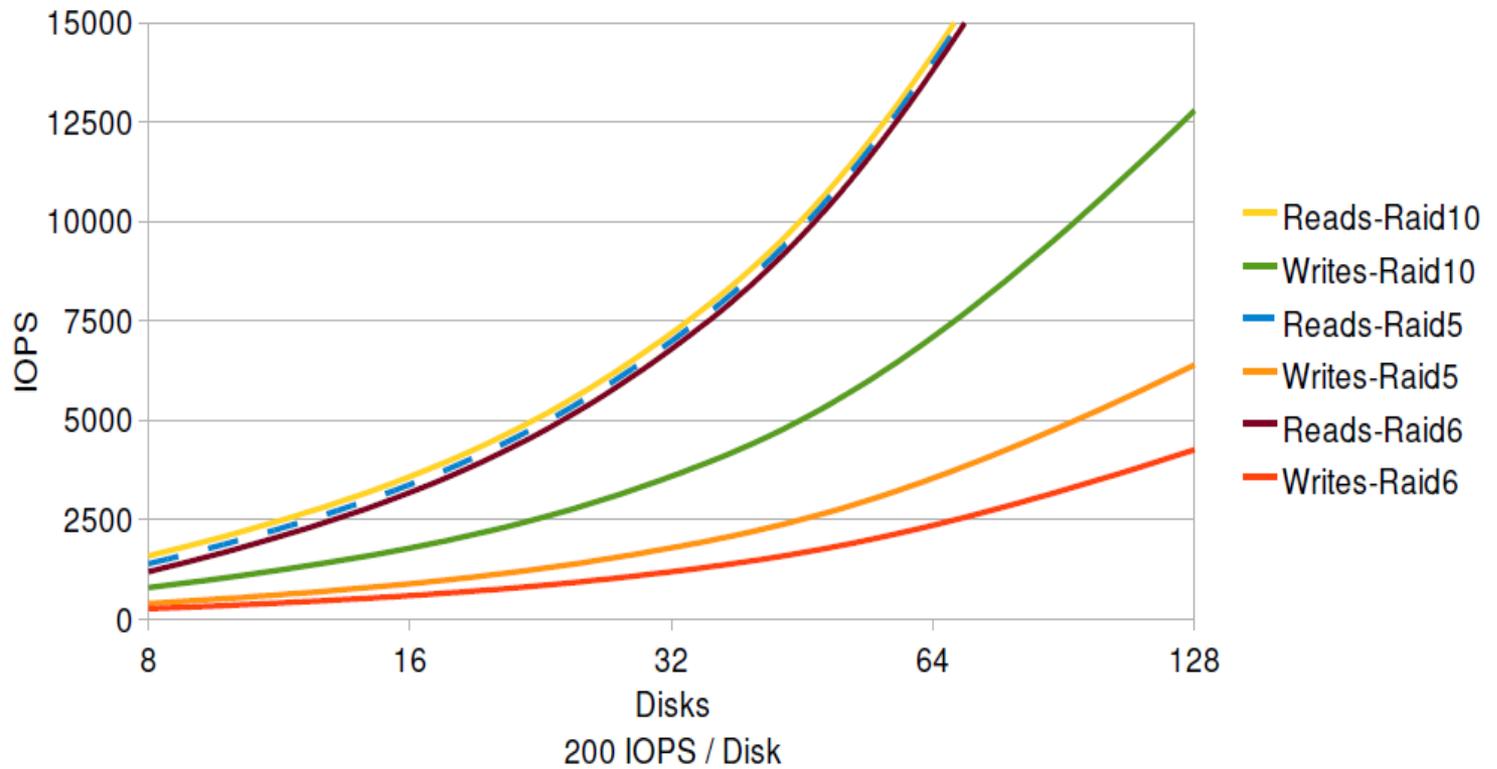


[5]



Vergleich RAID 10, 5 und 6

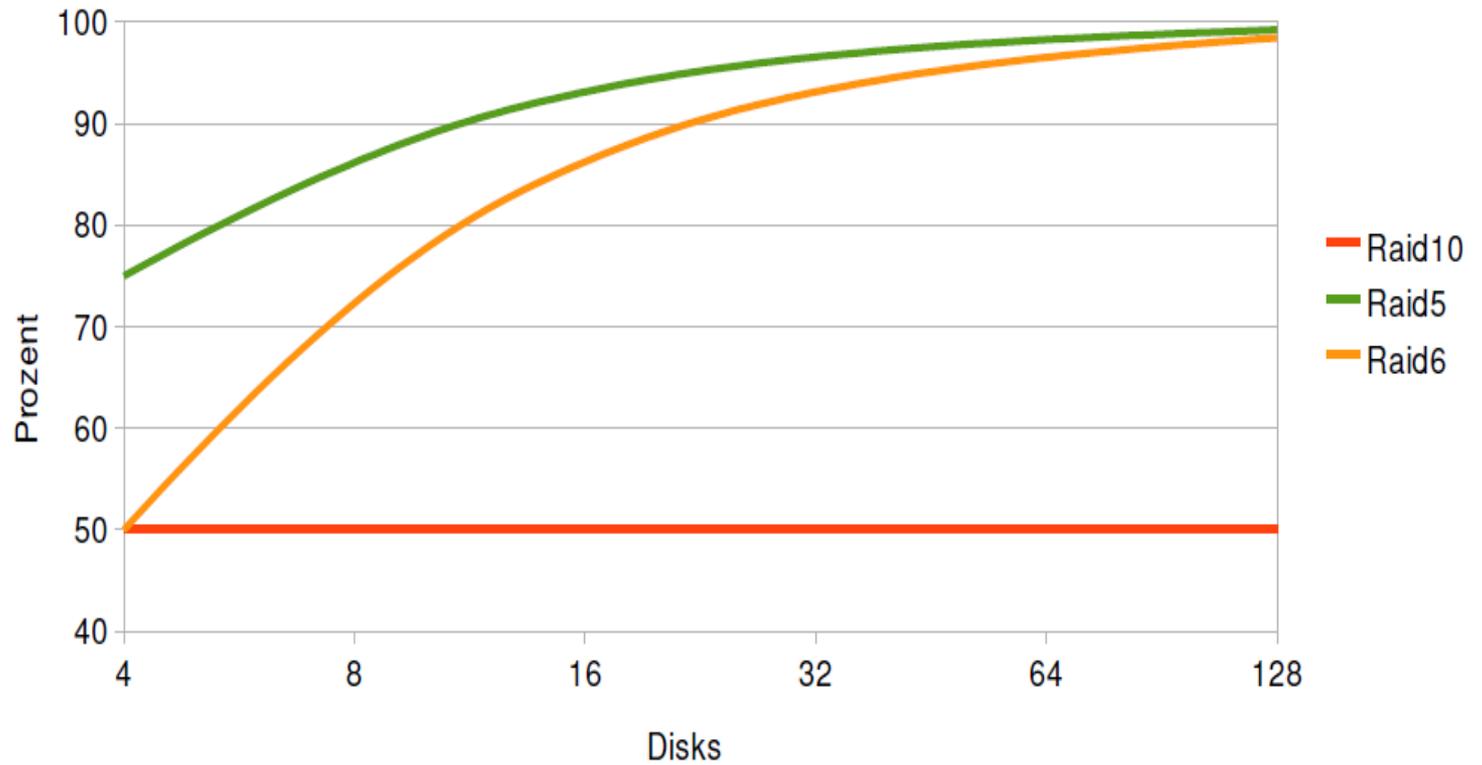
IO Durchsatz
RAID 10 : 5 : 6



[31]



Speichereffizienz Raid 10 : 5 : 6



[31]

Vergleich RAID 5 und 6

Ausfallwahrscheinlichkeit
Mean Time To Data Loss:

► Lebenszeit von RAID-5 und RAID-6

Disks**	4	6	8	10	20	40
RAID-5*	24.841	9.937	5.323	3.312	784	191
RAID-6*	129.381.945	25.876.389	9.241.568	4.312.732	453.972	52.381

* MTDL in Jahren

** MTF 250.000h

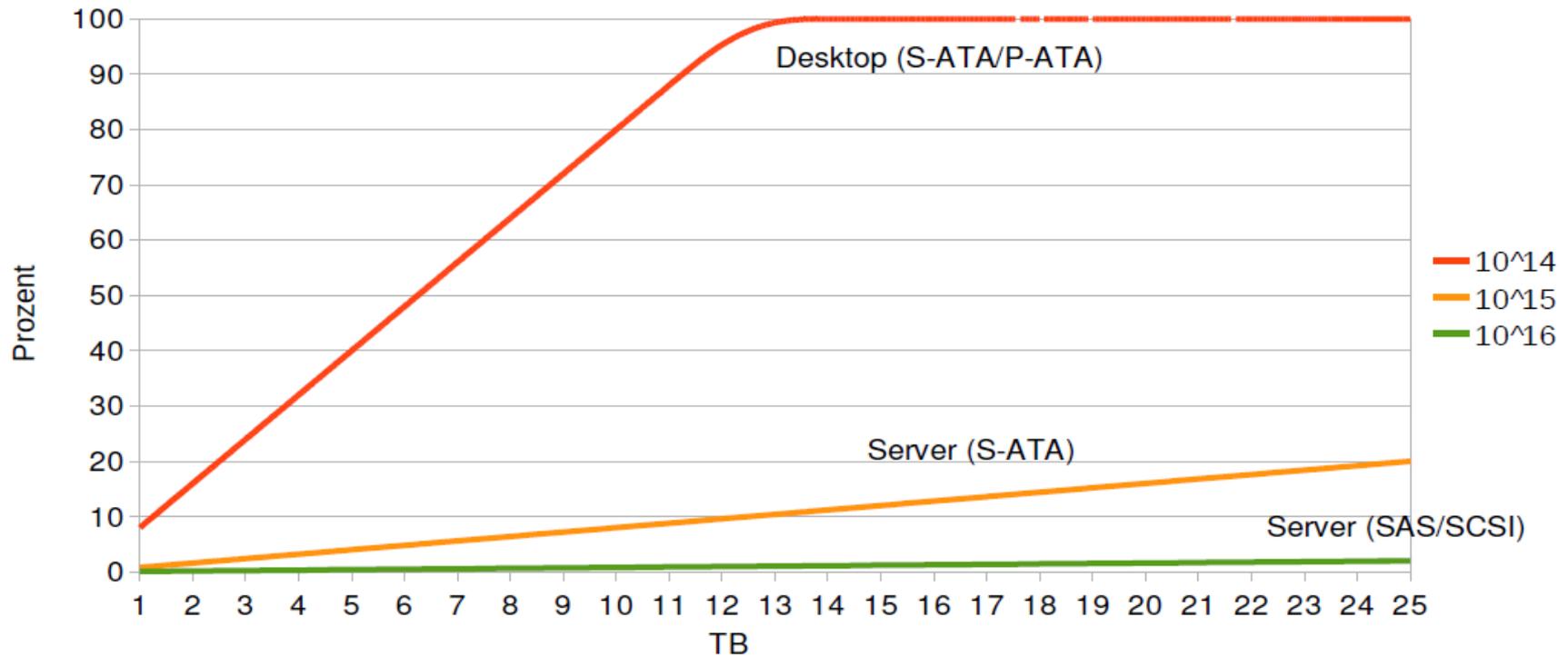
MTTR 24h

[31]

RAID 6

Ausfallwahrscheinlichkeit
Bitfehlerwahrscheinlichkeit:

Abhängig zur Gesamtkapazität



[31]

Ausfallwahrscheinlichkeit Mean Time To Data Loss mit Bit Error Rate:

► Lebenszeit von RAID-5 und RAID-6 mit BER

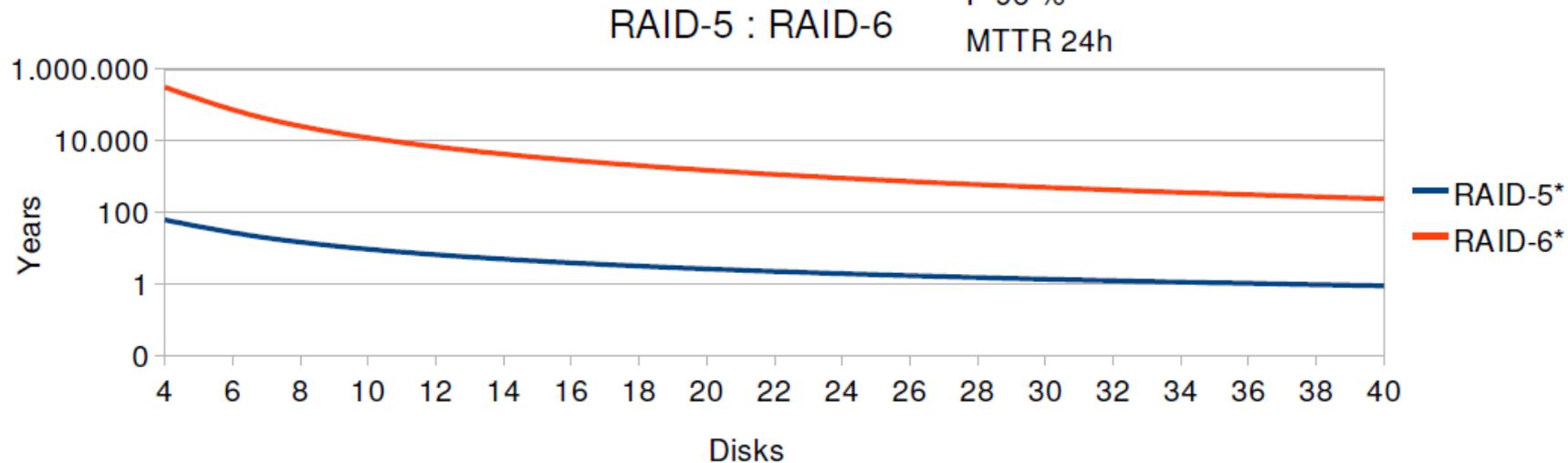
Disks**	4	6	8	10	20	40
RAID-5*	62	26	14	9	3	1
RAID-6*	316.854	65.956	24.503	11.888	1.507	242

* MTTDL in Jahren

** MTTF 250.000h

* P 96 %

MTTR 24h



[31]

- Zusammenfassung
- Anzahl der Festplatten
 - Die Anzahl der Festplatten n gibt an, wie viele Festplatten benötigt werden, um das jeweilige RAID aufzubauen.
- Nettokapazität
 - Die Nettokapazität k gibt die nutzbare Kapazität in Abhängigkeit von der Anzahl der verwendeten Festplatten n an. Dies entspricht der Anzahl der benötigten Festplatten ohne RAID, die die gleiche Speicherkapazität aufweisen.
- Ausfallsicherheit
 - Die Ausfallsicherheit s gibt an, wieviele Festplatten ohne Datenverlust ausfallen dürfen.
- Kombinations-RAIDs: Leg
 - Ein *Leg* (englisch für Bein) oder *lower level RAID* ist ein RAID-Array, welches mit anderen gleichartigen Legs über ein übergeordnetes RAID-Array (*upper level RAID*) zusammengefasst wird. Hierbei ist n in Leg die Anzahl der Festplatten in einem Leg und n of Leg die Anzahl der Legs im übergeordnetem Array.

Übersicht über die Standard-RAIDs

RAID	n	k	Sicherheit	Lesen	Schreiben
0	≥ 2	n	0	++	++
1	≥ 2	1	n-1	+	=
2	≥ 3	=1	2		
3	≥ 3 (2)	n-1	1		
4	≥ 3 (2)	n-1	1		
5	≥ 3	n-1	1	+	=
6	≥ 4	n-2	2	=	=
DP	≥ 3	n-2	2		

0	0 fail	Striping, Verschränkung
1	N-1 fail	Spiegelung, Mirroring
2	0 fail	Fehlererkennung auf HD selbst, keine Redundanz
3	1 fail	byte level striping with parity
4	1 fail	block level striping with parity
5	1 fail	block level striping with rotating parity
10	2 -> half fail	striped mirror, spiegeln und verschränken => Leistung Raid 1 x 2
6	2 fail	block level striping with rotating dual parity

- Nutzung
 - R0: Geschwindigkeit
 - R1: hohe Sicherheit, großer Platzverlust
 - R5: R+, W~, mit Caches und als 5+0 optimal
 - R1+0: sicher und groß, 50% Verlust, für Softwareraid

HW-RAID Anwendungsperformance

- Online-Transaktionssysteme (OLTP)

Konfiguration	I/O pro Lesezugriff	I/O pro Schreibzugriff	I/O pro 100 OLTP-Zugriffe*	I/O-Effizienz*
RAID-0	1	1	100	100 %
RAID-1	1	2	130	77 %
RAID-0+1	1	2	130	77 %
RAID-5	1	4	190	52 %
RAID-6	1	7	280	35 %

* basierend auf einem typischen OLTP-Mix aus 30 % Schreib- und 70 % Lese-Operationen

3. Hardware \leftrightarrow SW-RAID

- Software-Raid
 - Windows Server
 - Linux (LVM)
 - meist Raid 0,1 und 10
 - Direkter Laufwerkszugriff noch möglich
- Hardware-Raid
 - Systemunabhängiges Raid
 - Nur SCSI/FC-Raid ermöglicht komplette Plattenabstraktion

3. HW – RAID

- Plattenschnittstellen
 - FC 2 FC
 - FC 2 SCSI
 - FC 2 SATA
 - SCSI 2 SCSI
 - SCSI 2 SATA/PATA

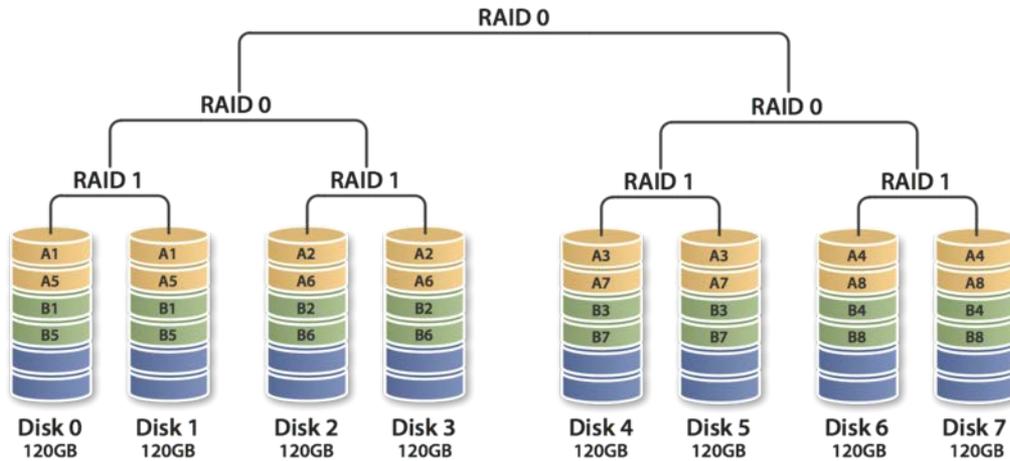
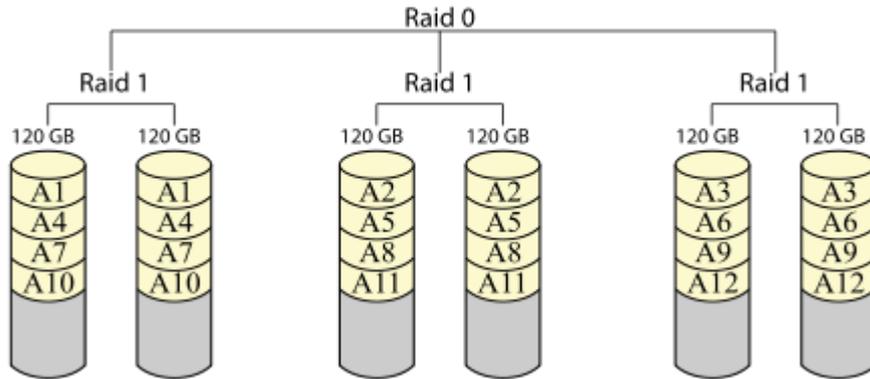
3. HW-RAID

- RAID sind kein Ersatz für eine Datensicherung
 - Backup und Backup-Strategien sind deshalb eine wesentliche Aufgabe des Systemadministrators

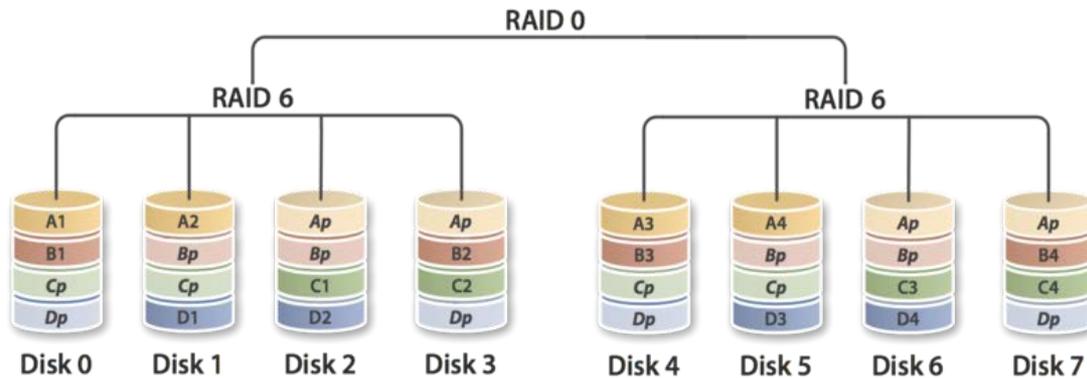
3.1 Verbundsysteme

- Alle RAID-Levels lassen sich zur Leistungsverbesserung kombinieren.
- RAID 1+0 oder 0+1 oder
- RAID 5+0 oder 5+1
- Beliebig viel Stufen sind möglich

3.1 Verbundsysteme



3.1 Verbundsysteme



n - Top Level Division

m - Bottom Level Division

Level	Description	Minimum # of disks	Space Efficiency	Fault Tolerance	Read Benefit	Write Benefit
RAID 0+1	Top Level RAID 1, Bottom Level RAID 0.	4	1/n	n - 1 to m(n - 1)	(n · m)X	mX
RAID 1+0	Top Level RAID 0, Bottom Level RAID 1.	4	1/m	m - 1 to n(m - 1)	(n · m)X	nX
RAID 5+0	Top Level RAID 0, Bottom Level RAID 5.	6	1 - 1/m	1 to n	n(m - 1)X*	n(m - 1)X*
RAID 5+1	Top Level RAID 1, Bottom Level RAID 5.	6	(1 - 1/m) / n	2n - 1 to m(n - 1) + 1	n(m - 1)X*	(m - 1)X*
RAID 6+0	Top Level RAID 0, Bottom Level RAID 6.	8	1 - 2/m	2 to 2n	n(m - 2)X*	n(m - 2)X*
RAID 6+1	Top Level RAID 1, Bottom Level RAID 6.	8	(1 - 2/m) / n	3n - 1 to m(n - 1) + 2	n(m - 2)X*	(m - 2)X*

* - Assumes hardware is fast enough to support

3.2 Jenseits von RAID

- SSPIRAL (Survivable Storage using Parity in Redundant Layouts)
 - Wachsendes Datenvolumen erhöht das Risiko, durch Ausfall mehrerer Platten Daten zu verlieren. Redundante Verfahren wie RAID 6 u. a. verwenden aufwendige Verfahren für die Berechnung mehrfacher Paritäten, um die Datenintegrität auch in solchen Fällen zu garantieren. Die Wiederherstellung ist zeitaufwendig. Im Normalbetrieb sind Schreibvorgänge verhältnismäßig langsam wegen der Neuberechnung der Paritäten.
 - Das SSPIRAL ist ein Schema redundanter Daten, das mit einfacher XOR-Paritätsberechnung arbeitet und gleichzeitig hohe Zuverlässigkeit und leichte Wartbarkeit garantiert.
 - Die Ausfallsicherheit wird stärker, der Nettokapazitätsverlust schwächer gewertet.
 - S. <http://www2.cs.uh.edu/~paris/MYPAPERS/Snapi07.pdf>

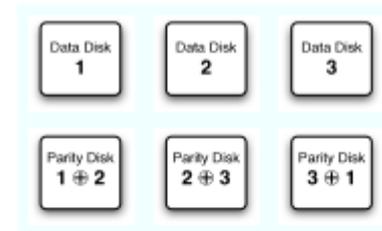
- SSPIRAL [24]

- Drei Parameter zur Beschreibung
 - Systemgrad (degree of the system)
 - X-Anordnung (x-order)
 - Gesamtknotenanzahl

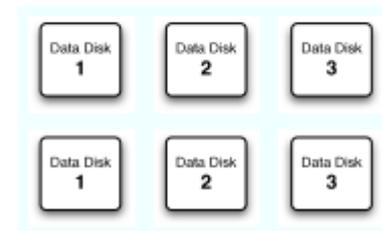
- Bsp. Rechts oben:

- Grad 3
- X-Anordnung 2 – zwei Knoten (hier Platten) werden zur Paritätsberechnung verwendet
- 6 Knoten

- Zum Vergleich RAID 1+0 mit gleichem Datenvolumen



(a) Pairwise-Parity (3+3 SSPIRAL)



(b) 3 pairs of mirrored disks

- SSPiRAL

- Auch bei Ausfall von zwei beliebigen Platten kein Datenverlust anders als bei RAID 1+0, wo bei gleichzeitigem Ausfall von Daten- und zugehöriger Spiegelplatte das RAID-System ausfällt; bessere MTDL als Raid 1+0
- Mehr Datenplatten sind möglich

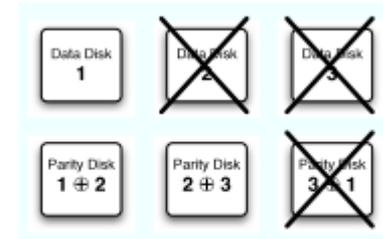


Figure 2: SSPiRAL data layout and the loss of three nodes.

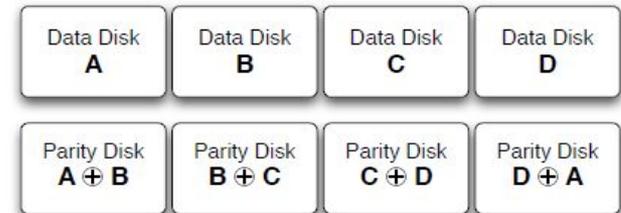


Figure 4. The 4+4 SSPiRAL layout (with $x = 2$).

- SSPiRAL

- Größere Sicherheit durch Paritätsbildung über mehrere Platten.

Bei dem Beispiel können im „optimalen“ Fall bis zu vier, auf jeden Fall bis zu drei Datenplatten ohne Datenverlust ausfallen:

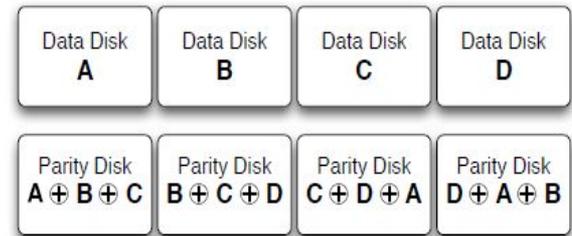


Figure 7. The 4+4 SSPiRAL layout with eight disks and $x = 3$.

Ich kürze XOR als Punkt „ \oplus “ ab, den ich auch noch weglasse. Dann gilt

$$\begin{aligned}
 A &= (ABC) (CDA) (DAB), \\
 B &= (DAB) (ABC) (BCD), \\
 C &= (ABC) (BCD) (CDA), \\
 D &= (BCD) (CDA) (DAB).
 \end{aligned}$$

Deshalb tritt kein Datenverlust ein. Bei Ausfall von vier Platten kann Datenverlust auftreten z.B. beim Ausfall von A, ABC, CDA und DAB kann A nicht mehr hergestellt werden.

3.3 Tape-RAID

- Tape-RAID
 - Auch beim Backup lassen sich die gleichen Überlegungen wie bei den Platten anwenden, um die Performance und Verlässlichkeit zu erhöhen.

3.4 Platten für RAID

- Auswahl der Platten für RAID-Einsatz
 - SATA-Platten
 - SAS-Platten
 - FC-Platten

4. Zusammenfassung

- Leistungsfähigkeit hängt von drei Parametern ab:
 - Anzahl möglicher I/O-Operation pro Zeit bei zufälligem Zugriff in IOPS (SATA-Platten Zugriffszeit 8 bis 10 ms, halbe Umdrehung 2 bis 4 ms. Rechnet man 10 ms bei nicht sequentiell Zugriff, so ergeben sich 100 IOPS. SCSI- und SAS-Platten liegen bei 200 IOPS.)
 - Datendurchsatz bei sequentiell Zugriff in MB/s abhängig von der Drehzahl, Schreibdichte und Anzahl beschriebener Oberflächen (Bei 10.000 Umdrehungen pro Minute etwa doppelt so schnell wie bei einer IDE-Platte mit 5.400 Upm. Samsung Spinpoint VL40 liefert etwa 33 MB/s; die WD Raptor WD740GD schafft 62 MB/s.)
 - Streifentiefe beim RAID
Bei kleinen Streifentiefe (*Streifengröße*, auch *stripe size*, *chunk size* oder *interlace size*; s. *Data Disc Format Seite 37*) werden bei gleichen Daten über mehr Platten angesprochen als bei großen. Je nachdem, ob es um Lesen oder Schreiben geht, sind die Ergebnisse unterschiedlich.

4. Zusammenfassung

– RAID 0 (aus zwei Platten):

Kleiner Streifentiefe: Beim Lesen sind praktisch beide Platten beschäftigt, d.h. die zweite Platte steht für einen weiteren IO-Befehl nicht zur Verfügung. Der Durchsatz verdoppelt sich, aber die IOPS bleiben bei 100 IOPS.

Großer Streifentiefe: Bei gleichen Daten ist es wahrscheinlicher, daß nur von einer Platte gelesen wird. Die zweite steht für den nächsten IO-Befehl bereit. Der Datendurchsatz bleibt, aber der IOPS-Werte verdoppeln sich. Server profitieren von dieser Anordnung. Bei zwei Platten mit 100 IOPS und 60 MB/s sind 200 IOPS bei 60 MB/s im RAID 0 erreichbar.

– RAID 1 (aus zwei Platten):

Verhält sich beim Schreiben wie die einzelnen Platten. Die Streifentiefe ist nicht von Belang.

Beim Lesen verhält es sich wie RAID 0, d.h. die Streifentiefe wirkt sich aus.

Bei zwei Platten mit 100 IOPS sind lesend bis 200 IOPS bei etwa 60 MB/s, schreibend bis 100 IOPS bei 30 MB/s im RAID 0 erreichbar.

– RAID 10:

Verdoppelt etwa die Leistung von RAID 1. Bei 8 Platten vom Typ 100 IOPS können theoretisch lesend bis 800 IOPS bei etwa 240 MB/s erreicht werden, während schreibend bis 400 IOPS bei etwa 120 MB/s möglich sind.

Verhält sich beim Schreiben wie die einzelnen Platten. Die Streifentiefe ist nicht von Belang.

werden die gleichen Daten aber mehr Platten angesprochen als bei RAID 0. Je nachdem ob es um Lesen oder Schreiben geht, sind die Ergebnisse unterschiedlich.

4. Zusammenfassung

– RAID 5 (aus mindestens drei Platten):

Bei drei Platten vom Typ 100 IOPS sind lesend bis 200 IOPS möglich. (Die Paritätsplatte wird nicht mitgerechnet.) Beim Schreiben ergibt sich eine größere Bandbreite abhängig von der Datenmenge, Streifentiefe und Plattenanzahl. Im *Idealfall* werden alle Daten eines Streifens und die Parität neugeschrieben. Dann sind 200 IOPS möglich. Im *Normalfall* sind pro Schreibvorgang immer zwei lesende (alter Sektor und alte Parität) und zwei schreibende (neuer Sektor und neue Parität) Zugriffe erforderlich. Im Mittel sind 50 bis 75 IOPS möglich.

Die Schreibleistung hängt stark von der Lage der Daten auf der Platte und der Streifengröße ab. Zur Erzielung guter Schreibleistung ist die Streifentiefe klein zu wählen, um dem Idealfall nahe zu kommen. Ist die Streifentiefe groß und werden überwiegend kleine Datenmenge geschrieben, so kann die Leistung deutlich geringer als beim RAID 1 sein.

– RAID 6 (aus mindestens vier Platten):

Beim RAID 6 sind die Zusammenhänge ähnlich komplex wie beim RAID 5. Die Nettokapazität ist durch die zweite Paritätsplatte geringer.

5. Weitere Begriffe

- Cache

Die Leistung des RAID-Systems hängt vom Cache ab. Wir unterscheiden

- Betriebssystem
- RAID-Controller
- Plattenarray

- RAID 5-Schreibloch

Kann ein Schreibvorgang einschließlich Paritätsspeicherung nicht abgeschlossen werden, so entstehen Inkonsistenzen, die nicht gleich auffallen müssen. Kann z.B. die neue Parität nicht mehr gespeichert werden, wenn vorher der Strom ausfällt und das System keine USV oder die Platten auch keine Batterie haben. Die nicht alte Parität bleibt dann erhalten. Wenn jetzt eine Datenplatte ausfällt, können die Daten an dieser Stelle nicht rekonstruiert werden. Werden die Datenblöcke mit fehlerhaftem Paritätsblock im Betrieb wieder überschrieben und kann der Schreibvorgang einschließlich Paritätsblock fehlerlos beendet werden, so wirkt sich der zwischenzeitliche Fehler nicht aus. Auf jeden Fall darf solch eine Situation für ein Transaktionssystem nicht eintreten.

Die Situation kann beim RAID 5 auftreten. Sie wird als RAID 5-Schreibloch bezeichnet.

5. Weitere Begriffe

- RAID 5-Schreibloch (Forts.)

Die Situation kann beim RAID 5 auftreten. Sie wird als RAID 5-Schreibloch bezeichnet. Vermieden kann sie durch Cachekontroller mit Batterie oder Platten mit Batterie oder einer Unterbrechungsfreien Stromversorgung (USV) oder durch den Übergang zu RAID-Systemen, die diese Situation schon bei der Entwicklung berücksichtigten (RAID-Z, RAID-Z2, RAID-Z3; s. [13]-[15]).



Literatur

1. Jon Elerath, Hard Disk Drives: The Good, the Bad and the Ugly!
<http://queue.acm.org/detail.cfm?id=1317403>
2. <http://de.wikipedia.org/wiki/RAID>
3. 15 Years Of Hard Drive History: Capacities Outran Performance
<http://www.tomshardware.com/reviews/15-years-of-hard-drive-history,1368.html>
4. Adam Leventhal, Triple-Parity RAID and Beyond, Dez 17, 2009
<http://queue.acm.org/detail.cfm?id=1670144>
5. Jason Resch, Solving Data Loss in Massive Storage Systems, 2010,
http://www.snia.org/sites/default/files2/sdc_archives/2010_presentations/tuesday/JasonResch_%20Solving-Data-Loss.pdf
6. MTBF (Mean Time Between Flareups, er, Failures) <http://www.faqs.org/faqs/arch-storage/part2/section-151.html>
7. Hard Disk Drive Reference Guide
http://www.storagereview.com/hard_disk_drive_reference_guide
8. Estimating Drive Reliability in Desktop Computers Consumer Electronics Systems, March 27, 2011, <http://ixbtlabs.com/articles/storagereliability/>
9. Jeff Whitehead, Calculating Mean Time To Data Loss (and probability of silent data corruption) <http://info.zetta.net/2009/06/>



10. David A. Patterson, Garth Gibson, and Randy H. Katz: A Case for Redundant Arrays of Inexpensive Disks (RAID), in *International Conference on Management of Data (SIGMOD)*, Seiten 109-116, Juni 1988; s. a.
<http://www.eecs.berkeley.edu/Pubs/TechRpts/1987/CSD-87-391.pdf>
11. P. M. Chen, E. K. Lee u.a.: RAID: High-Performance, Reliable Secondary Store, in *ACM Computing Surveys*, vol. 26, Nor. 2, June 1994, Seiten 145-185; s. a.
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.17.7866&rep=rep1&type=pdf>
12. Common Raid Disk Data Format (DDF), siehe
http://www.snia.org/tech_activities/standards/curr_standards/ddf
13. Jeff Bonwick's Blog: RAID-Z, 2005; http://blogs.oracle.com/bonwick/entry/raid_z
14. Adam Leventhal's Weblog: Double-Parity RAID-Z, Jun 18, 2006;
http://blogs.oracle.com/ahl/entry/double_parity_raid_z
15. Adam Leventhal's Weblog: The need for triple-parity RAID, Dec 21, 2009;
http://blogs.oracle.com/ahl/entry/acm_triple_parity_raid
16. J.E. Angus, On computing MTBF for a k-out-of-n:G repairable system. *IEEE Trans. Reliab.*, 37 3 (1988), pp. 312–313.

17. By Seagate Technology: Estimating Drive Reliability in Desktop Computers Consumer Electronics Systems, March 27, 2001; <http://ixbtlabs.com/articles/storagereliability/>
18. http://de.wikibooks.org/wiki/Computerhardware:_HDD:_Ausfall#Elektrostatik
19. Marcus Schuster: Neue RAID-Level im Überblick: Doppelt gesichert, 10.07.2009; <http://www.admin-magazin.de/content/download/1087/10167/file/Neue-RAID-Level-im-Ueberblick.pdf>
20. David S. H. Rosenthal: Keeping Bits Safe: How Hard Can It Be?, October 1, 2010; <http://queue.acm.org/detail.cfm?id=1866298>
21. Lakshmi N. Bairavasundaram, Garth R. Goodson, Shankar Pasupathy, Jiri Schindler: An Analysis of Latent Sector Errors in Disk Drives. In *Proceedings of the 2007 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, USA, 2007. s. http://delivery.acm.org/10.1145/1260000/1254917/p289-bairavasundaram.pdf?ip=160.45.113.47&acc=ACTIVE%20SERVICE&CFID=52330834&CFTOKEN=73408946&acm=1317746959_1300859595b9e735061ab3eb67663540
22. Robin Harris, Latent sector errors in disk drives, 2008, <http://storagemojo.com/2008/02/18/latent-sector-errors-in-disk-drives/>



23. E. Pinheiro, W.-D. Weber and L. A. Barroso: Failure Trends in a Large Disk Drive Population. In Proc. Of the FAST '07 Conference of File and Storage Technologies, 2007. s.
http://www.usenix.org/events/fast07/tech/full_papers/pinheiro/pinheiro_html/
24. Ahmed Amer, Darrell D. E. Long, Jehan-François Pâris and Thomas Schwarz, "Increased Reliability with SSPIRAL Data Layouts," Proceedings of the 16th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2008), Baltimore, MD, USA: IEEE September 2008.
24. Ahmed Amer u. a. Outshining Mirrors: MTTDL of Fixed-Order SSPIRAL Layouts. s.
<http://www2.cs.uh.edu/~paris/MYPAPERS/Snapi07.pdf>
25. Dr. V. Tiederle, Bewertung der Zuverlässigkeit von komplexen Baugruppen s.
http://p58105.typo3server.info/Archiv/2003/Vortrag_Dr_Tiederle_Teil1.pdf
26. Wendy Torell, Victor Avelar, Mittlerer Ausfallzeitraum: Erläuterung und Normen, APC White Paper / technische Dokumentation Nr. 78, 2004,
http://www.apcmedia.com/salestools/VAVR-5WGTSB_R0_DE.pdf

27. Wendy Torell, Victor Avelar, Effektive MTBF-Vergleiche in Datencenter-Infrastrukturen, APC White Paper / technische Dokumentation Nr. 112, , <http://pdfsuche.info/view/aHR0cDovL3d3dy5tZ2V1cHMuZGUvX3doaXRlcGFwZXJzL2RvY3MvMTEyJTlwLSUyMEVmZmVrdGI2ZSUyME1UQkYtVmVyZ2xlaWN0ZSUyMGluJTlwRGFOZW5jZW50ZXIIMjBJbmZyYXN0cnVrdHVyZW4ucGR>
28. Hubert Kirrmann, Fault Tolerant Computing in Industrial Automation, 2. Auflage 2005, http://lamspeople.epfl.ch/kirrmann/Pubs/FT_Tutorial_HK_050418.pdf
29. Michael Glaß, Visualisierung von Zuverlässigkeit, 2006, <http://www12.informatik.uni-erlangen.de/edu/qmz/ss06/docs/visualisierung.pdf>
30. Oliver Ott, RAID – Ein Überblick, 2006, <http://www.gernoth.net/rdf/raid.pdf>
31. Holger Uhlig, Fakten statt Bauchgefühl: RAID-Mathematik für Admins, 2008, http://www.heinlein-support.de/upload/unsere_vortraege/HS_RAID_Mathematik_fuer_Admins.pdf
32. Hannes Federrath, Safety und Fehlertoleranz, <http://www-sec.uni-regensburg.de/intern/lecturenotes/security/20safety.pdf>
33. Robin Harris, Google's Disk Failure Experience, 2007, <http://storagemojo.com/2007/02/19/googles-disk-failure-experience/>
34. Henry Newman, Get Ready for RAID-6, 2008, <http://www.serverwatch.com/tutorials/article.php/3723911/Get-Ready-for-RAID6.htm>

35. Jay White & Chris Lueth, RAID-DP: NetApp Implementation of Double-Parity RAID for Data Protection, NetApp Technical Report , May 2010 | TR-3298
http://www.bedrock-tech.com/wp-content/uploads/2010/05/wp_raid-dp.pdf
36. M. Blaum, J. Brady u.a., EVENODD: an optimal scheme for tolerating double disk failures in RAID architectures, Proceeding ISCA '94 Proceedings of the 21st annual international symposium on Computer architecture, page 245-254
<http://dl.acm.org/citation.cfm?id=192033>
37. Edward K. Lee and Randy H. Katz, Performance consequences of parity placement in disk arrays, 1991, <http://dl.acm.org/citation.cfm?id=106992>
38. J.-F. Paris, T. Schwarz, A. Amer, and D. D. E. Long , [Highly Reliable Two-Dimensional RAID Arrays for Archival Storage](#), *Proceedings of the 31st International Performance of Computers and Communication Conference (IPCCC 2012)*, Austin, TX, Dec. 2012. 107
39. André Brinkmann, Theoretical Aspects of Storage Systems, Autumn 2009, Chapter 2: Double Disk Failures