Freie Universität Berlin

Systemverwaltung (Vorl. und Blockveranstaltung) SS 2012

06.08. bis 24.09.2012

Arnold Kühnel,
Rolf Dietze, Robert Franke

Systemverwaltung

Organisation

- Veranstalter
  A. Kühnel, R. Dietze, R. Franke
- 06.08.-24.09.2012 (2x3h täglich)

1 Einleitung

- 1. Einleitung (Arnold)
  - RAID
- 2. Shells und mehr
  - Korn-Shell (Robert),
  - 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/ArtLehreSystemverwaltung2012

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

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

    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

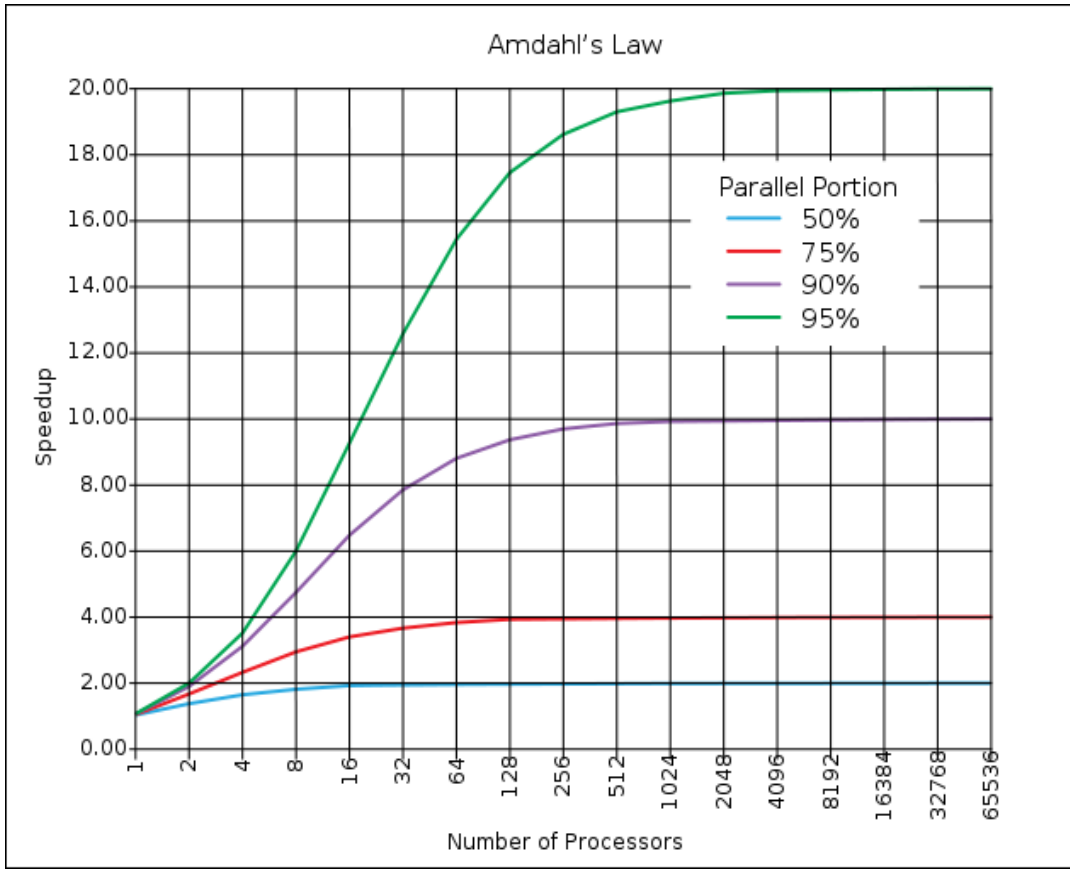
$$S = \frac{1}{(1-p)+pk} \quad Speedup$$

wobei

$p =$ *Anteil parallel ausführbaren Codes,*

$k =$ *# CPUs*

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



- Gordon Moores Gesetz (1965):
  - $\#Transistoren = 2^{Jahr-1964}$

2.1 Entwicklungen

- 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.

2.1 Entwicklungen

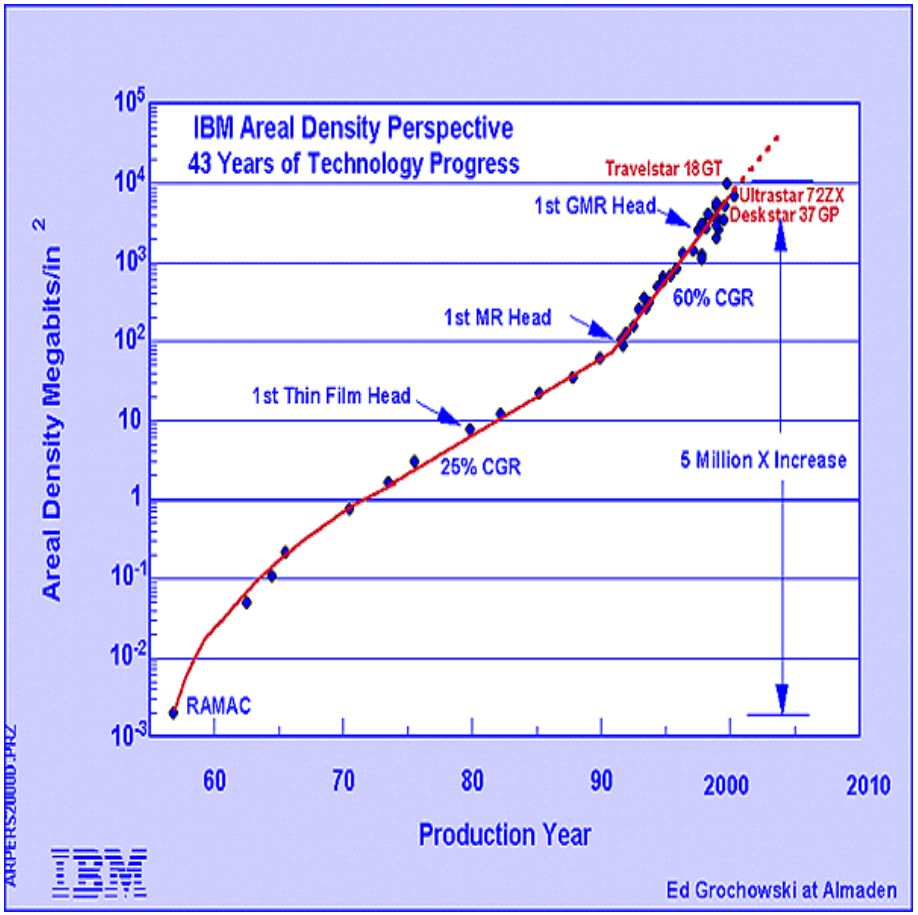
- Folgerungen aus dem Amdahlschen Gesetz
  - Quantitativer Aspekt: Große Verbesserungen im Mikroprozessorbereich verbessern die Gesamtperformanz nur, wenn gleichzeitig der Sekundärspeicher verbessert wird.
  - Qualitativer Aspekt: Schnelle Mikroprozessoren ermöglichen neue Anwendungen in der Bildverarbeitung,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

    Existierende Anwendungen können mehr Daten verwenden.
- Illustrationen zu Kryders Gesetz

Speicherdichte

- Steigerung der Speicherdichte von
- 1991 bis 2006 um das 10.000-fache

- 1956: IBM 305 RAMAC: Erster kommerzieller Rechner mit Festplatten mit beweglichem Kopf

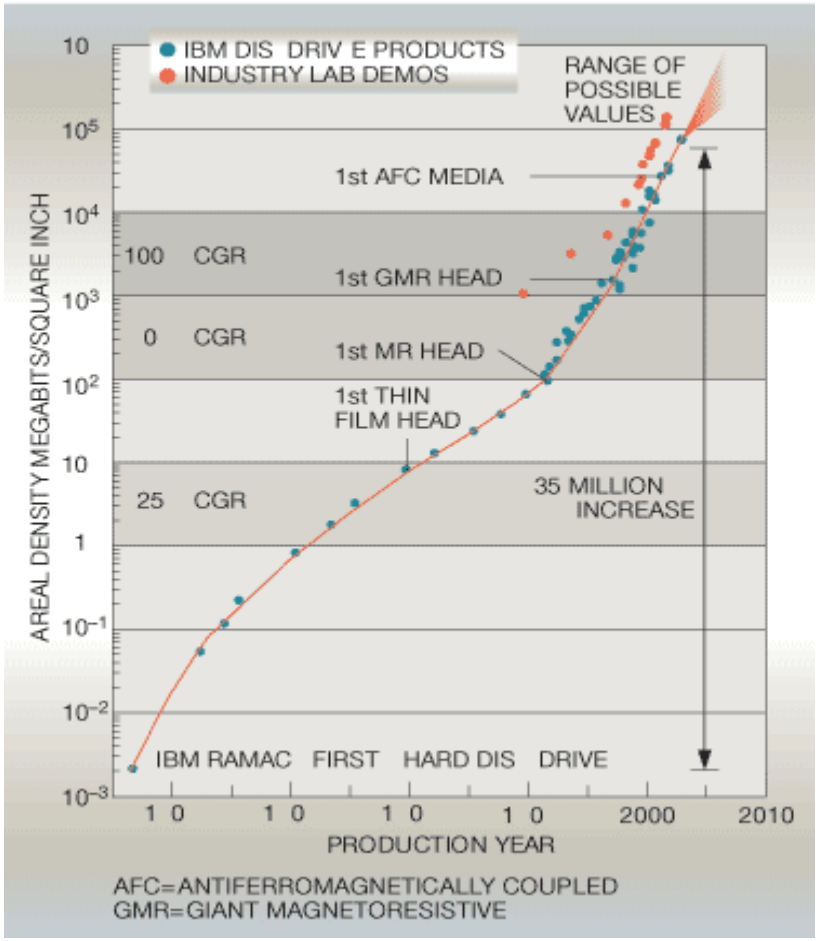


Plattenkapazität

- 1991: 40 MB, 130 MB maximal
- 2006: 750 GB pro HD
- Steigerung auf das 5.700-fache in 15 J.

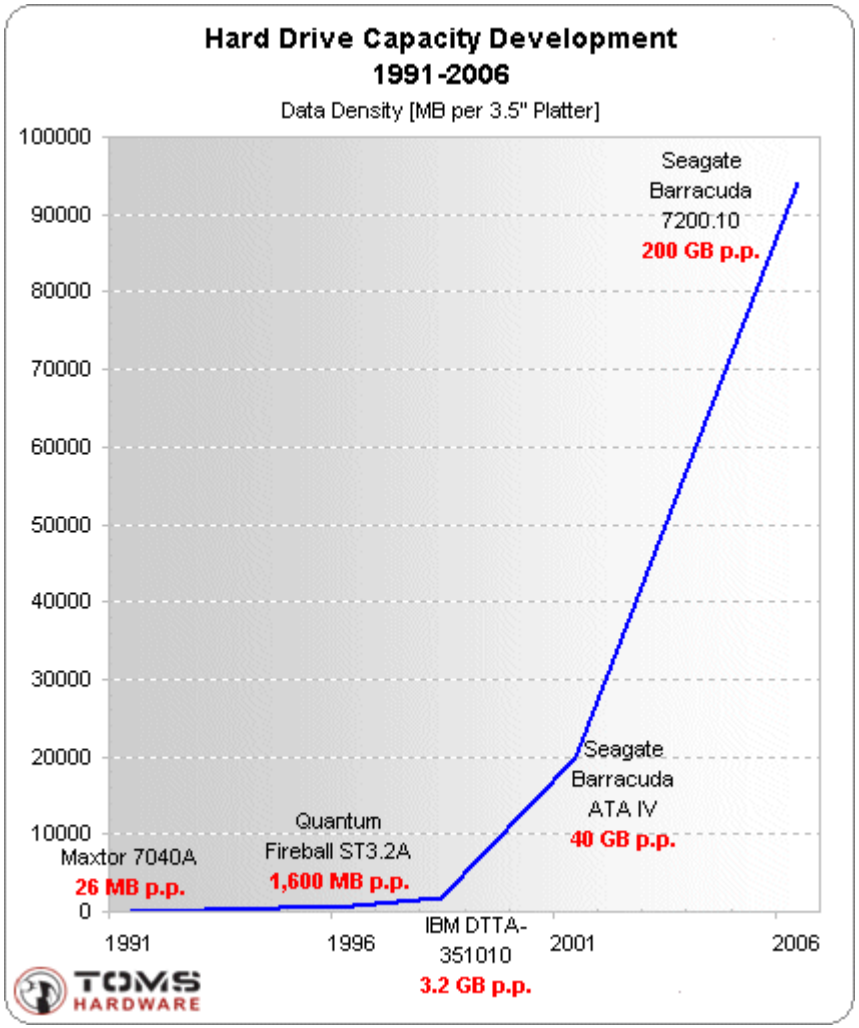
- Magneto-Resistive (MR) Heads
- 2000: Giant Magneto-Resistive (GMR) Heads
- Perpendicular Magnetic Recording
- 2001: Hitachi Antiferromagnetically-coupled (AFC)Media

Figure 1 Hard disk drive areal density trend



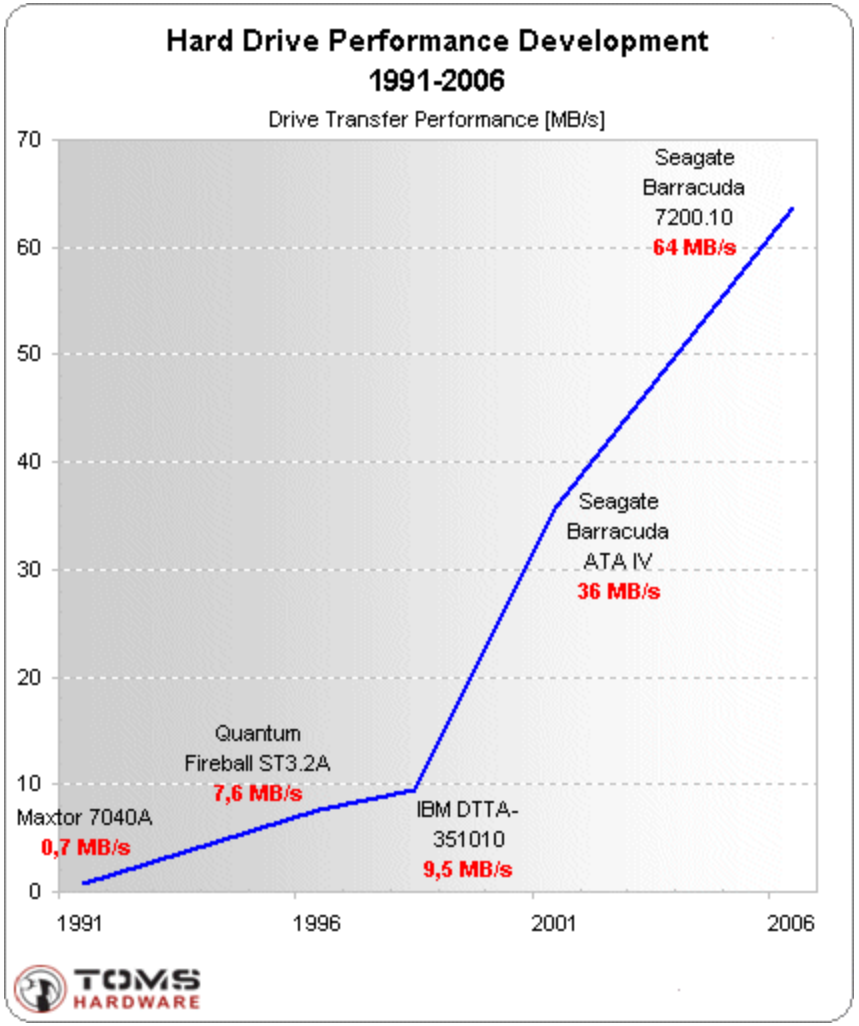
Plattenkapazität

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



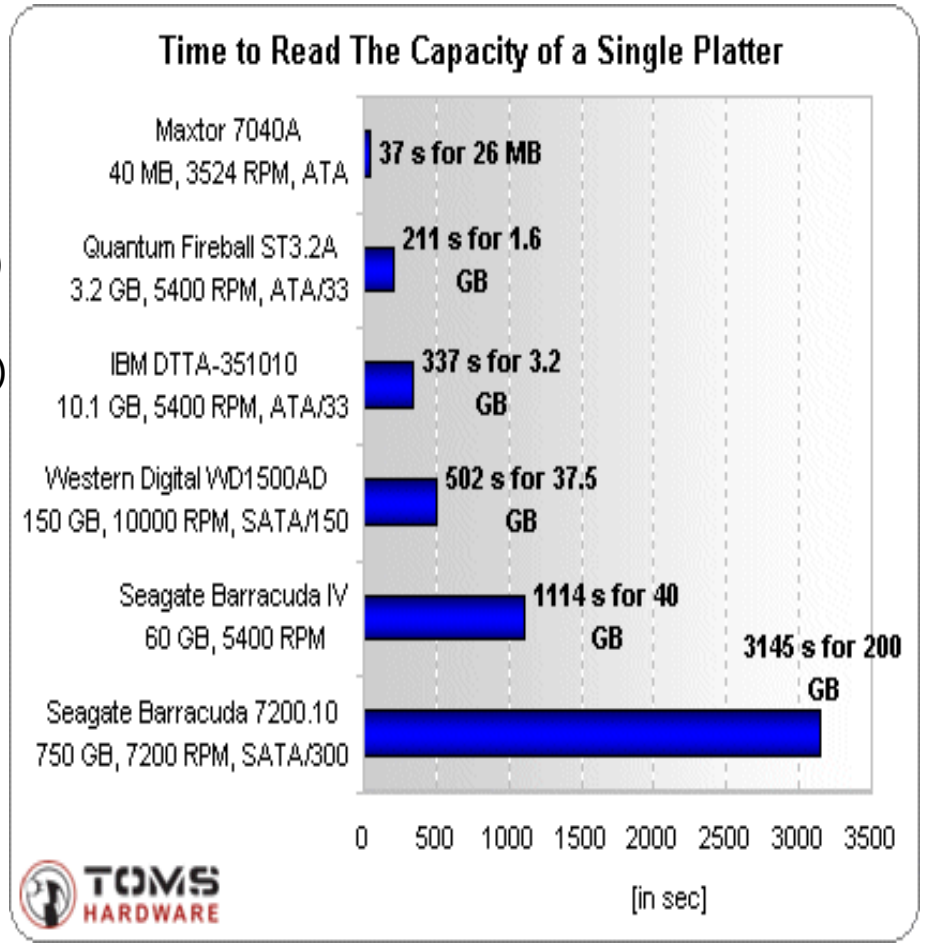
Plattendurchsatz

- 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



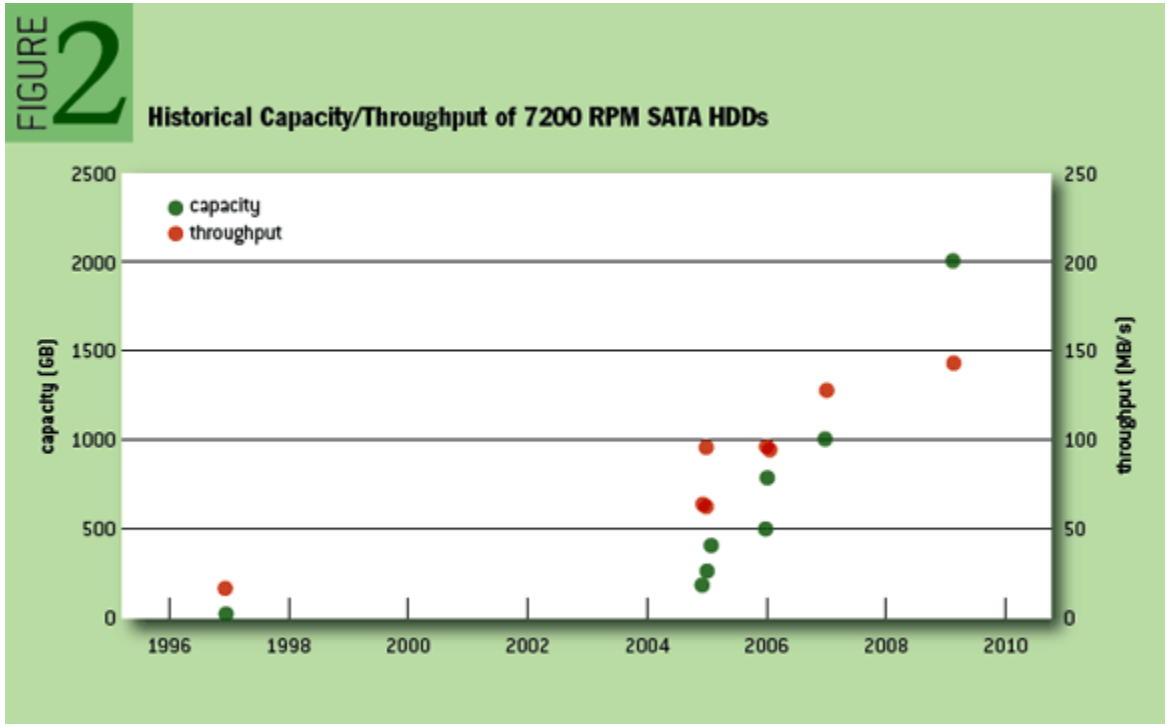
Plattendurchsatz

- 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

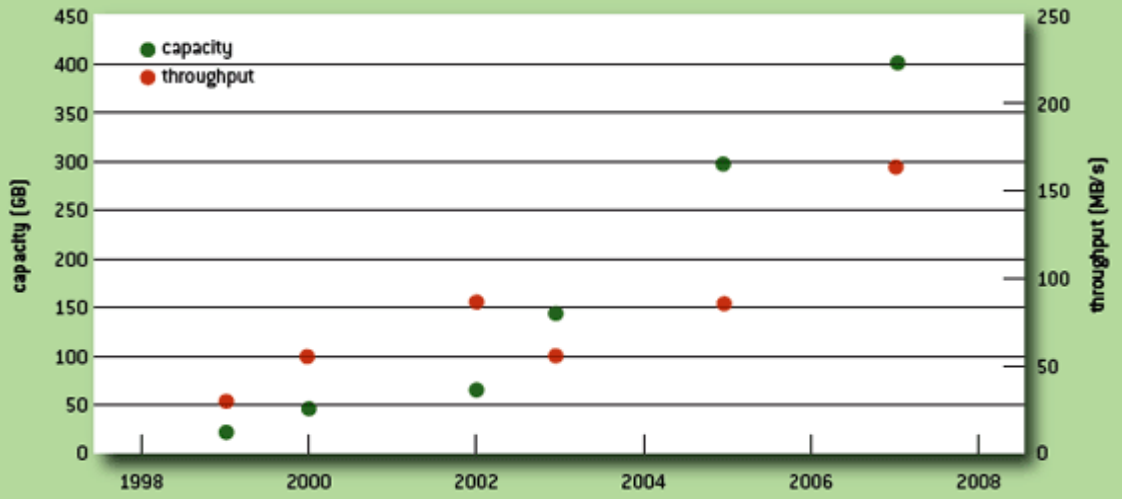
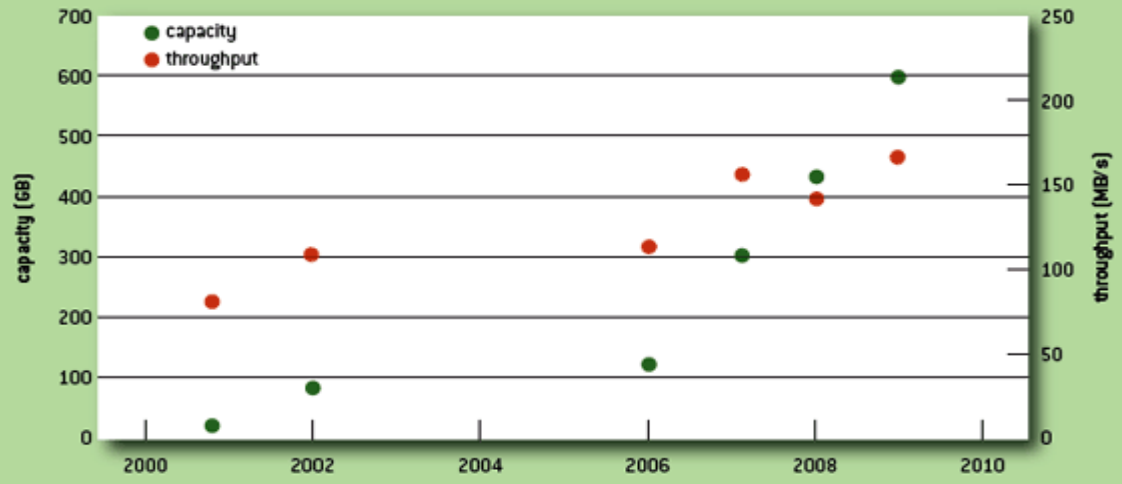


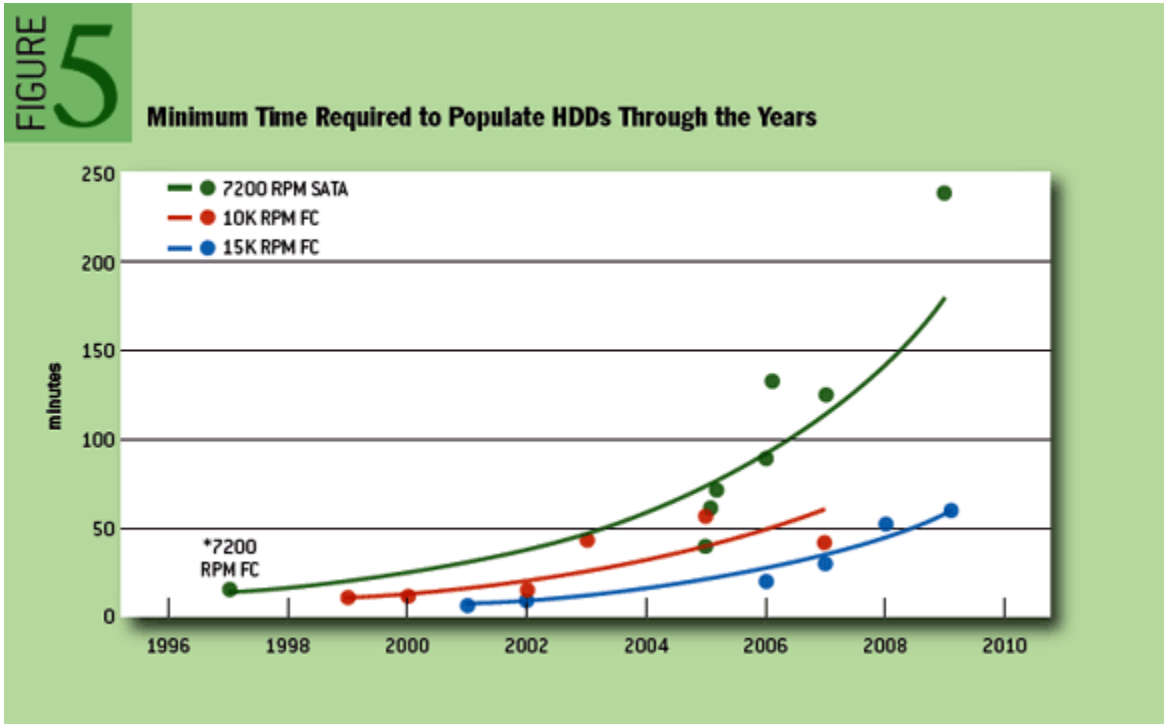
FIGURE 4

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)



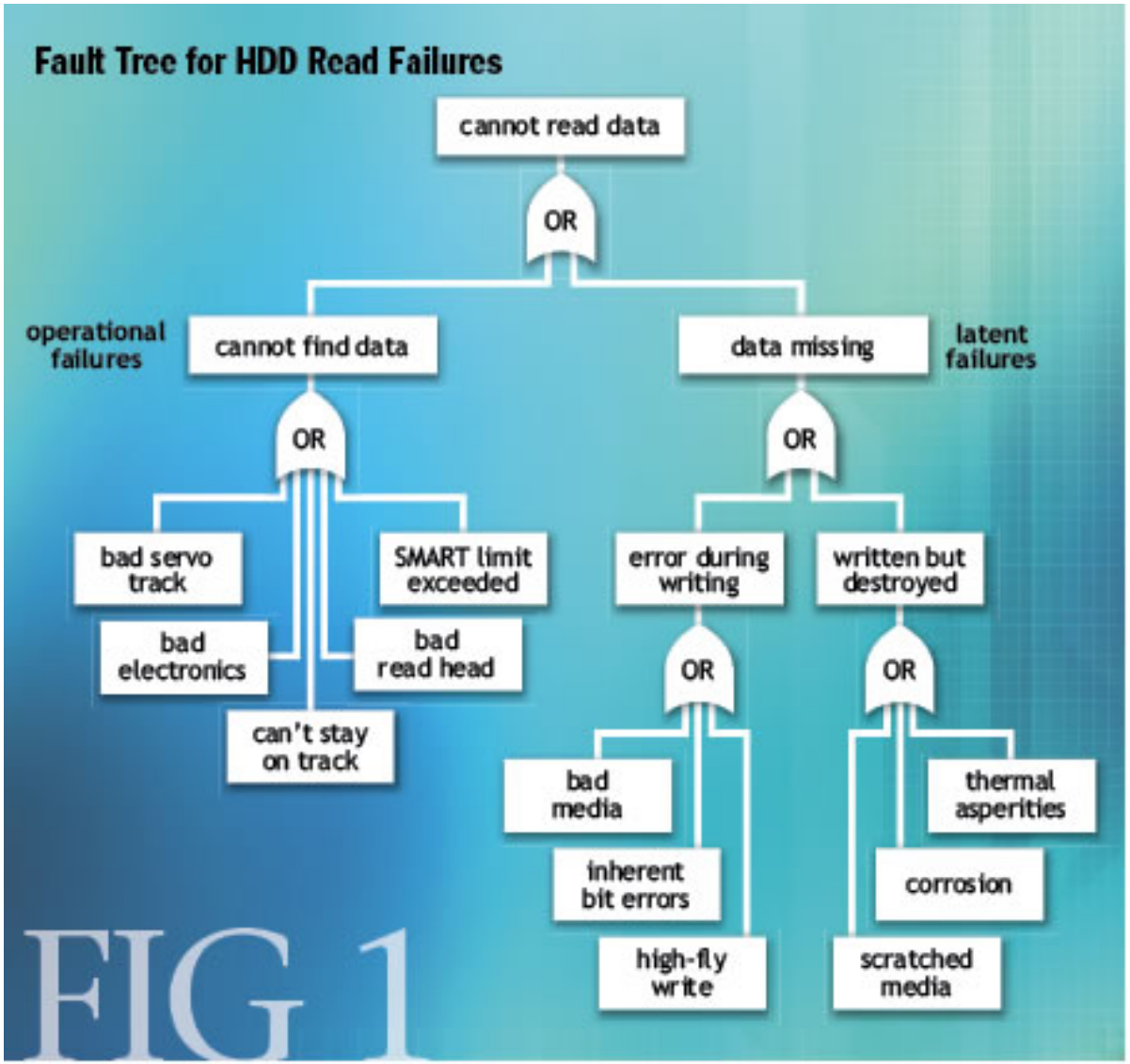
2.1 Entwicklungen

- 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 RAIDs 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)

2.1 Entwicklungen

- 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 Oberflächen-schäden abgespeichert. Die Elektronik der Platte enthält ein Programm zur Auswertung der Liste.
- 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]

2.1 Entwicklungen

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)

2.1 Entwicklungen

- 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.

– Die LSE-Rate wächst mit der Kapazität

2.1 Entwicklungen

- 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 Daten-korrumpierung 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, ….

2.1 Entwicklungen

- 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?

2.1 Entwicklungen

- 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

2.2 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 behebbar sein/das System reparierbar sei oder sie führen zum Totalausfall/das System ist dann nicht reparierbar.

2.2 Qualität technischer Systeme

Bestimmung der Zuverlässigkeit (reliability)

Ausfall (Forts.)

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

Angaben zur MTBF gibt es seit über 60 Jahren. 20 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?

2.2 Qualität technischer Systeme

Bestimmung der Zuverlässigkeit (reliability)

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 eine 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].

2.2 Qualität technischer Systeme

Bestimmung der Zuverlässigkeit (reliability)

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.

2.2 Qualität technischer Systeme

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 (Forts.)

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: 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. Die Änderung der Kurve wird als *Fehler*- oder *Ausfallrate* definiert:

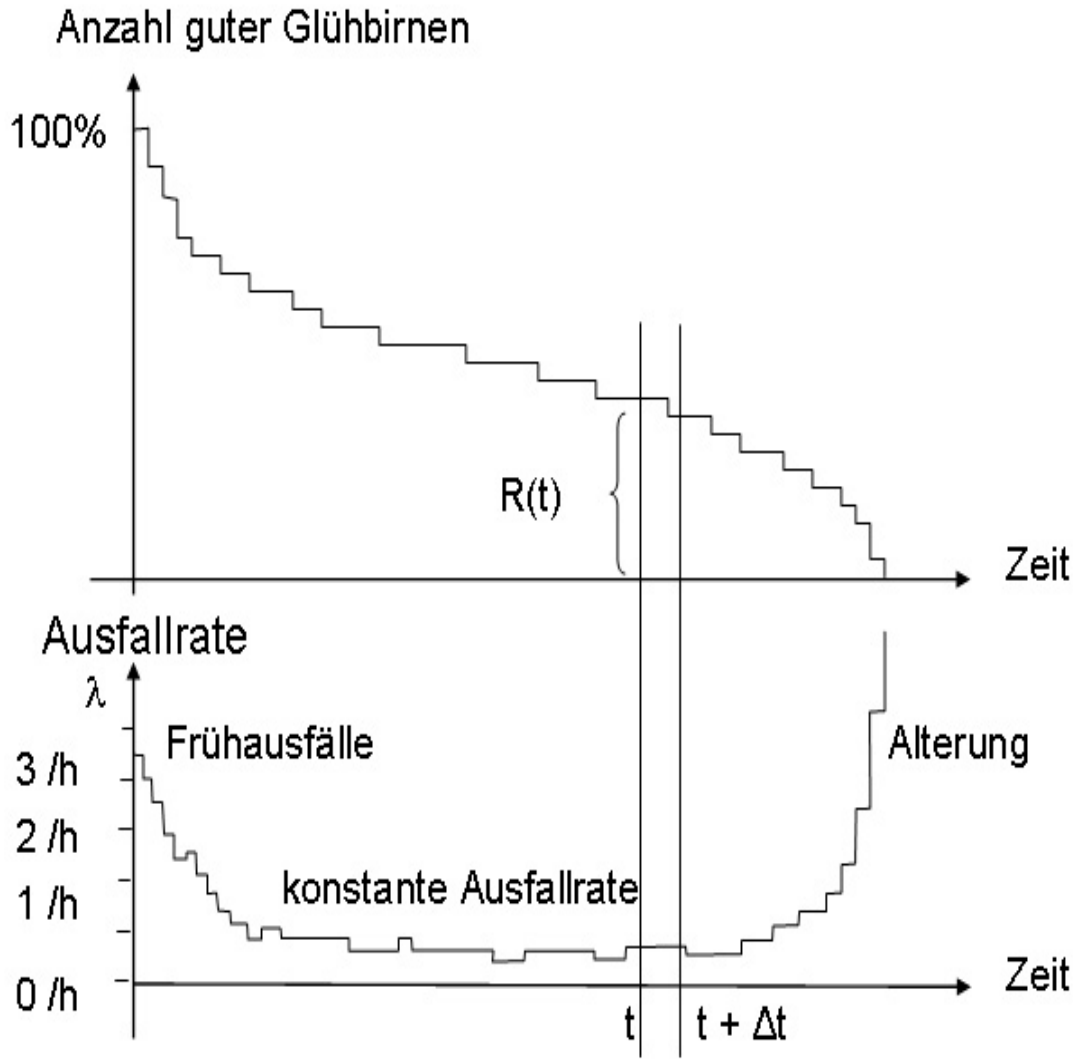
$$\lambda = \frac{Anzahl\ ausgefallener\ Lampen}{Anzahl\ heiler\ Lampen\ bei\ Beginn\ des\ betrachteten\ Zeitraums}$$

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

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

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

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


Anzahl guter Glühbirnen
100%
R(t)
Zeit
Ausfallrate
λ
Frühausfälle
Alterung
3 /h
2 /h
konstante Ausfallrate
1 /h
0 /h
Zeit
t
t + Δt

Bestimmung der Zuverlässigkeit (Forts.)

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:

Miltitary Reliability Handbook, MIL-HDBK-217D

MIL-HDBK-718

Siemens SN29500

RAC

RDP

Bellcore

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 in „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.

Bestimmung der Zuverlässigkeit (Forts.)

Die folgende Tabelle aus Wikipedia zeigt Fehlerraten für einige elektronische Komponenten:

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.)

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

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

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 $R(t)$

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

Die Funktion

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

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

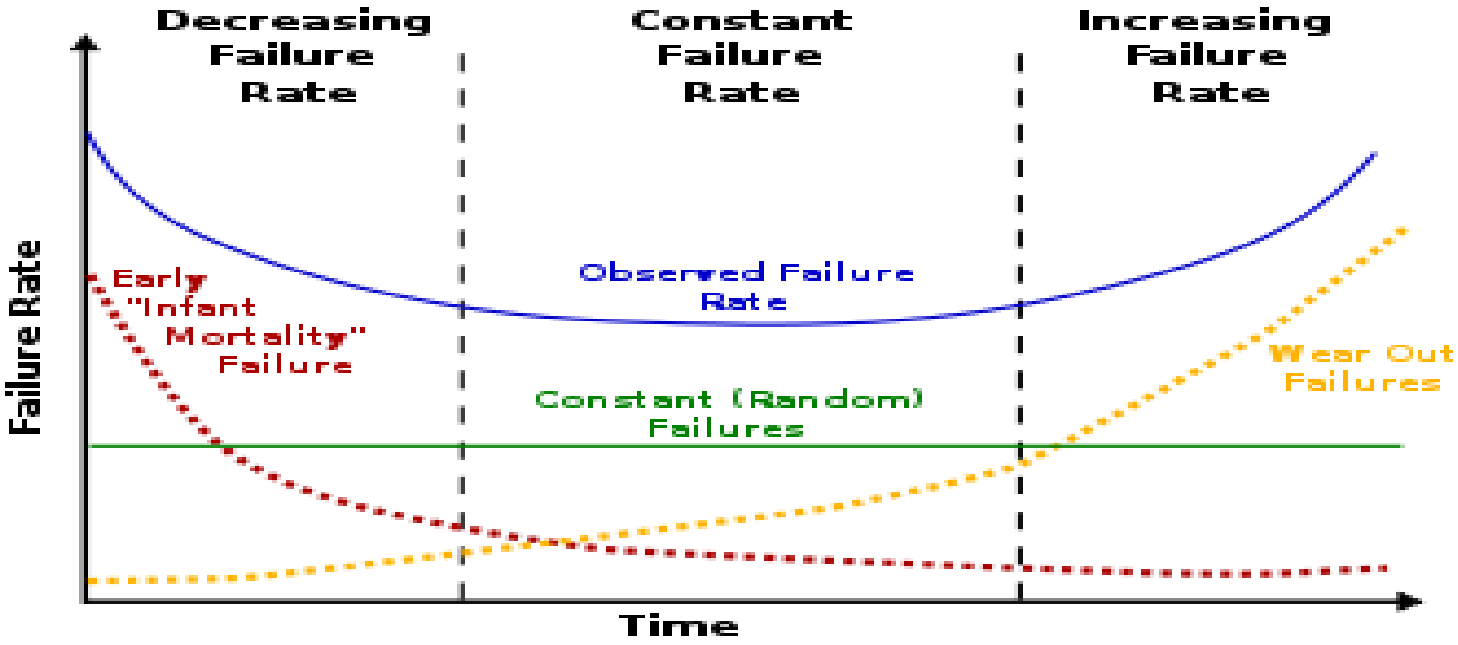
Theoretische Zuverlässigkeit (Forts.)

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:

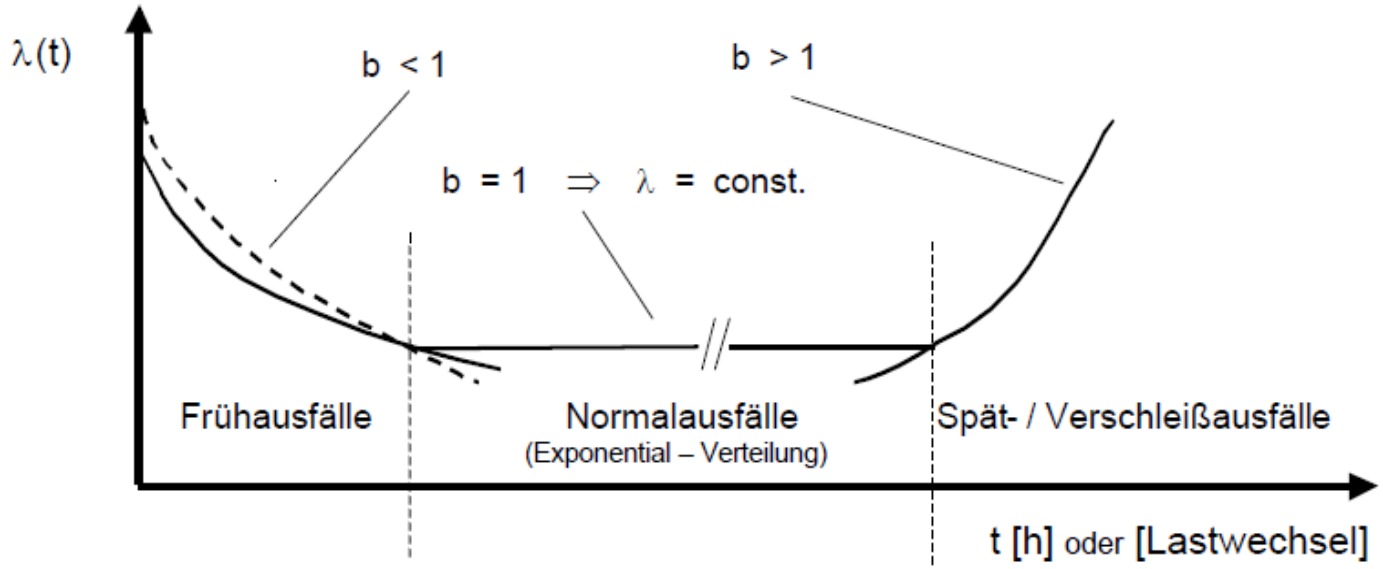
Theoretische Zuverlässigkeit (Forts.)



- FR (Failure Rate) = Fehleranzahl / Komponentenanzahl pro Zeitintervall

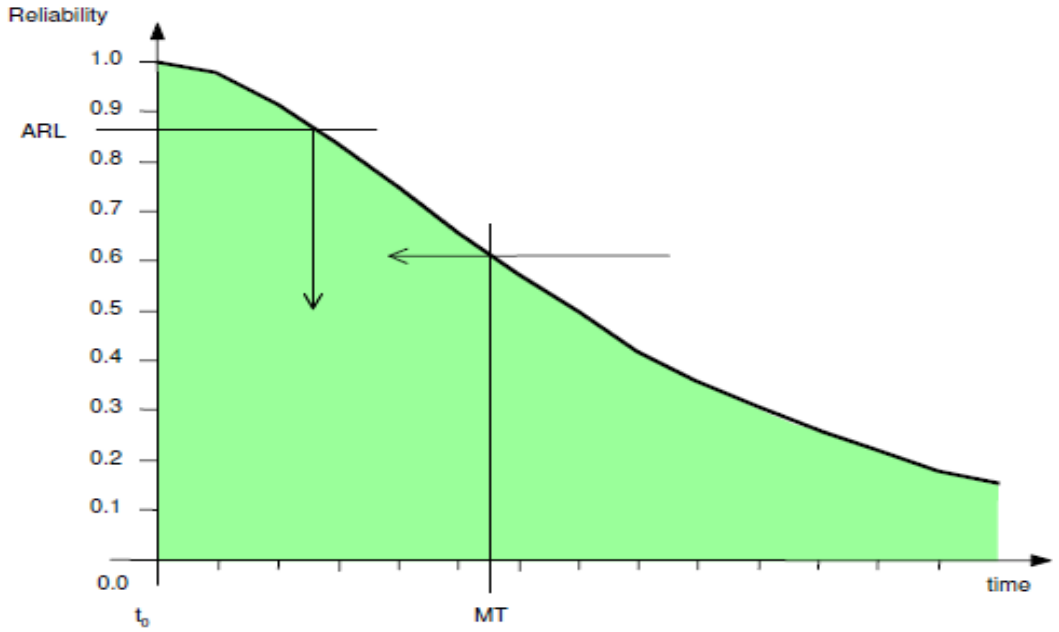
Theoretische Zuverlässigkeit (Forts.)

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.



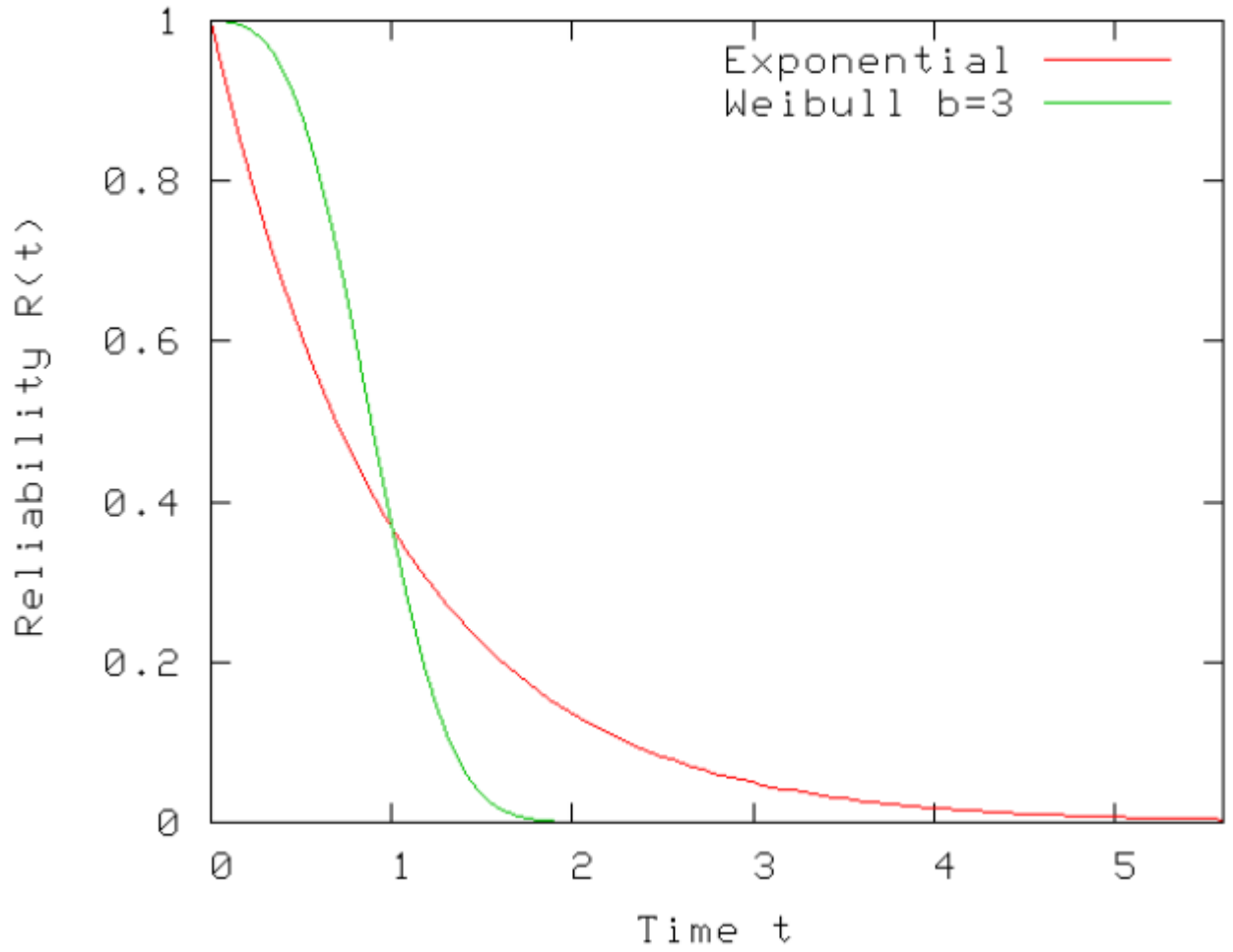
Dichtefunktion $f(t) = \frac{b}{T} . (\frac{t}{T})^{b-1} . e^{-(\frac{t}{T})^b}$ für t, T, b ≥ 0

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

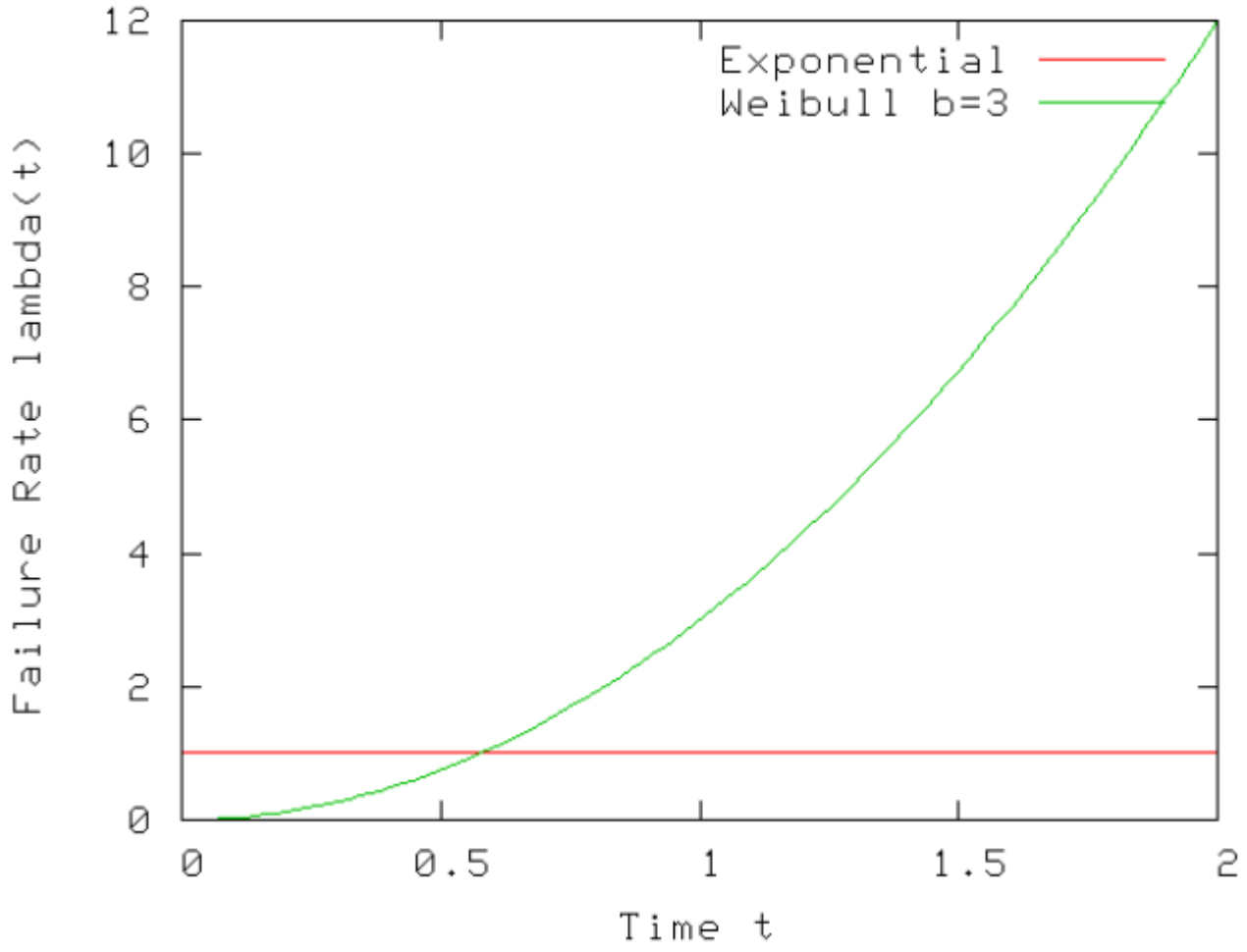


MT Mission Time: Zeitpunkt, bis zu dem man an der Zuverlässigkeit interessiert ist
ARL Accepted Reliability Level: Unterhalb dieses Wertes wird das Produkt als unsicher betrachtet

$$\text{MTTF} = \int_0^\infty R(t)dt$$


Exponential
Weibull b=3
Reliability R(t)
Time t
0
0.2
0.4
0.6
0.8
1
0
1
2
3
4
5

[29]


Exponential
Weibull b=3
Failure Rate lambda(t)
Time t
0
2
4
6
8
10
12
0
0.5
1
1.5
2

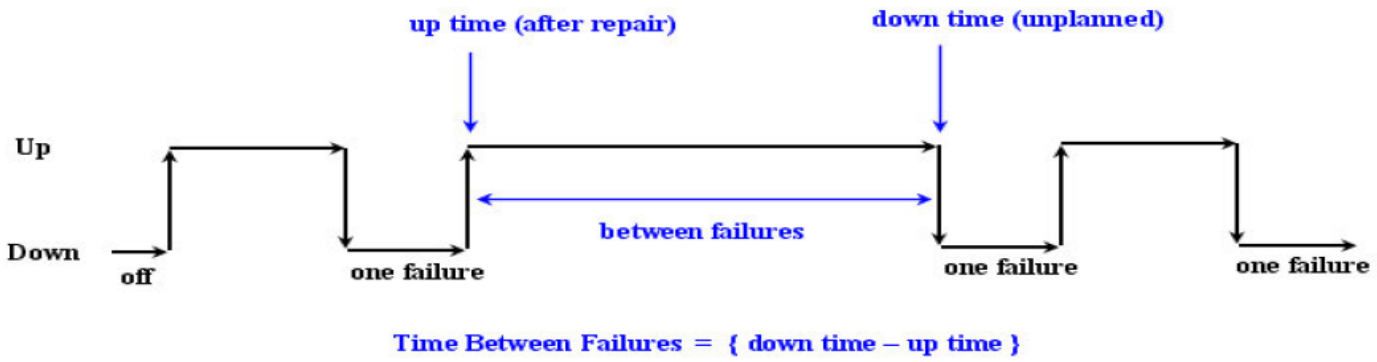
[29]

2.2 Qualität technischer Systeme

- Fehlerfunktionen (Zusammenstellung)
  - MTBF (Mean-Time-Between-Failure [mittlerweilen auch Mean-Time-Before-Failure])

    Mittlere Betriebsdauer/Erwartungswert der Betriebsdauer zwischen zwei auf einander folgenden Ausfällen. Wenn up-time der Einschalt- und down-time der nächste Ausfallzeitpunkt sowie n die Zahl der Ausfälle während des Beobachtungszeitraums ist:

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



2.2 Qualität technischer Systeme

- Fehlerfunktionen (Zusammenstellung)

  – 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

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

$$AFR = 8760/MTBF$$

  – POH (Power-On Hours)

$$AFR = {}^{1}/_{MTBF} \cdot POH$$

  – Reparaturrate μ = μ(t) = const

  μ = 1/ MTTR Mittlere Reparaturrate

2.2 Qualität technischer Systeme

- Fehlerfunktionen (Zusammenstellung)
  - 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 sind über 6 Monate im Einsatz
      Die Einsatzdauer als Summe beträgt 100*0,5 = 50 Plattenjahre.
      Wir würden 50 / 20 = 2,5 Ausfälle erwarten.

  - Fehler- oder Ausfallrate $\lambda = \lambda(t) = const$
    $\lambda = 1/ MTTF$ Mittlere Zeit bis zum Ausfall

2.2 Qualität technischer Systeme

- Fehlerfunktionen (Zusammenstellung)
  - MTTDL (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 MTTDL = MTTF. (Schlecht!)

    Normalerweise sollte MTTDL >> MTTF sein.

  - Availabiltliy = MTBF / (MTBF + MTTR)

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

  - Wie verhalten sich MTTR 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.

2.2 Qualität technischer Systeme

- Wie verhalten sich MTTR 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 veröffentlich, 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.

2.2 Qualität technischer Systeme

- 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

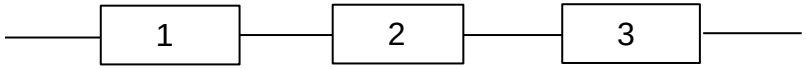
2.2 Qualität technischer Systeme

Fehlerabschätzungen bei zusammengesetzten Systemen

Welche Fehlerabschätzungen kann man für ein Produkt angegeben, 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.

Serienschaltung



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:

$$R_s(\mathrm{t}) = R_1(\mathrm{t}) \times R_2(\mathrm{t}) \times R_3(\mathrm{t}) \times \ldots$$

Für die Exponentialverteilung gilt

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

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

mit Fehlerrate λ_s fürs Gesamtsystem

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

2.2 Qualität technischer Systeme

Fehlerabschätzungen bei zusammengesetzten Systemen (Forts.)

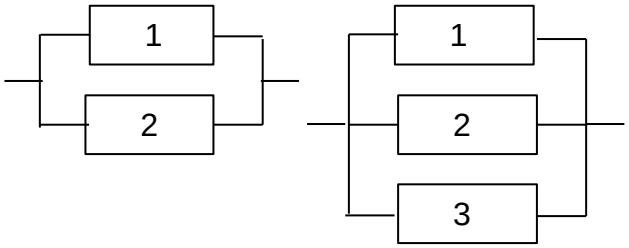
MTBF bei zwei seriell zusammengesetzten Komponenten *a* und *b* beträgt , wenn $MTBF_a = MTBF_b$ ist,

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

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

Fehlerabschätzungen bei zusammengesetzten Systemen (Forts.)

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.

Fehlerabschätzungen bei zusammengesetzten Systemen (Forts.)

1. Redundanz, nicht reparierbar

Die einzelnen nichtreparierbaren Komponenten werden als unabhängig von einander in ihrem Fehlerverhalten 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

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

$$F_s(\mathrm{t}) = F_1(\mathrm{t}) \times F_2(\mathrm{t}) \times F_3(\mathrm{t}) \times \ldots$$

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

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

Fehlerabschätzungen bei zusammengesetzten Systemen (Forts.)

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

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

$$R_s(\mathrm{t}) = 2 \cdot exp(-\lambda t) - exp(-2\,\lambda t)$$

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

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]).

Fehlerabschätzungen bei zusammengesetzten Systemen (Forts.)

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(System) = MTTR(Komponente a) = MTTR(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

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

Fehlerabschätzungen bei zusammengesetzten Systemen (Forts.)

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(System) = MTTR(Komponente a) / 2 = MTTR(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 Markoffsystem erhält man mit einigem Aufwand die andere Abschätzung

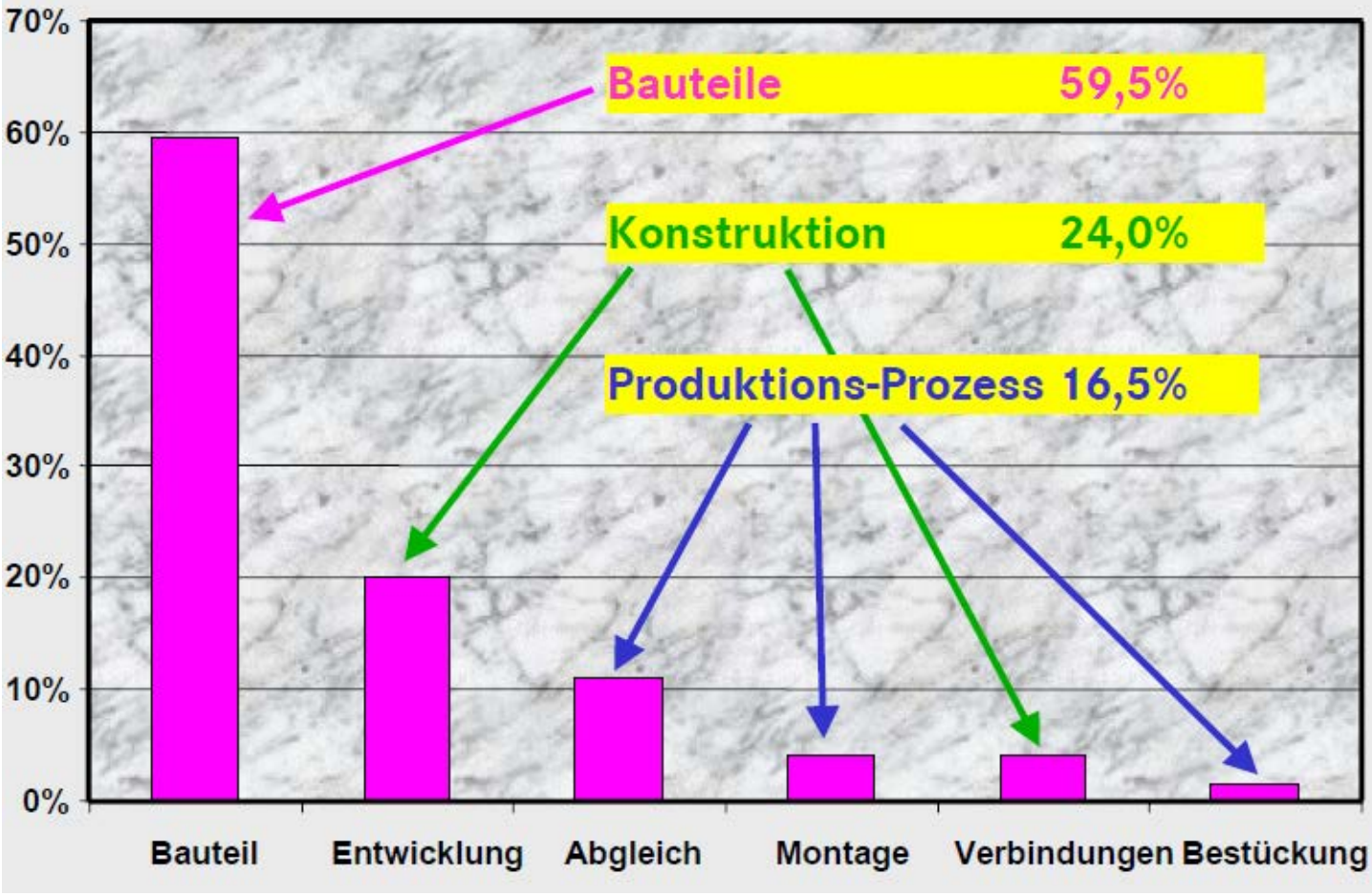
$$\lambda_{parallel} = \frac{2\mu}{\lambda^2} + \frac{3}{2\lambda}$$

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 aif den beiden folgenden Folien zu sehen ist.

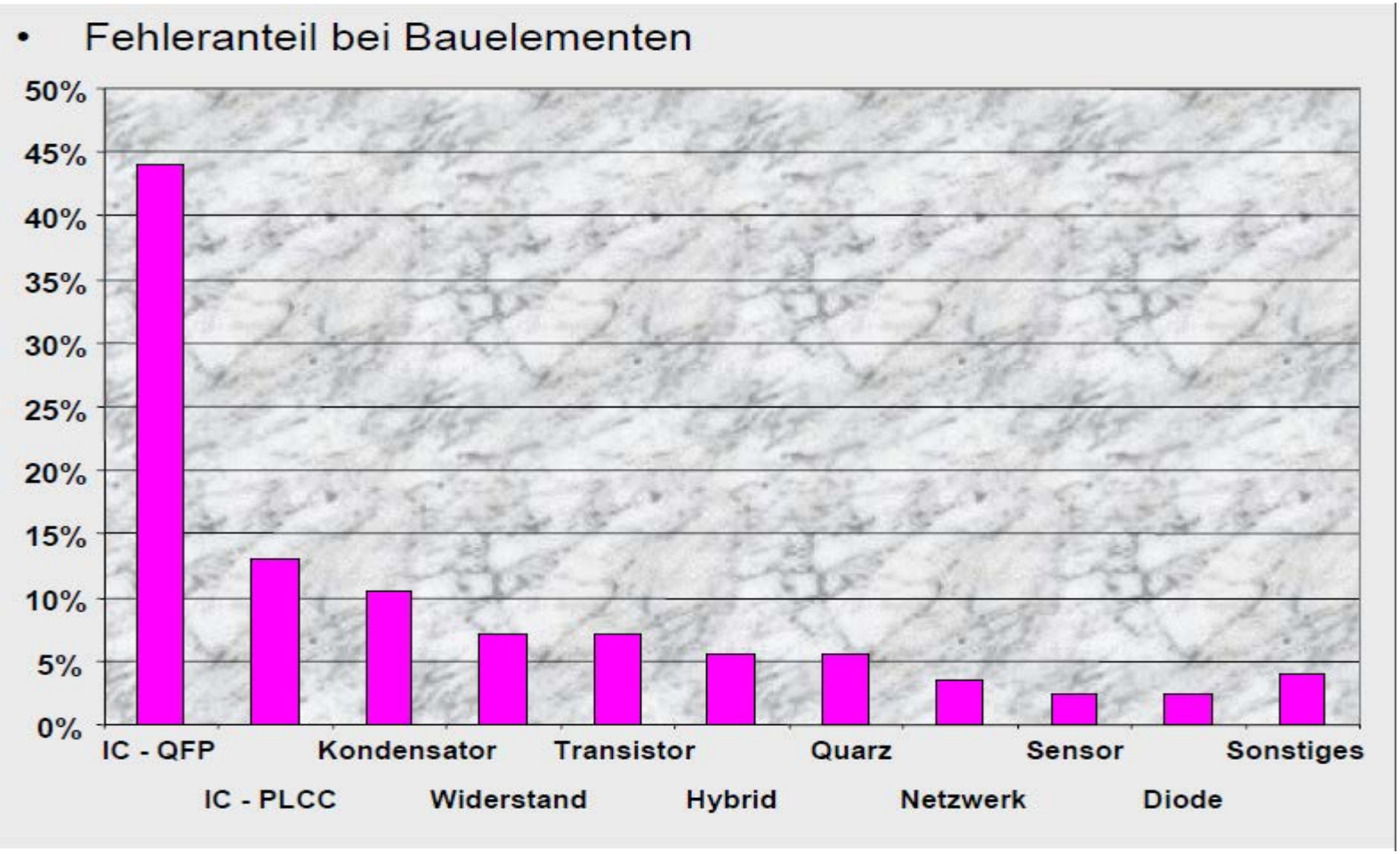
Fehlerabschätzungen bei zusammengesetzten Systemen (Forts.)

- Anteil an Ausfällen



[25]

Fehlerabschätzungen bei zusammengesetzten Systemen (Forts.)



[25]

2.3 Qualität magnetischer Speichermedien

- 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 MTTDL
  - Für das ST5800 Honeycomb-Storagesystem der Fa. Sun wurde eine MTTDL von 2,4 x 10^6 Jahren angegeben. Das macht scheinbar jedes Backup überflüssig. Dem würde keiner folgen. Angenommen wir haben 10 ST5800 im Einsatz, von denen n Systeme, 1<=n<=5, bereits nach einem Jahr ausfallen, 10-2n Systeme nach 2,4 x 10^6 Jahren und n Systeme nach 4,8 x 10^6 Jahren. Trotzdem wäre die MTTDL der ST5800 2,4 x10^6 Jahre.

2.3 Qualität magnetischer Speichermedien

- Beschränkte Aussagekraft von MTTDL (Forts.)
  - 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 Untersuchen in der Praxis zeigen. U.a. haben Google und NetApp haben großflächige Untersuchungen durchgeführt.

- Spreadsheet für MTTDL-Berechnung für verschiedene Platten und Fehlertypen, s. Zetta_MTTDL_June10_2009.xls in [9]

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 angeben 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.

2.3 Qualität magnetischer Speichermedien

- 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 Google und NetApp

    NetApp: 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:

      Die Beobachtung lief über 41 Monate. Dabei wurden über 1.5 Mill. Platten untersucht. Dabei ergab sich eine Fehlerquote von 4x 10^5 Silent Corruption. 3x 10^4 wurden erst bei der RAID-Rebuild entdeckt. Sie hätten zu Datenverlust geführt.

    Google: Die Aussage der Hersteller zu den MTBF-Angaben, die Aussagekraft von SMART-Statistiken sowie Verhalten der Platten abhängig von der Belastung, Lebensalter und Temperaturen. [33]

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 fast 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.

3. RAID

- Redundant Array of Inexpensive Disks
  - RAID 6 kann den Ausfall zweiter 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.
  - Mittlerweilen wird der Begriff als „Redundant Array of Independent Disks“ gelesen.

3. RAID

- Redundant Array of Inexpensive Disks

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]

3. RAID - DDF

- Ist der RAID-Aufbau von bestimmten Produkten wie Kontroller abhängig? Oder - anders gefragt - ist das RAID-Format 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

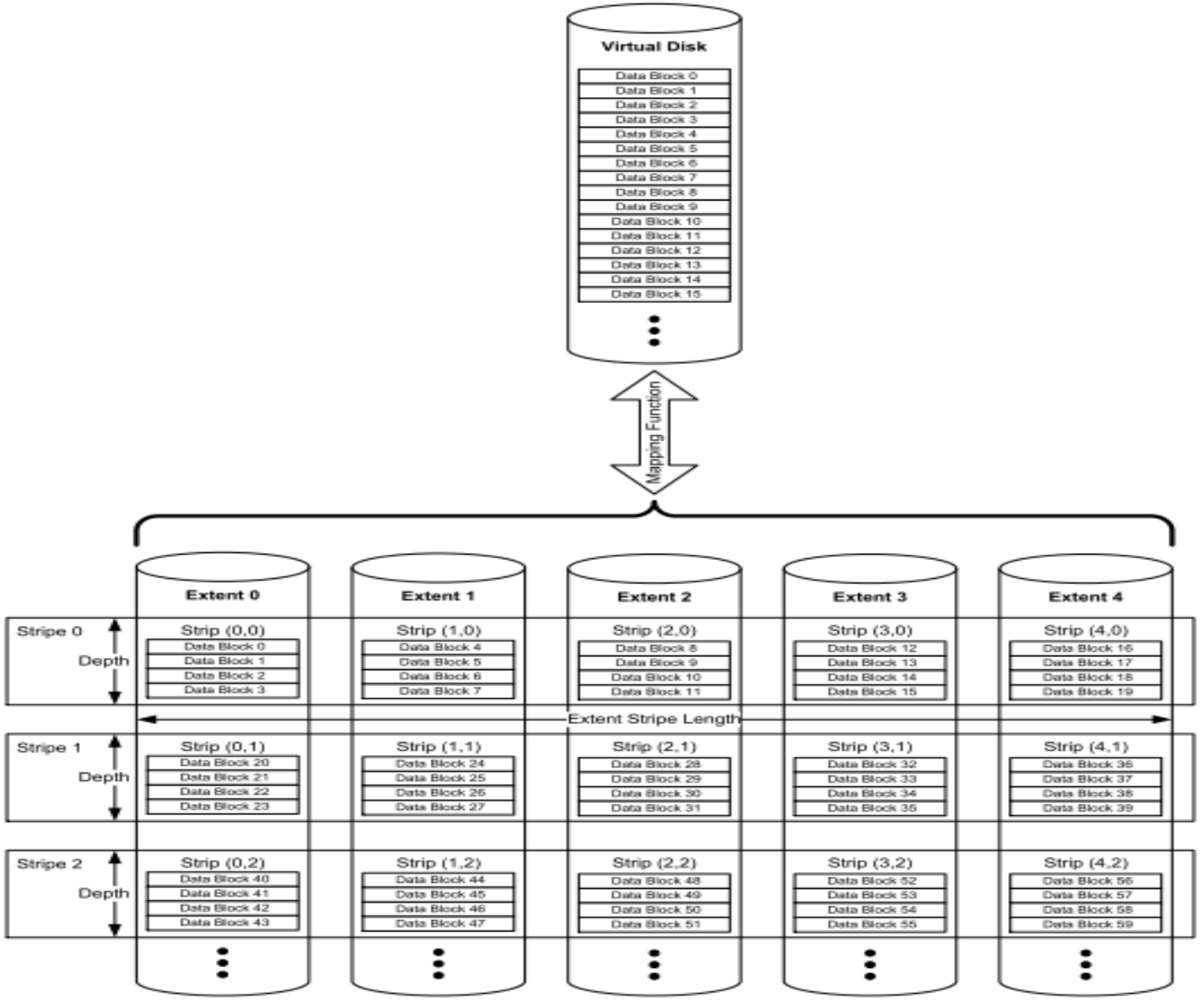
3. RAID - DDF

- 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 Kontrollers
  - 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

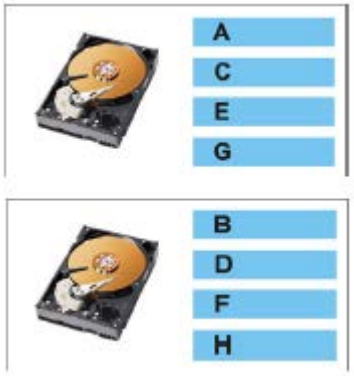
- Disk Data Format (DDF)
  - Varianten mit Hot-Spare-Bereiche
    - Zusätzliche Platte
    - Reservierung auf Daten- und Paritätsplatten

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



3. HW-RAID

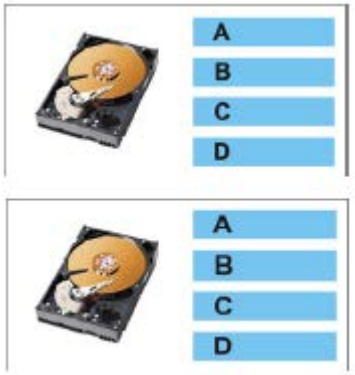
- RAID 0
  - Striping
  - Kein echtes RAID-Verfahren
  - Daten parallel auf alle Festplatten verteilt
  - Bei Plattenausfall sind die Daten des kompletten Raids verloren.



Mindestanzahl n benötigter Festplatten: n=2
Maximal verfügbare Nutzkapazität k: k=n
Ausfallsicherheit s (in #Festplatten): s=0

3. HW-RAID

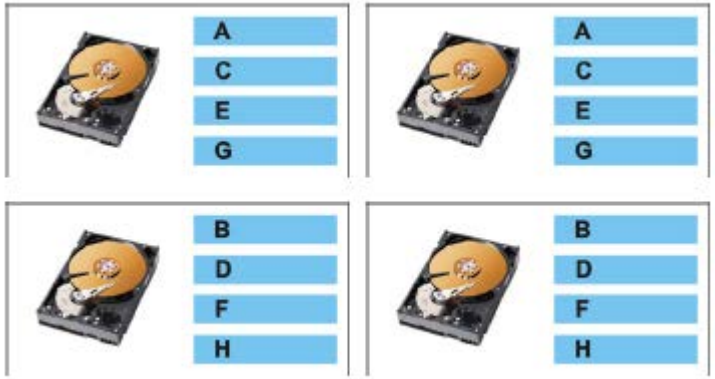
- RAID 1
  - Mirroring/Spiegelung
  - 100% Redundanz
  - Daten parallel auf zwei Festplatten gespeichert
  - Bei Plattenausfall werden Daten vom Spiegel gelesen.



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

3. HW-RAID

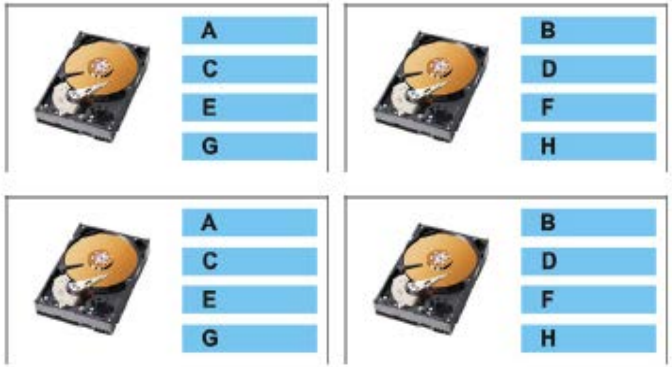
- RAID 01
  - Weitere Bezeichnungen RAID 0+1, RAID 0,1 oder RAID 0/1
  - Kombiniert ein RAID 0 mit einem RAID 1 (ein RAID 0 wird gespiegelt), um die Transferrate zu erhöhen bei gleichzeitiger 100%iger Redundanz.



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

3. HW-RAID

- RAID 10
  - Kombiniert ein RAID 1 (Spiegelung) mit einem RAID 0 (Striping), um die Transferrate zu erhöhen bei gleich-zeitiger 100%iger Redundanz.



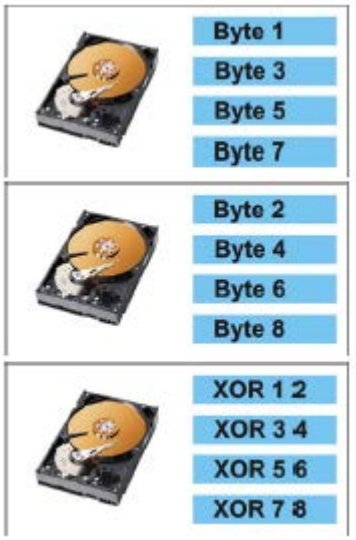
  - Der Unterschied zwischen RAID 10 und RAID 01 zeigt sich im Fehlerfall:
    - Fällt bei RAID 01 eine Festplatte aus, so ist das betroffene RAID 0 unbrauchbar. Die Daten sind aber über das zweite RAID 0 (den Spiegel) verfügbar. Fällt eine weitere Platte aus dem zweiten Verbund aus, so sind die Daten 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.

3. HW-RAID

- 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.

Mindestanzahl n benötigter Festplatten:	n=10
Maximal verfügbare Nutzkapazität k:	k=8
Ausfallsicherheit s (in #Festplatten):	s=2

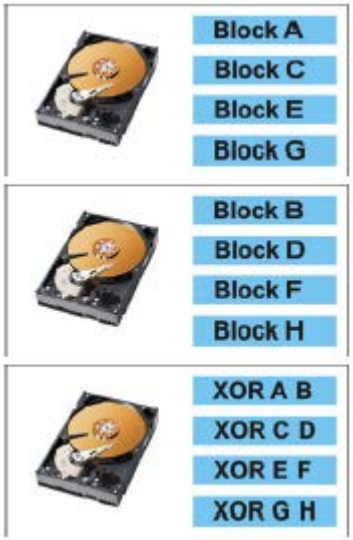
3. HW-RAID



- RAID 3
  - Vorgänger von RAID 4 und 5
  - Verfügt über Redundanz
  - Daten werden byteweise auf mindestens zwei Festplatten gespeichert.
  - Über XOR werden aus den Daten Paritätsbits und byte-weise auf eine dedizierte Paritätsplatte gespeichert.
  - Die Paritätsplatte ist der ‚Flaschenhals'.
  - Synchronisiert die Kopfbewegungen der Laufwerke beim Schreiben berechnet wird über Bei Plattenausfall werden Daten vom Spiegel gelesen.

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

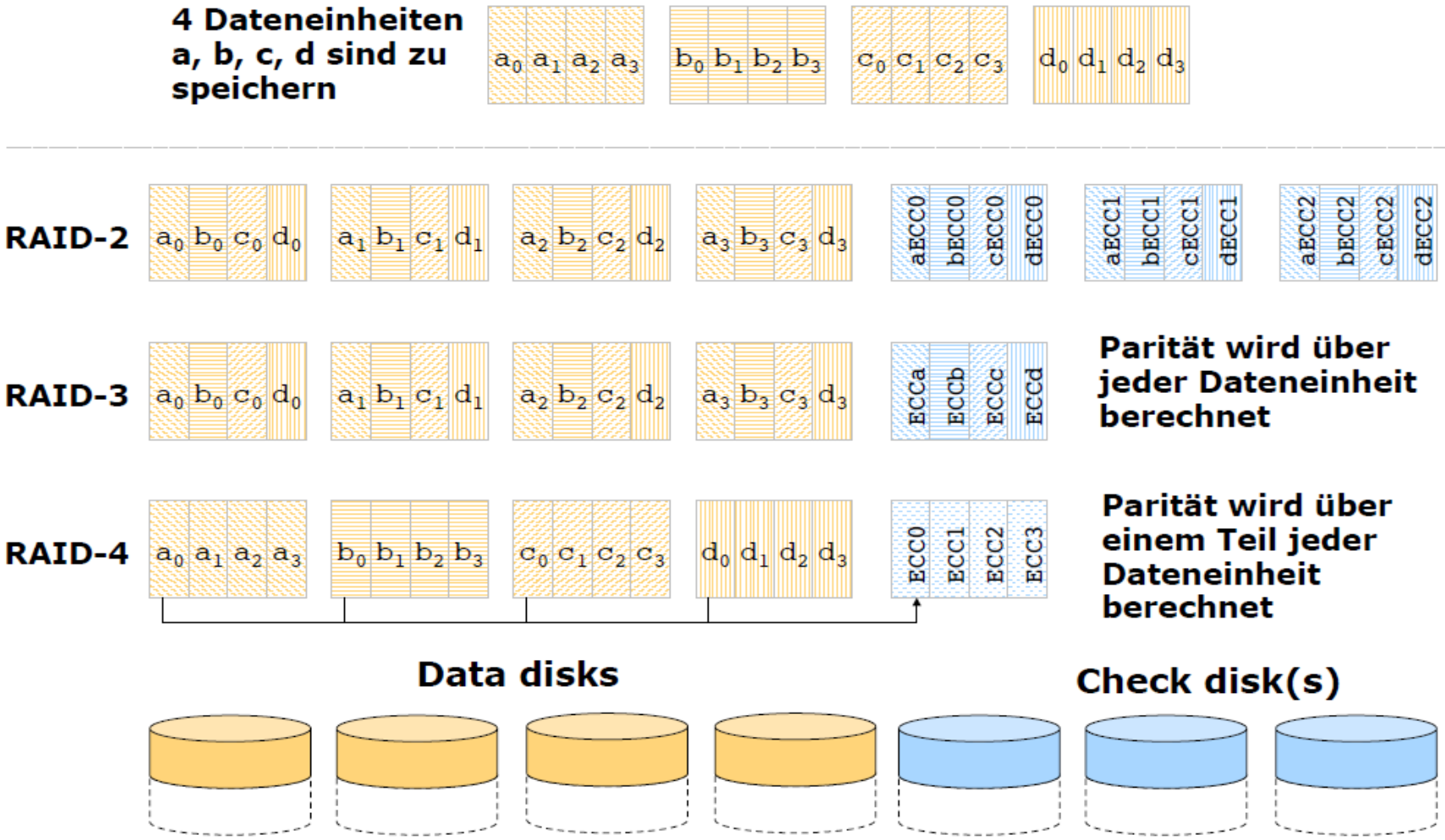
3. HW-RAID



- 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

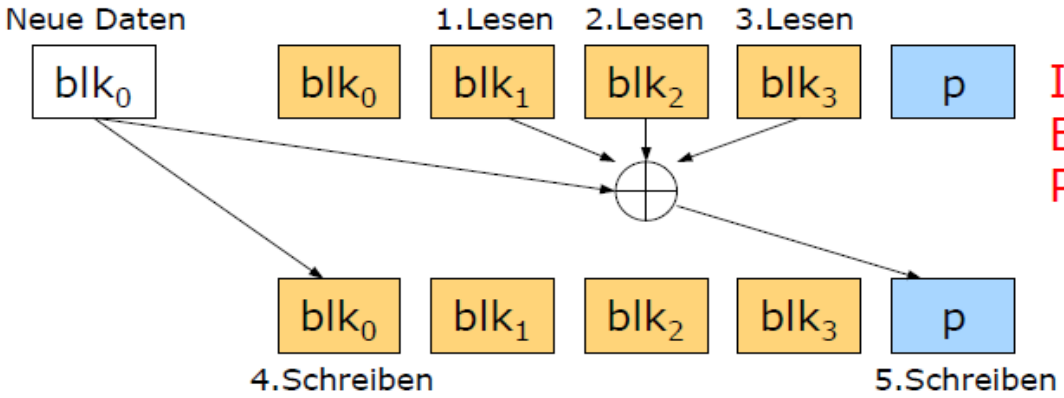
Vergleich der RAID-Leves 2, 3 und 4



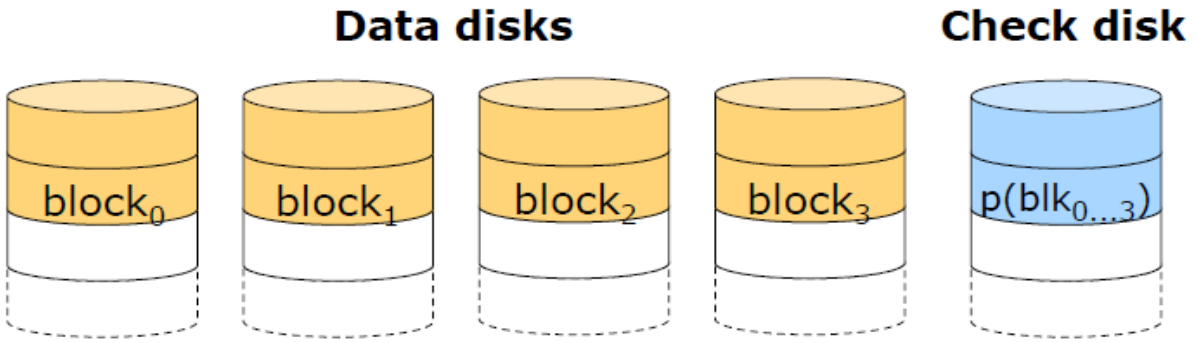
[32]

RAID 4

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



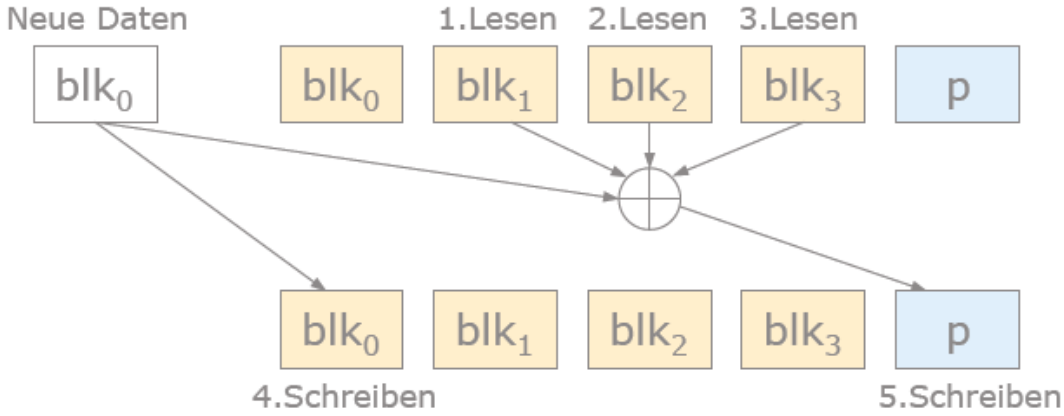
Ineffiziente Berechnung der Parität

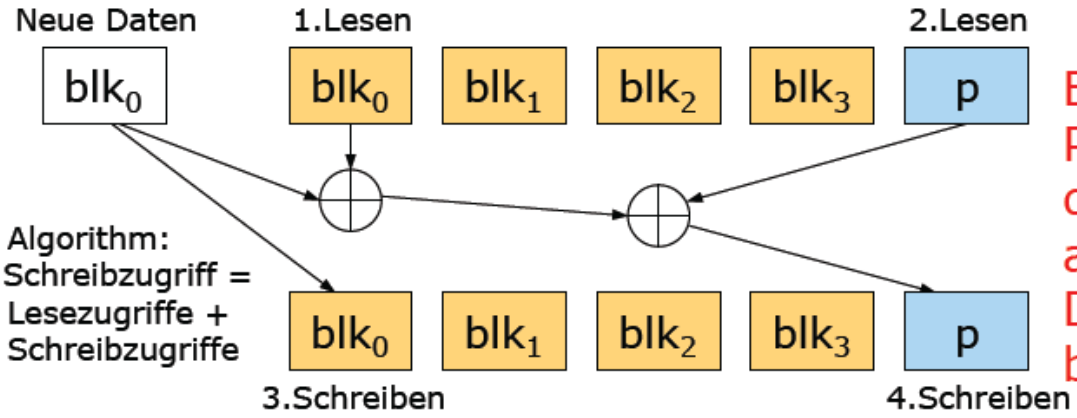


[32]

RAID 4

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



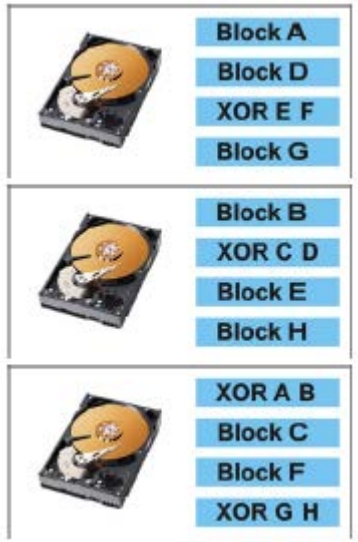


Small-Write Algorithm:
1 Logischer Schreibzugriff =
2 physische Lesezugriffe +
2 physische Schreibzugriffe

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

[32]

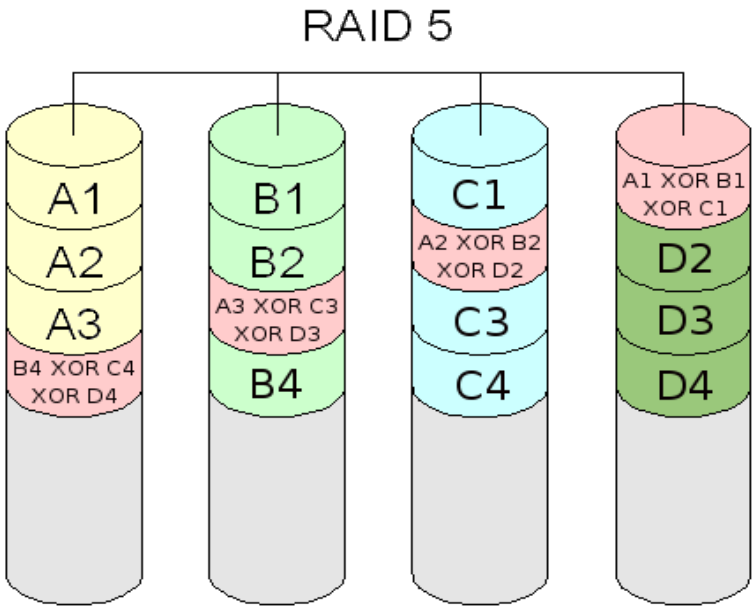
3. HW-RAID



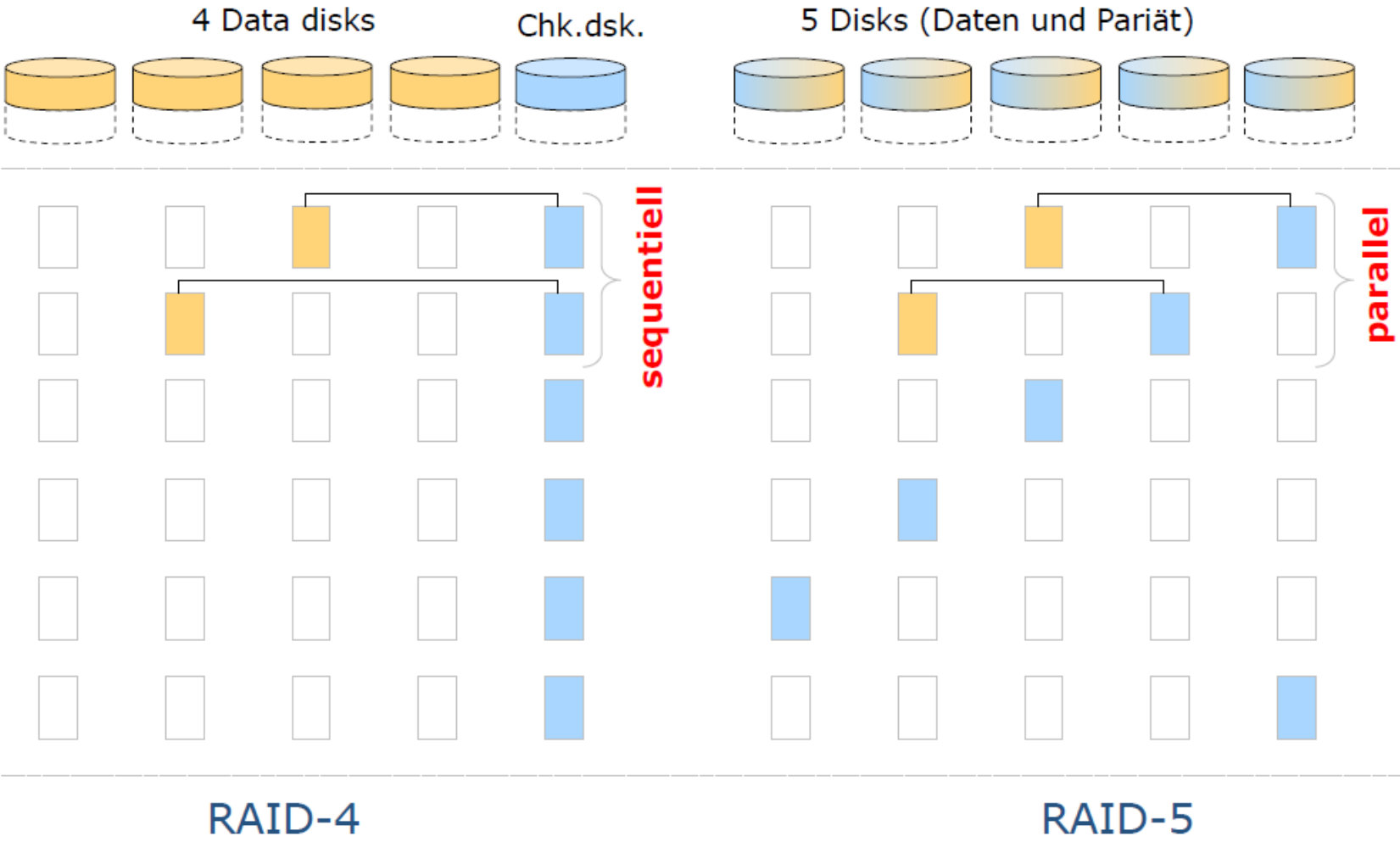
- RAID 5
  - Vorgänger RAID 3 und 4
  - Verfügt ebenfalls über Redundanz
  - Die Daten werden blockweise verteilt. Die dedizierte Paritätsplatte entfällt. Der Flaschenhals entfällt. Die ECC-Daten werden gleichmäßig über alle Laufwerke verteilt.
  - Beim Lesen Transferraten wie RAID 4. Der Schreibzugriff 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

3. HW-RAID


RAID 5
A1
A2
A3
B4 XOR C4
XOR D4
B1
B2
A3 XOR C3
XOR D3
B4
C1
A2 XOR B2
XOR D2
C3
C4
A1 XOR B1
XOR C1
D2
D3
D4

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



[32]

3. HW-RAID

- RAID 6 und 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.

  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 Total-ausfall und einem zusätzlichen Bad Block Failure gerechnet werden.

  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 andere in Mitleidenschaft gezogen, was sich erst beim Rebuild zeigt. Dann war die Annahme der Unabhängigkeit von Fehlern eventuell zu optimistisch.

3. HW-RAID

- RAID 6 und weitere RAID-Verfahren

  Deshalb wurden weitere auf RAID 5 aufbauende RAID-Verfahren entwickelt, die den Ausfall von maximal 2 Platten kompensieren können.

- RAID 6

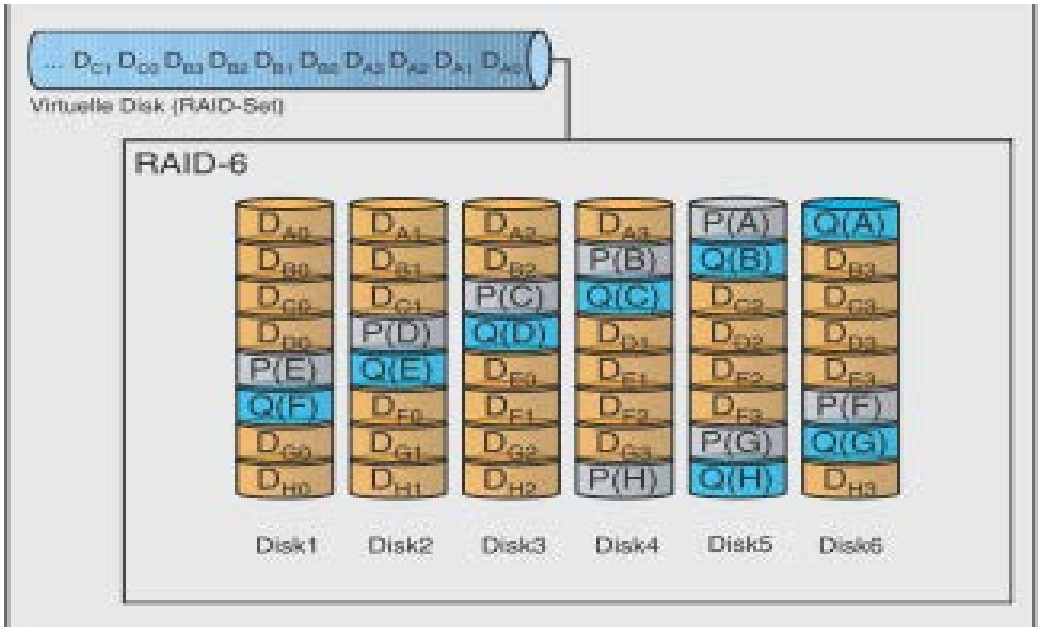
  RAID 6 arbeitet mit zwei Paritätsplatten. Da die Paritäten unabhängig von einander sind, stehen zwei unabhängige Werte zur Fehlerkorrektur zur Verfügung. Maximal zwei Fehler können damit korrigiert werden.

  Beide Paritäten werden gleichmäßig über alle Platten verteilt. Die erste Parität wird wieder mit XOR ermittelt und erfordert nicht viel Aufwand. Bei der zweiten Parität ist es anders. Der Rechenaufwand zum Erstellen der Korrekturwerte ist erheblich höher.

3. HW-RAID

- RAID 6

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

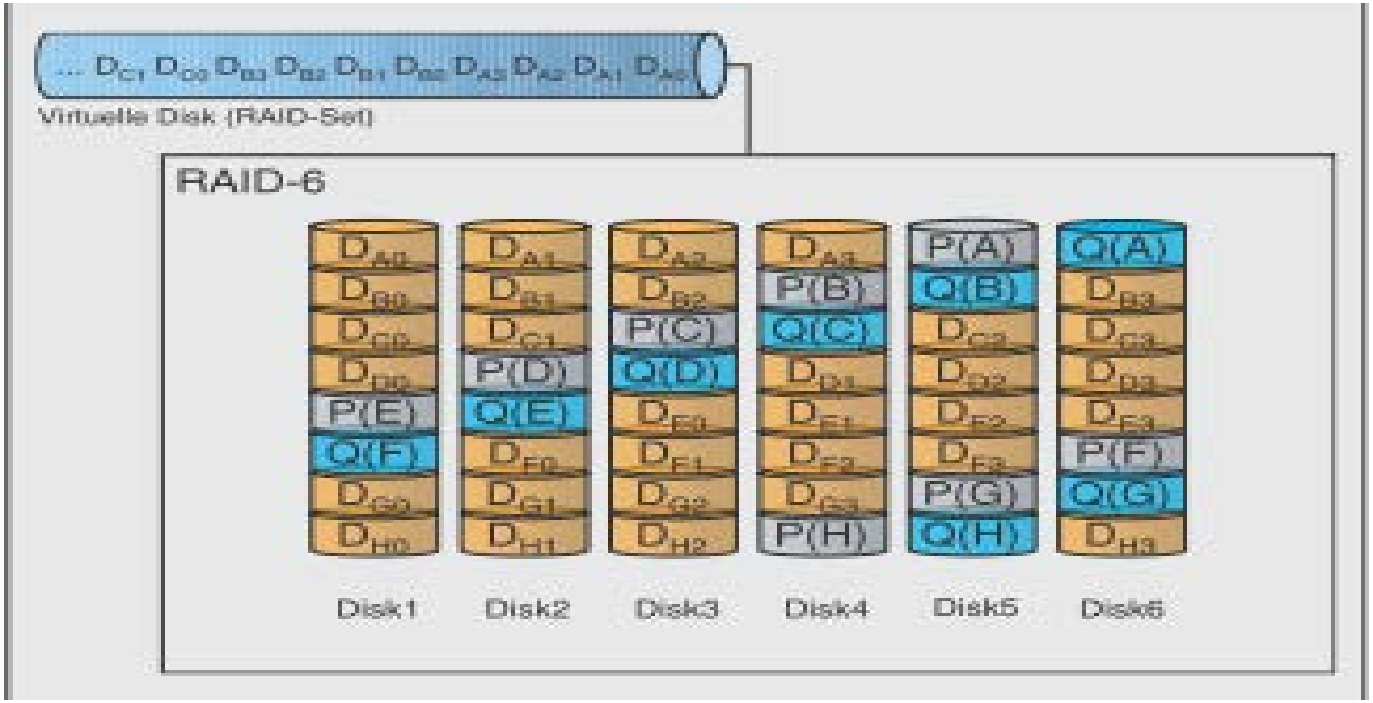


Mindestanzahl n benötigter Festplatten: $n \geq 3$ Maximal verfügbare Nutzkapazität k: $k = n-2$

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

3. HW-RAID

- RAID 6



3. HW-RAID

- 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

3. HW-RAID

- RAID DP
  - Network Applicance

Weiterentwicklung von RAID 4: Zwei Paritäten P und Q, wobei die erste P wie bisher horizontal, die zweite Q 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.

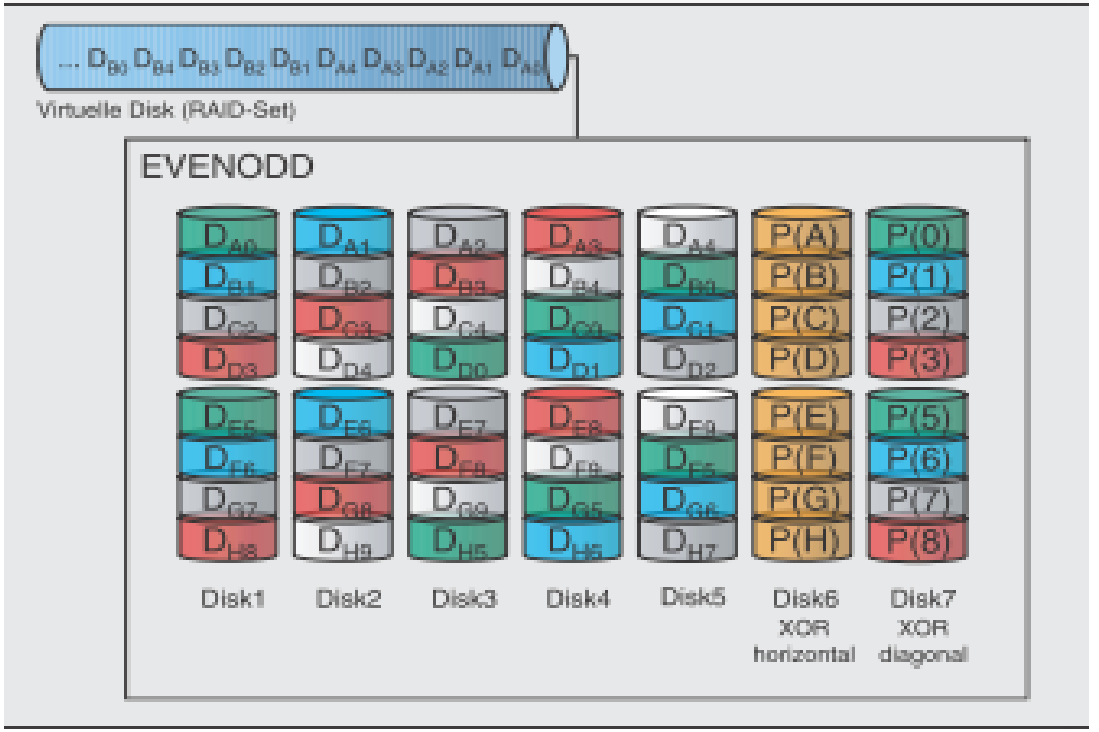
$P_1 = XOR(A_1, B_1, C_1)$
$P_2 = XOR(A_2, B_2, C_2)$
$P_3 = XOR(A_3, B_3, C_3)$

$Q_1 = XOR(P_1, A_2, B_3, 0)$
$Q_2 = XOR(P_2, A_3, 0, C_1)$
$Q_3 = XOR(P_3, 0, B_1, C_2)$
$Q_4 = XOR(P_4, A_5, B_6, 0)$
...

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

3. HW–RAID

- Zweifaches XOR im RAID-4 Stil
  - RAID 5DP (RAID 5 Double Parity von HP)
  - RAID ADP von HP
  - EVENODD – Anzahl der Datenplatten muß eine Primzahl sein



3. HW-RAID

- Verläßlichkeit
  - Performance -> RAID 10 (0+1)
  - Hohe Kapazitäten und besten Schutz gegen Auszeiten und Datenverlust -> RAID 6

- Datenverlust beim Rebuild (Beispiel von Intel):

  Bei einem Array mit 60 Laufwerken muß unter normalen Betriebsbedingungen und gleichen Fehlerwahrscheinlichkeiten damit gerechnet werden für

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

  RAID 6: nach 25 Jahren

3. HW-RAID

- Fehlerbetrachtungen:
  - Beim Rebuild müssen alle Sektoren fehlerfrei bearbeitet werden können
  - Probleme durch defekte Platten, defekte Plattensektoren, Einzelbitfehler

  Fehler hängen von der Kapazität ab. Sie ist leicht zu berechnen.
  MTTDL ist ein Maßstab für Reliability, Availability und Serviceability.

  - Für nichtgeschützte Daten (Dynamisches Striping, RAID-0):
    MTTDL = MTTF / N

  - Für Syteme mit einfache Parität (RAID-1, RAID-5, 2-fach Spiegel)
    MTTDL = MTTF(disk1) * MTTF(disk2) / (N*(N-1)*MTTR(disk1))

  - Für Systeme mit zweifacher Parität (RAID-6, RAID-DP, RAID-Z2)
    MTTDL = MTTF(disk1)*MTTF(disk2)*MTTF(disk3) /
    (N*(N-1)*(N-2)*MTTR²(disk))

3. HW-RAID

- Fehlerbetrachtungen bei LSE und N Platten:

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

Sei f der Fehler während des Rebuilds. Dann gilt

$$f = 1 - (1 - LSE_{Rate})\text{^}BitsPerDisk(N-1)$$

$$MTTDL_{RAID5\_LSE} = MTTF_{Erster\ Fehler} / f$$
$$= MTTF_{Platte} / (f \cdot N)$$

Die Summe der beiden Fehlergrößen ergibt für RAID 5

$$MTTDL_{RAID5} = ((MTTDL_{RAID5\_DF})^{-1} + (MTTDL_{RAID5\_LSE})^{-1})^{-1}$$

3. HW–RAID
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.

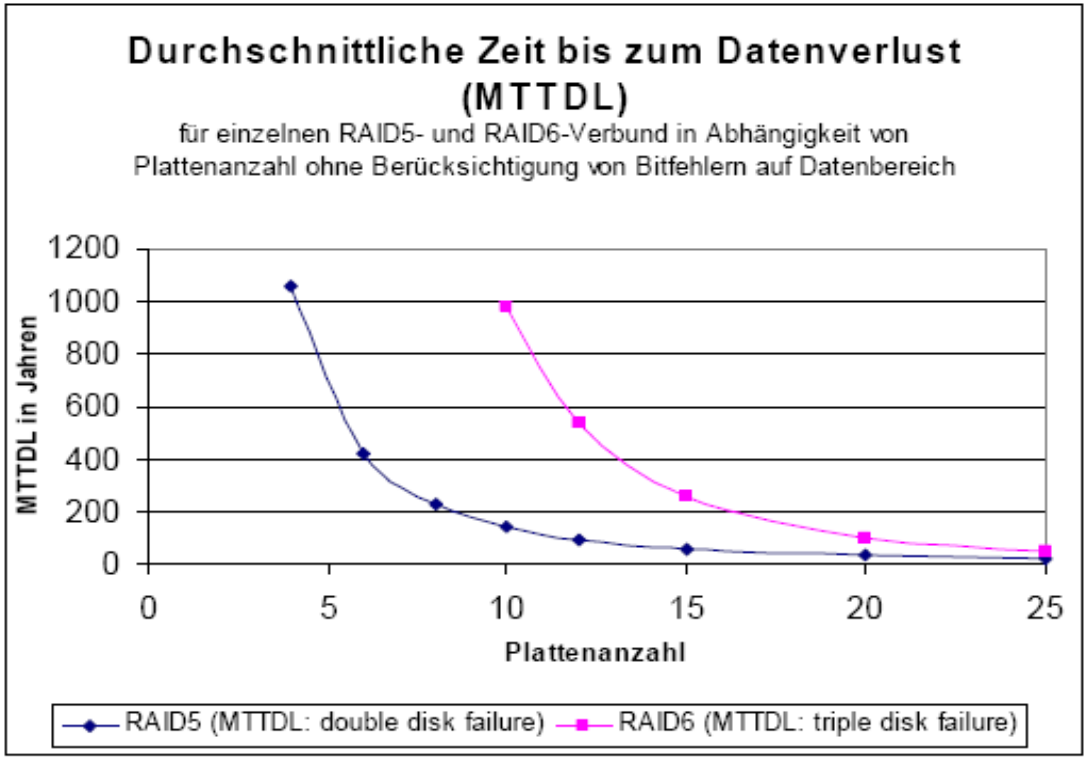


Abbildung 1: Berechnungsgrundlage MTTF(disk1)=200000h, MTTF(disk2)=20000h, MTTF(disk3)=2000h, MTTR=36h

[19]

3. HW–RAID

Eine Prüfsumme ist nicht genug

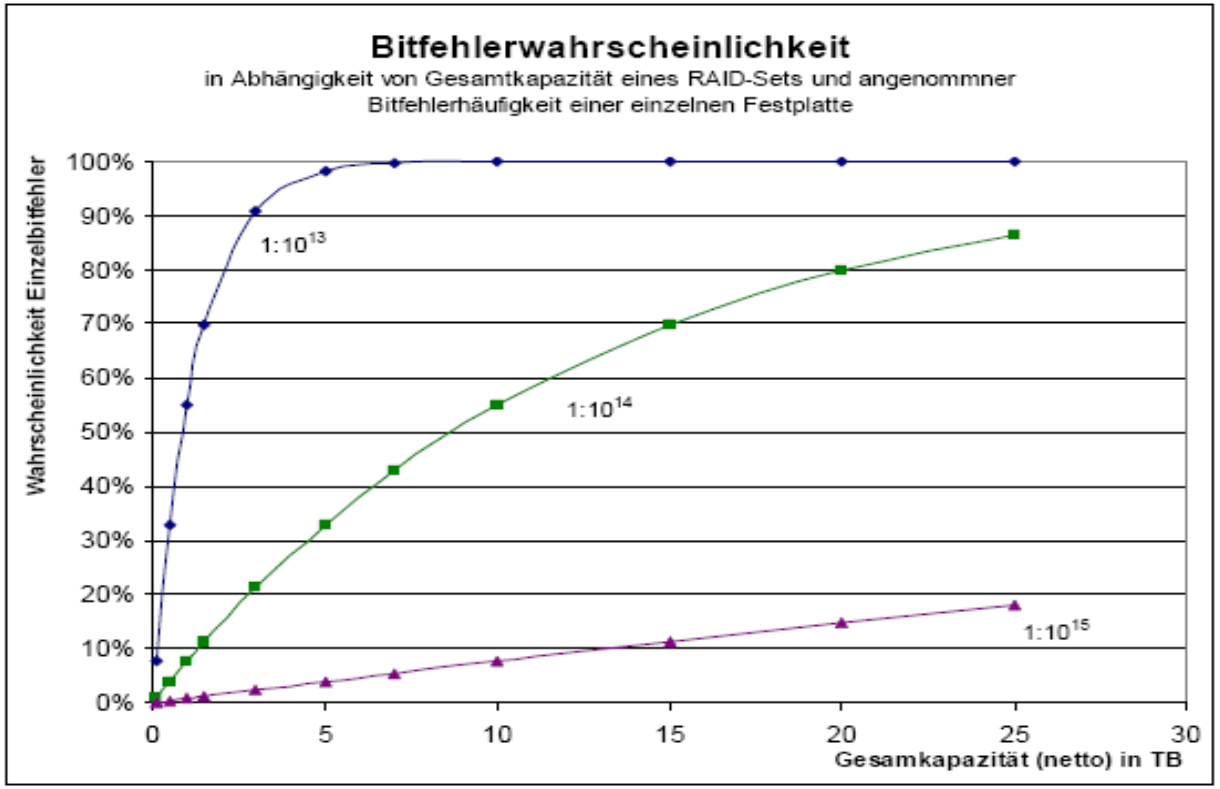


Abbildung 2: Bitfehlerwahrscheinlichkeiten für unterschiedliche Bitfehlerraten

[19]

3. HW-RAID

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

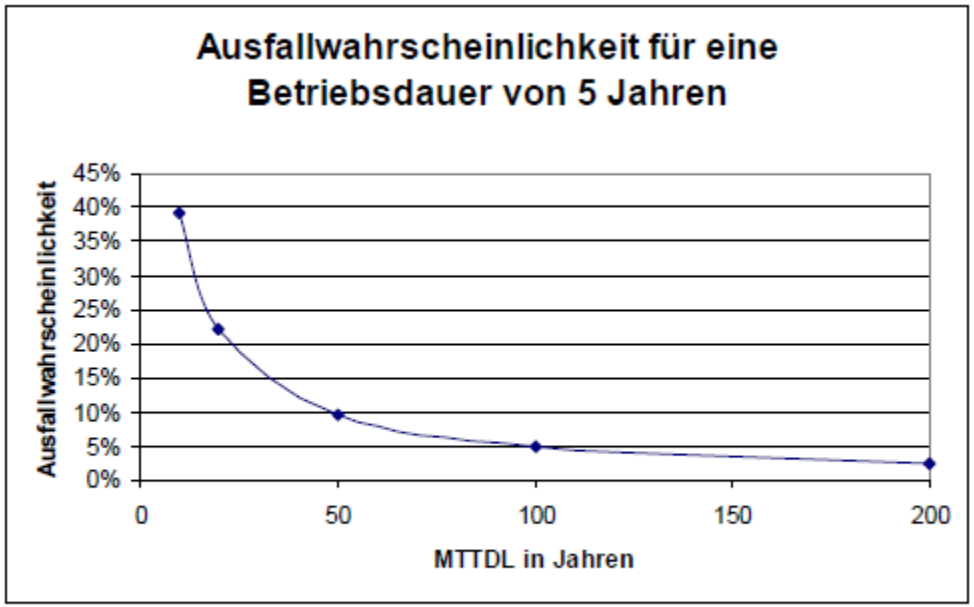
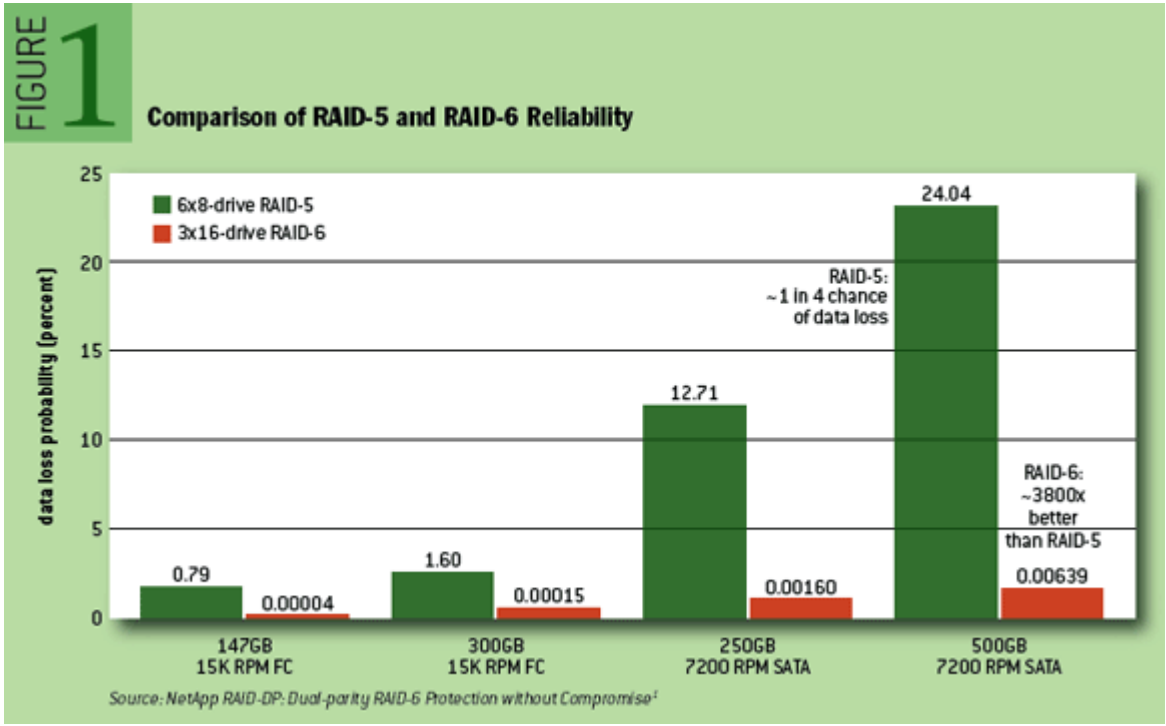


Abbildung 3: MTTDL und Ausfallwahrscheinlichkeit

[19]

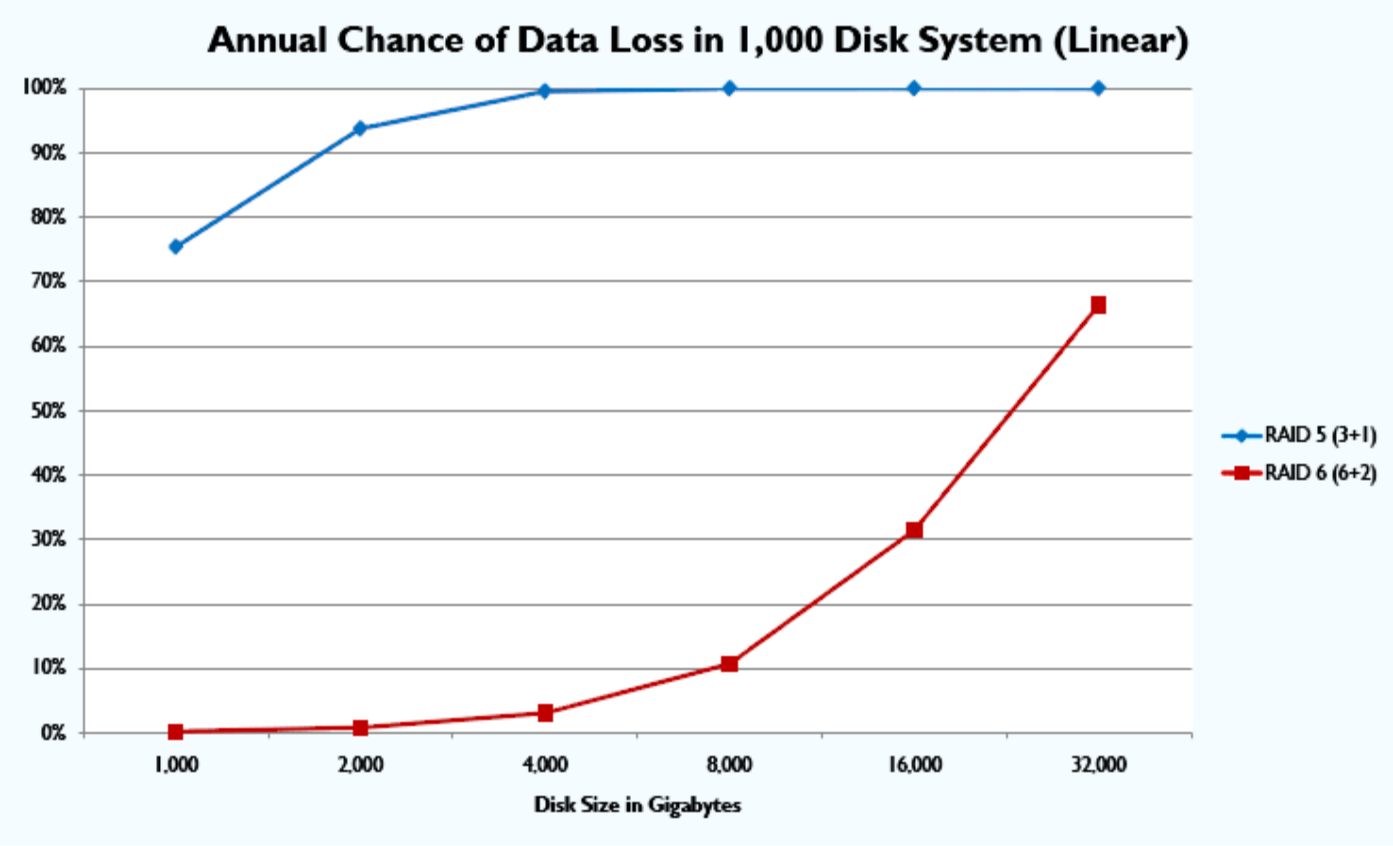
3. HW–RAID

- Vergleich RAID-5 mit RAID-6



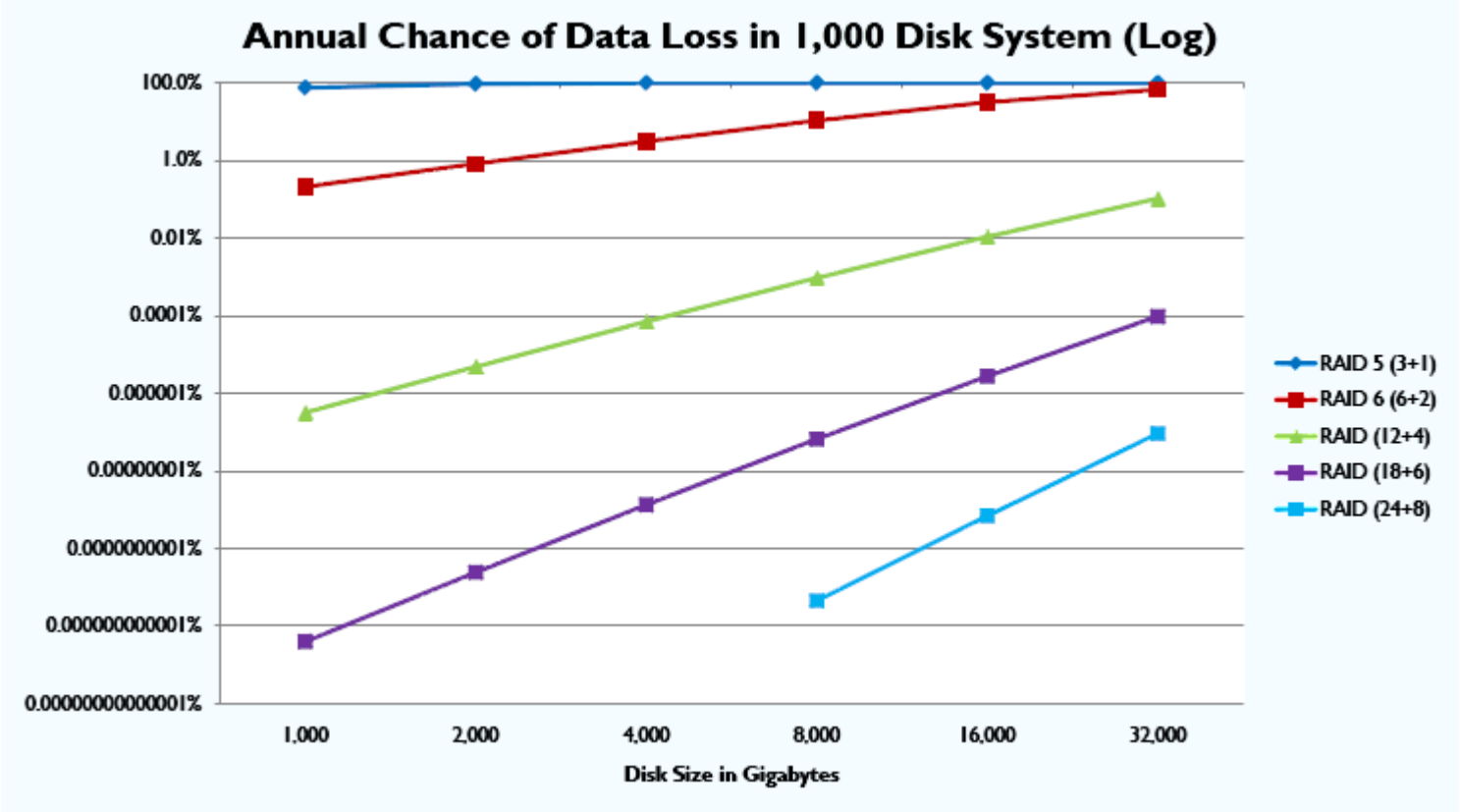
[4]

3. Grenzen von RAID 5 und RAID 6



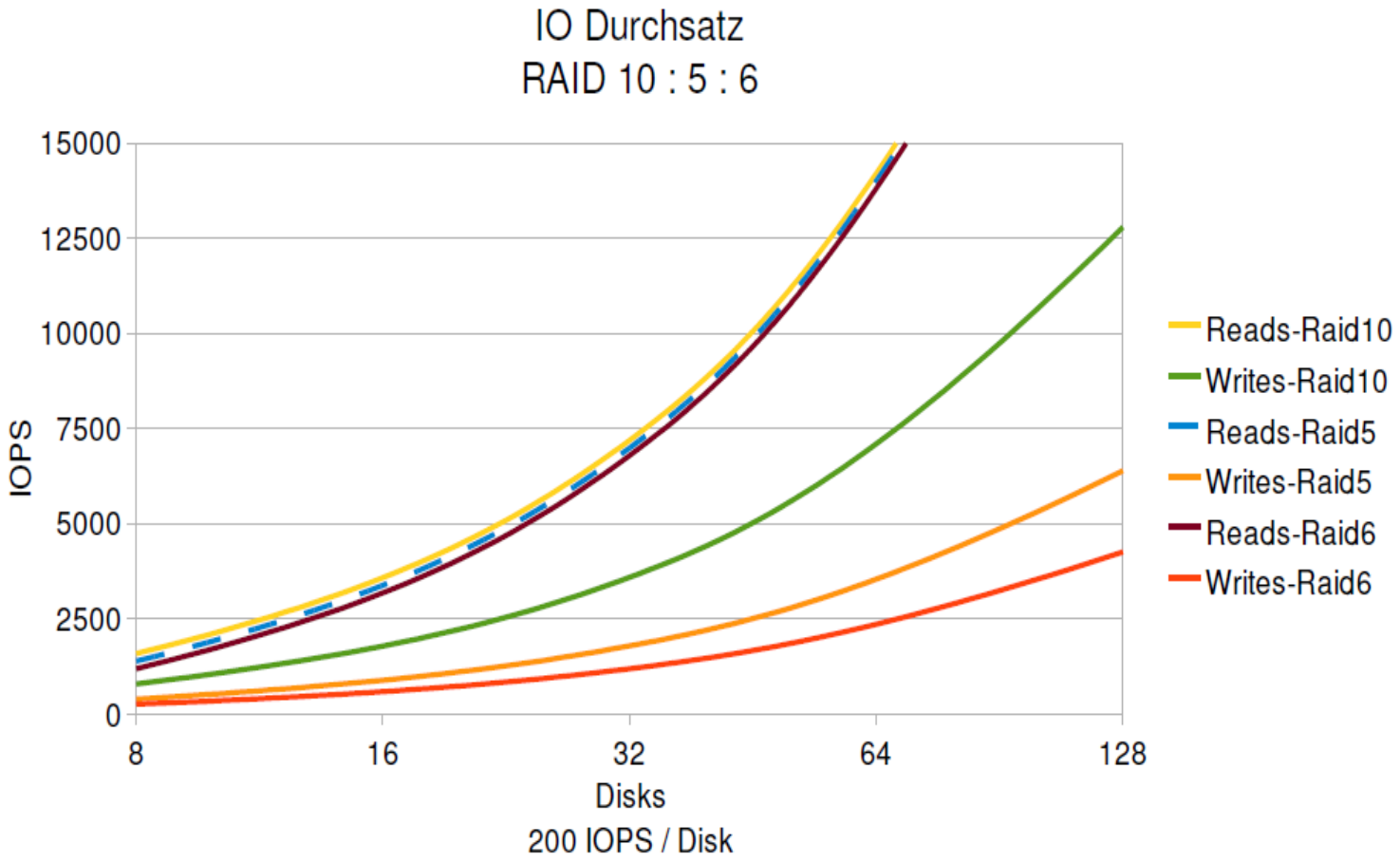
[5]

3. Grenzen von RAID 5 und RAID 6



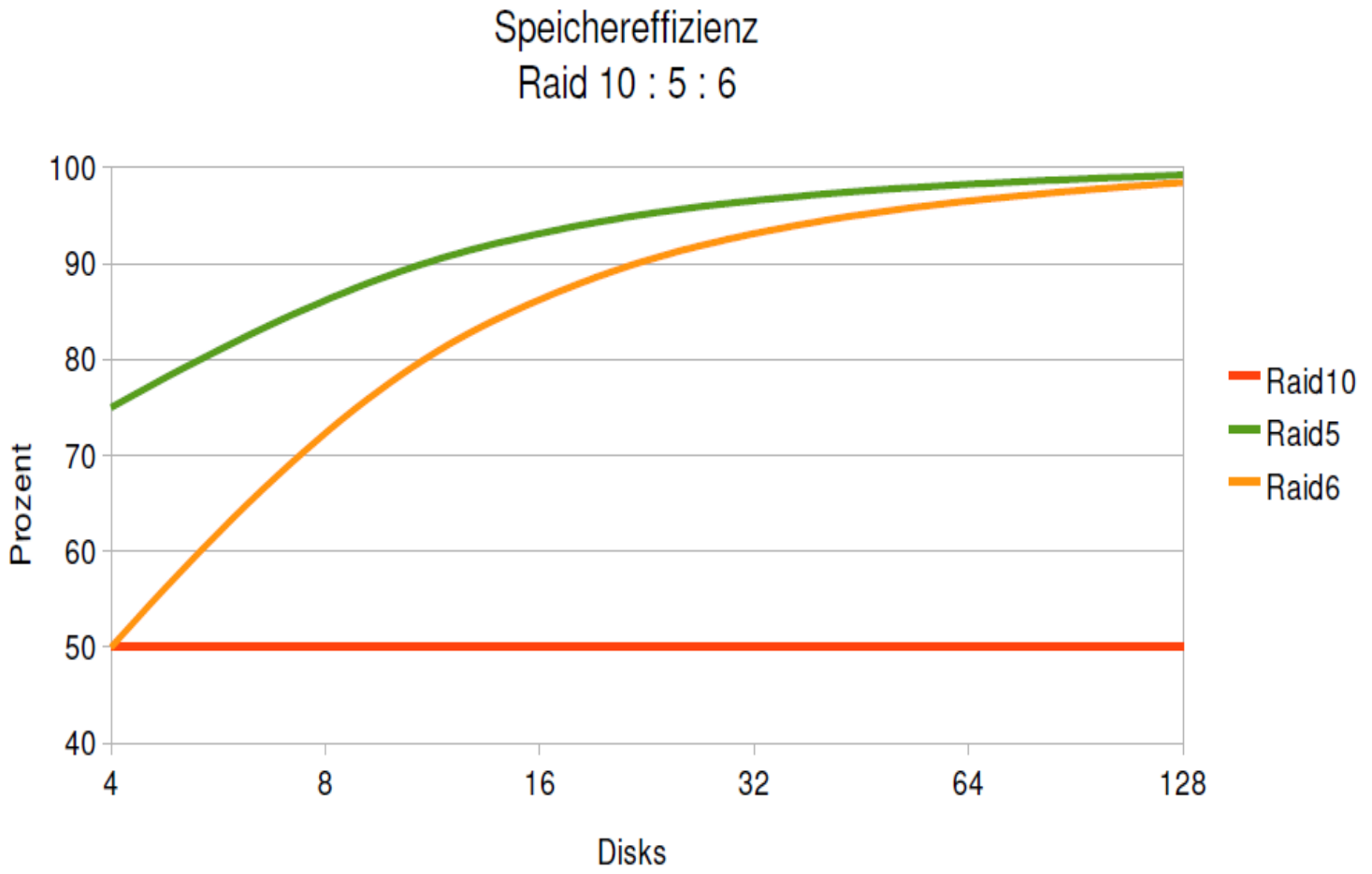
[5]

3. Vergleich RAID 10, 5 und 6



[31]

3. Vergleich RAID 10, 5 und 6



[31]

3. 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

* MTTDL in Jahren

** MTTF 250.000h

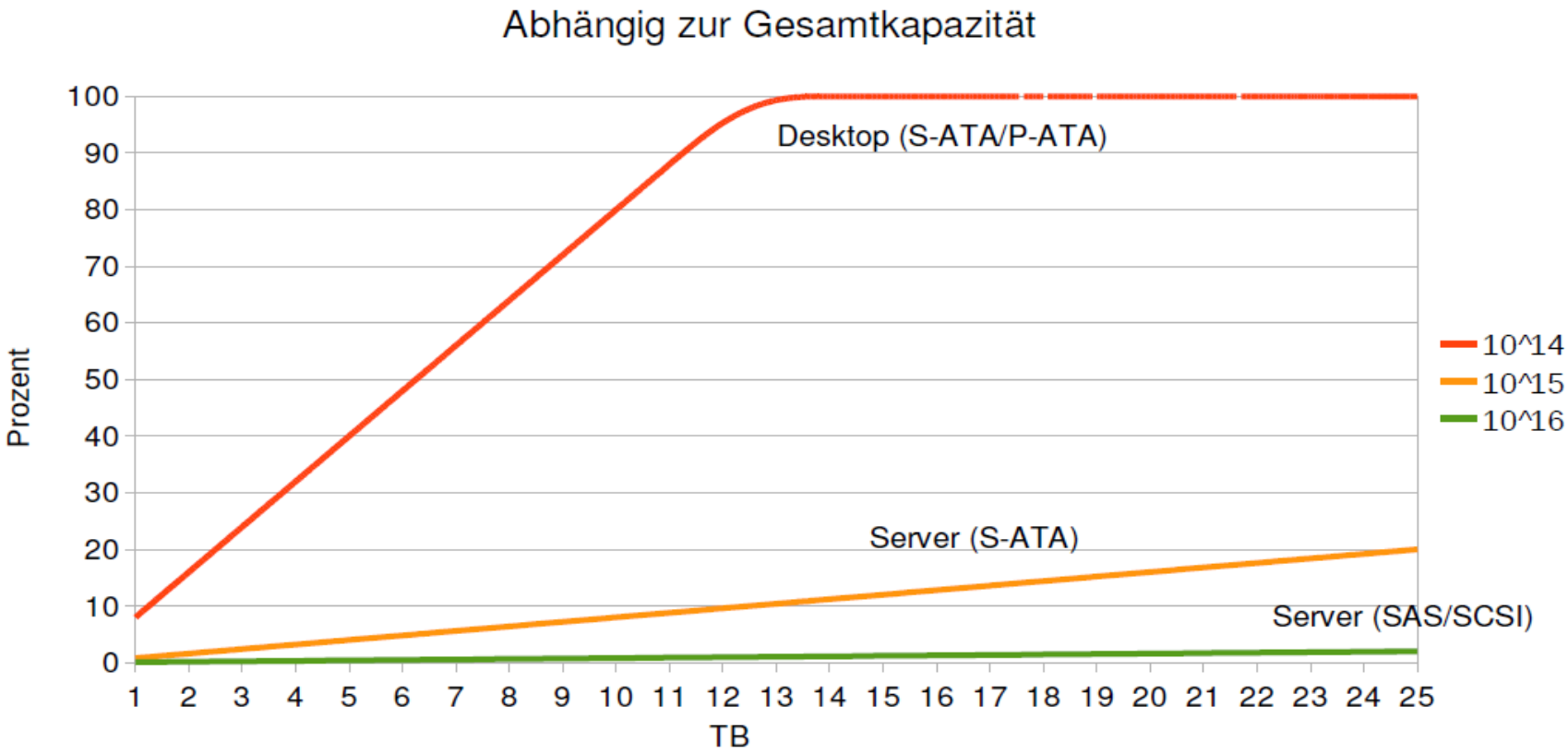
MTTR 24h

[31]

3. RAID 6

Ausfallwahrscheinlichkeit

Bitfehlerwahrscheinlichkeit:



[31]

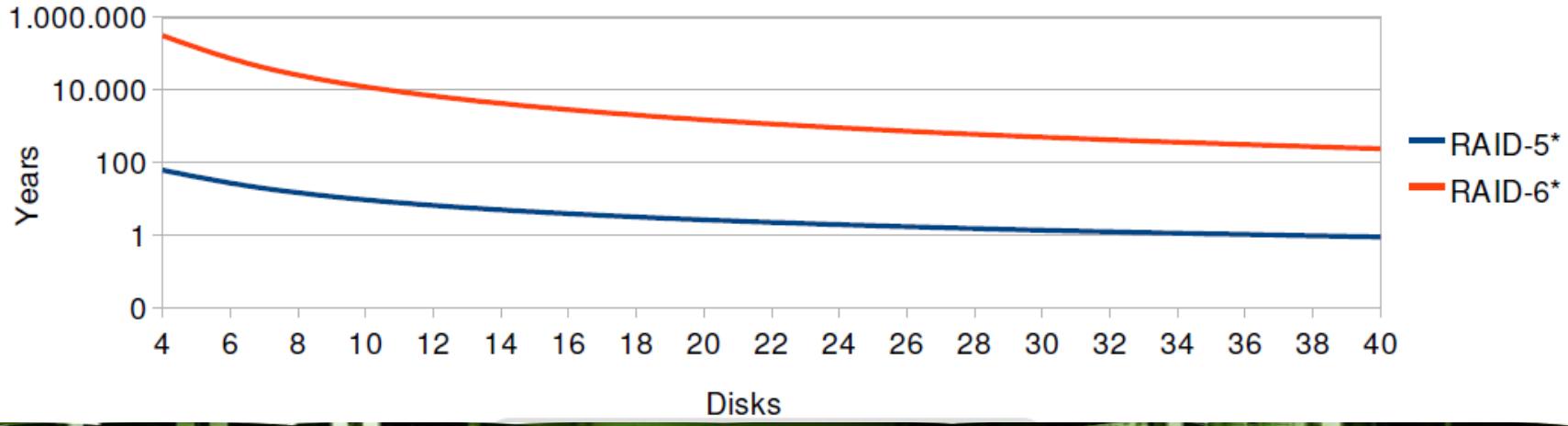
Ausfallwahrscheinlichkeit

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]

3. HW–RAID

- 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.

3. HW–RAID

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		

3. HW–RAID

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

3. HW–RAID

- 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

3. 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 <–> 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

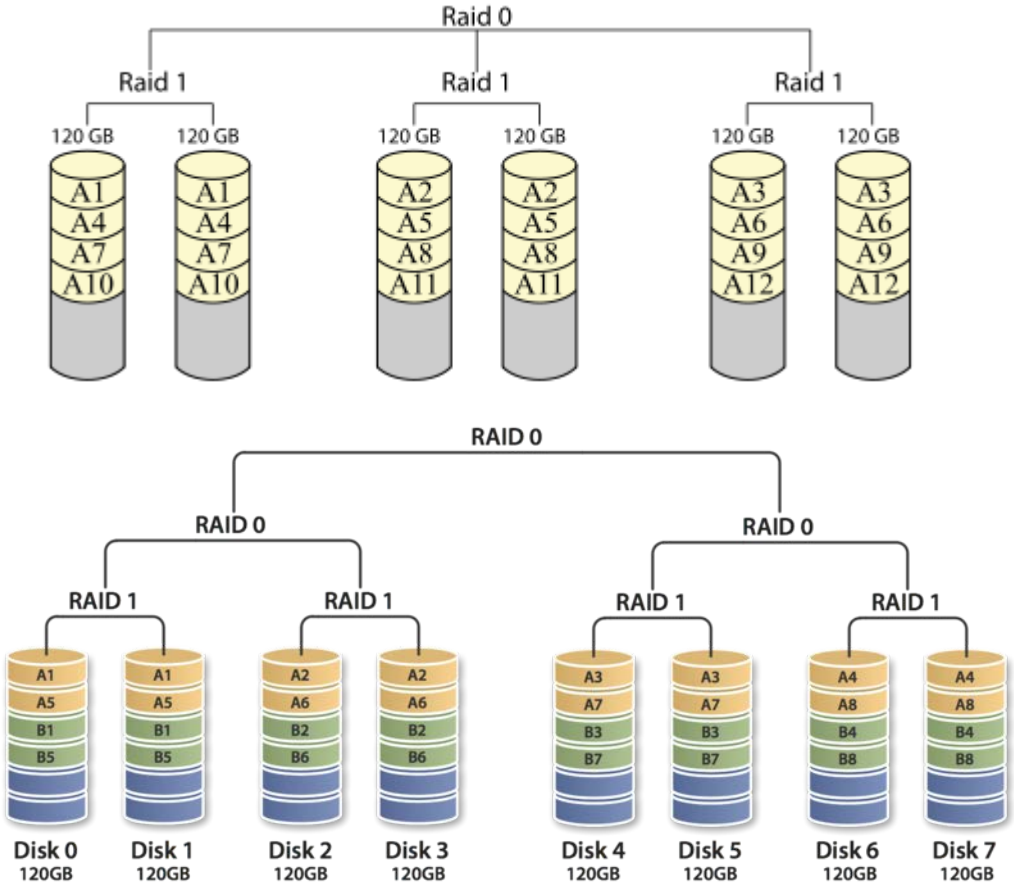
3. HW–RAID

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

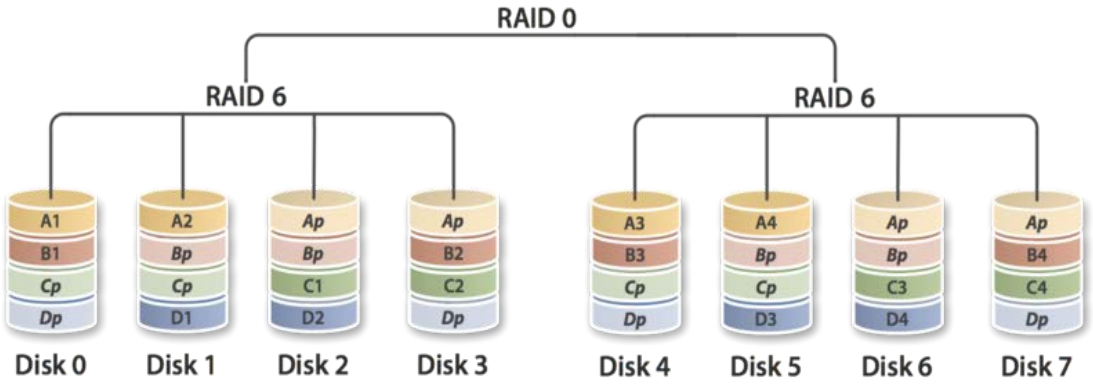
3.1 Verbundsysteme

- 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 SSPiRAL

- 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 Normal-betrieb 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.

3.2 SSPiRAL

- 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

3.2 SSPiRAL

- 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

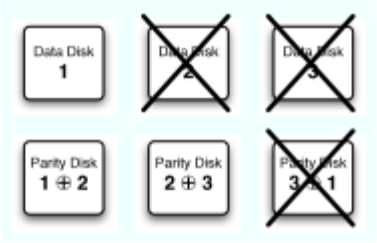


Figure 2: *SSPiRAL data layout and the loss of three nodes.*

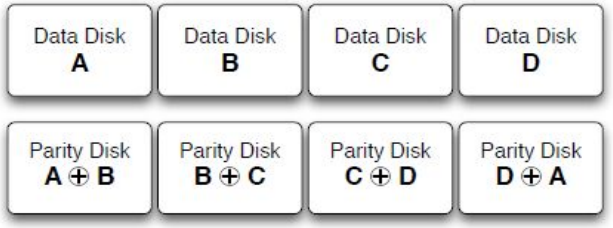


Figure 4. *The 4 + 4 SSPiRAL layout (with x = 2).*

3.2 SSPiRAL

- SSPiRAL
  - Größere Sicherheit durch Paritäts-bildung über mehrere Platten.
    Bei dem Beispiel können im „optimalen“ Fall bis zu vier, auf jeden Fall bis zu drei Datenplatten ohne Datenverlust ausfallen:

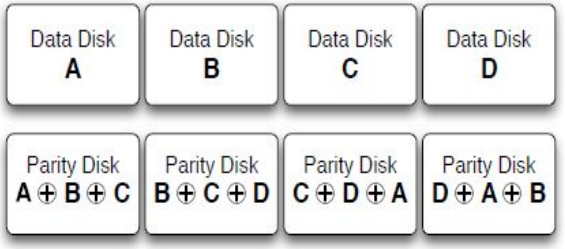


Figure 7. *The 4+4 SSPiRAL layout with eight disks and $x = 3$.*

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

A = (ABC) (CDA) (DAB),
B = (DAB) (ABC) (BCD),
C = (ABC) (BCD) (CDA),
D = (BCD) (CDA) (DAB).

Deshalb tritt kein Datenverlust ein. Bei Ausfall von vier Platten kann Daten-verlust 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 Verläßlichgkeit 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 sequentiellem Zugriff, so ergeben sich 100 IOPS. SCSI- und SAS-Platten liegen bei 200 IOPS.)
  - Datendurchsatz bei sequentiellem 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 DIE-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 (*Streifengranualität*, 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.

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 den 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
  - Betriebsystem
  - 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/

Literatur, Forts.

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.

Literatur, Forts.

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

Literatur, Forts.

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

Literatur, Forts.

27. Wendy Torell, Victor Avelar, Effektive MTBF-Vergleiche in Datencenter-Infrastrukturen, APC White Paper / technische Dokumentation Nr. 112, , http://pdfsuche.info/view/aHR0cDovL3d3dy5tZ2V1cHMuZGUvX3doaXRlcGFwZXJzL2RvY3MvMTEyJTIwLSUyMEVmZmVrdGl2ZSUyME1UQkYtVmVyZ2xlaWNoZSUyMGluJTIwRGF0ZW5jZW50ZXIlMjBJbmZyYXN0cnVrdHVyZW4ucGR
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/