Speichergeräte und -verbünde

Hochleistungs-Ein-/Ausgabe

Michael Kuhn

Wissenschaftliches Rechnen
Fachbereich Informatik
Universität Hamburg

2018-04-06

E/A-Schichten

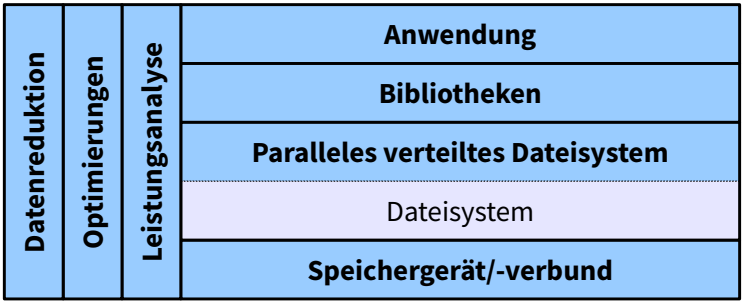


Abbildung: E/A-Schichten und orthogonale Themen

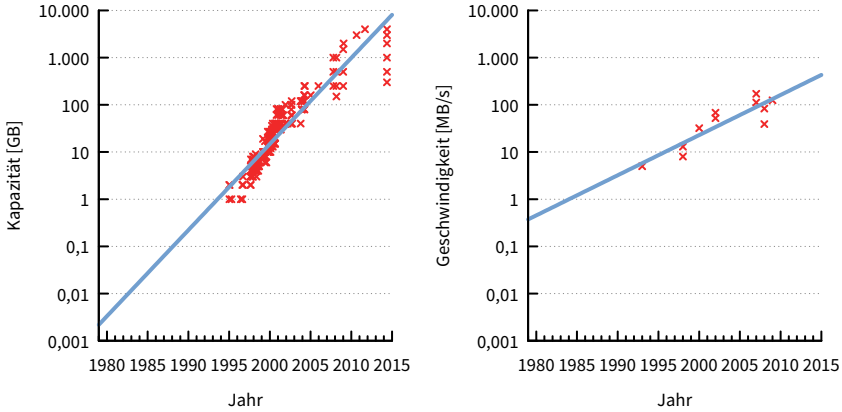
Festplatten

- Die erste Festplatte 1956
  - IBM 350 RAMAC
  - 3,75 MB
  - 8,8 KB/s
  - 1.200 RPM
- Festplattenentwicklung
  - Kapazität: Faktor 100 alle 10 Jahre
  - Geschwindigkeit: Faktor 10 alle 10 Jahre

Festplattenentwicklung [3]

Parameter	Started with	Developed to	Improvement
Capacity (formatted)	3.75 megabytes[9]	eight terabytes	two-million-to-one
Physical volume	68 cubic feet (1.9 m^3)[c][3]	2.1 cubic inches (34 cc)[10]	57,000-to-one
Weight	2,000 pounds (910 kg)[3]	2.2 ounces (62 g)[10]	15,000-to-one
Average access time	about 600 milliseconds[3]	a few milliseconds	about 200-to-one
Price	US$9,200 per megabyte[11][*dubious – discuss*]	< $0.05 per gigabyte by 2013[12]	180-million-to-one
Areal density	2,000 bits per square inch[13]	826 gigabits per square inch in 2014[14]	> 400-million-to-one

Festplatten…

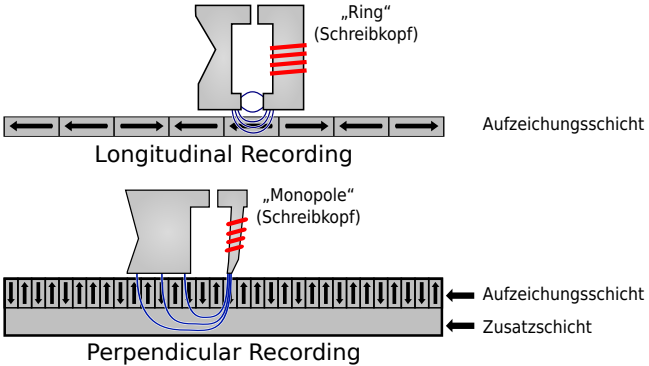


(a) HDD-Kapazitäten (b) HDD-Geschwindigkeiten

Festplatten…

- Physikalische Grundlagen
  - Riesenmagnetowiderstand, magnetischer Tunnelwiderstand
- Magnetische Platter
  - Nicht-magnetisches Material
    - Aluminium-Legierung oder Glas
  - Überzogen mit magnetischer Schicht
    - Dicke im Bereich von Nano- bis Mikrometern
  - Schutzschicht aus Karbon
- Lese-/Schreibkopf
  - Bewegt sich über die Platter
  - Abstand im Nanometerbereich

Festplatten… [2]



- Neu: “heat-assisted magnetic recording” und “shingled magnetic recording”

SSDs

- Festplatten werden zunehmend durch Solid-State-Drives (SSDs) abgelöst
  - Früher: MP3-Player mit kleiner Festplatte
  - Heute: Smartphones mit Flashspeicher
- Vorteile
  - Lesegeschwindigkeit: Faktor 15
  - Schreibgeschwindigkeit: Faktor 10
  - Latenz: Faktor 100
  - Energieverbrauch: Faktor 1–10

SSDs…

- Nachteile
  - Preis: Faktor 10
  - Haltbarkeit: 10.000–100.000 Schreibzyklen
  - Komplexität
    - Unterschiedliche optimale Zugriffsgrößen für Lese- und Schreibzugriffe
    - Adressübersetzung
    - Überhitzung

Technik

- Energieverbrauch
  - Heliumfüllung reduziert Reibungswiderstand
- Kapazität
  - Heliumfüllung erlaubt Verringerung des Abstandes zwischen Plattern (sieben statt fünf Platter)
- Diagnose
  - S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology)
  - Vielzahl von Parametern: Ein-/Ausschaltvorgänge, Parkvorgänge, Anlaufzeit, fehlerhafte Sektoren, Betriebszeit, Temperatur

RAID

- Speicherverbünde für höhere Kapazität, Geschwindigkeit und Zuverlässigkeit
- 1988 an der University of California, Berkeley vorgeschlagen
  - Redundant Array of Inexpensive Disks
  - Heute: Redundant Array of Independent Disks
- Ursprünglich fünf Varianten
  - Genannt RAID 1–5

RAID…

- Kapazität
  - Speicherverbund lässt sich wie eine sehr große Festplatte ansprechen
  - Problem: Organisation/Verteilung der Daten
- Geschwindigkeit
  - Alle Speichergeräte können zur Geschwindigkeit beitragen
  - Problem: Bus muss schnell genug sein
- Zuverlässigkeit
  - Daten können redundant gespeichert werden
  - Problem: Anzahl der Geräte erhöht Ausfallwahrscheinlichkeit

RAID…

- Formen der Redundanz
  - Spiegelung
  - Hamming-Code
  - Parität

RAID 1

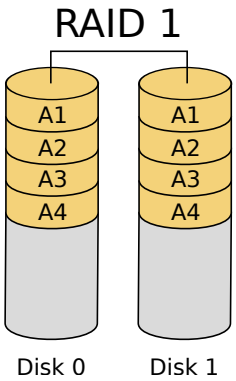


Abbildung: RAID 1 – Spiegelung [4]

RAID 1…

- Erhöhung der Zuverlässigkeit durch Spiegelung
- Vorteile
  - Eine Festplatte kann ausfallen
  - Lesegeschwindigkeit kann erhöht werden
- Nachteile
  - Doppelter Speicherplatzbedarf
  - Doppelte Kosten
  - Schreibgeschwindigkeit entspricht einer einzigen Festplatte

RAID 2

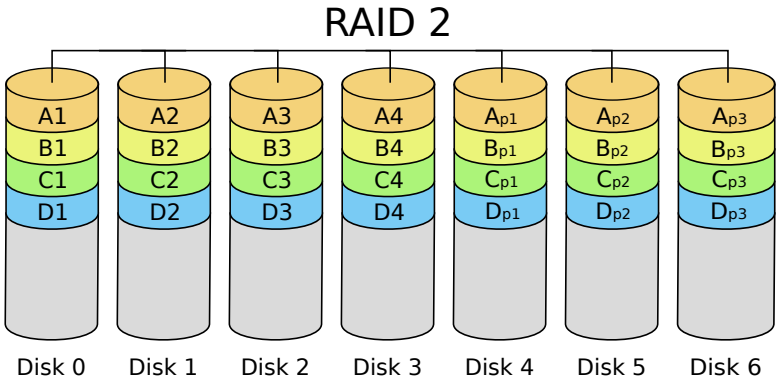


Abbildung: RAID 2 – Bit-Striping mit Hamming-Codes [4]

RAID 2…

- Erhöhung der Zuverlässigkeit durch Hamming-Codes
  - Vier Nutzbits, drei Kontrollbits
- Vorteile
  - Geschwindigkeit kann erhöht werden
- Nachteile
  - Bei jedem Zugriff alle Speichergeräte aktiv
  - Spindeln müssen synchronisiert werden
  - Speicherplatzbedarf fast genauso hoch wie bei RAID 1
- RAID 2 in der Praxis nicht relevant
  - Mehrfachbitfehler kommen kaum vor
  - Festplatten implementieren intern Hamming-Codes

RAID 3

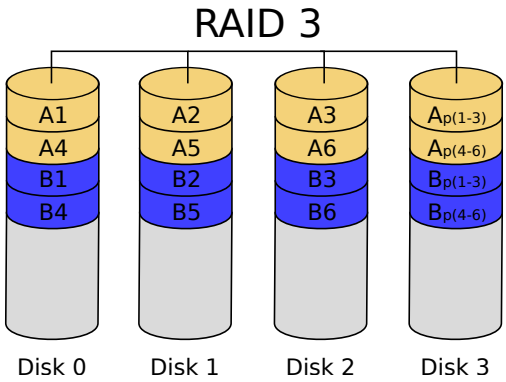


Abbildung: RAID 3 – Byte-Striping [4]

RAID 3…

- Erhöhung der Zuverlässigkeit durch Parität
- Vorteile
  - Geschwindigkeit kann erhöht werden
- Nachteile
  - Bei jedem Zugriff alle Speichergeräte aktiv
  - Spindeln müssen synchronisiert werden

Paritätsberechnung

- Paritätsberechnung z. B. mit Hilfe von XOR ($\oplus$)
  - $A \oplus B = 1 \Leftrightarrow A \neq B$
- Wichtige Eigenschaft
  - $A \oplus B = P \Rightarrow A \oplus P = B$
- Analog für mehrere Werte
  - $A \oplus B \oplus C \oplus D \oplus E = P$

RAID 4

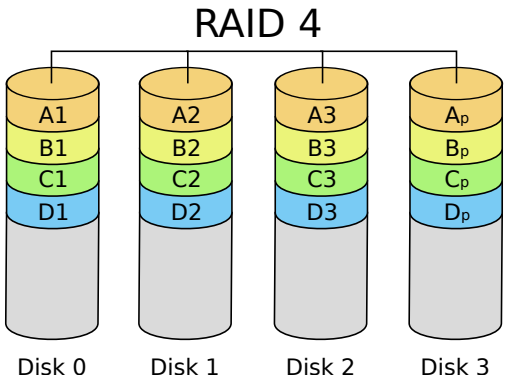


Abbildung: RAID 4 – Block-Striping [4]

RAID 4…

- Erhöhung der Zuverlässigkeit durch Parität
- Vorteile
  - Geschwindigkeit kann erhöht werden
- Nachteile
  - Paritätsfestplatte wird übermäßig beansprucht
  - Schreibgeschwindigkeit durch einzelne Festplatte für Parität beschränkt

RAID 5

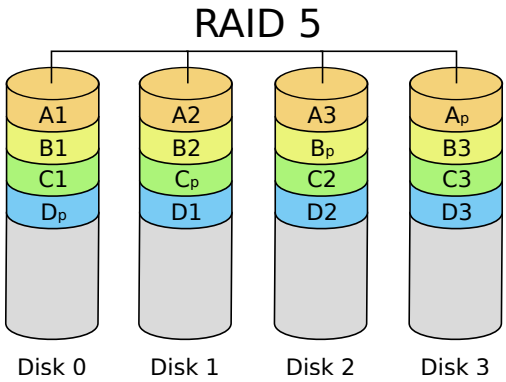


Abbildung: RAID 5 – Block-Striping mit verteilter Parität [4]

RAID 5…

- Erhöhung der Zuverlässigkeit durch Parität
- Vorteile
  - Geschwindigkeit kann erhöht werden
  - Parallele Abarbeitung möglich
  - Schreiblast durch Parität wird auf alle Festplatten verteilt

RAID 0

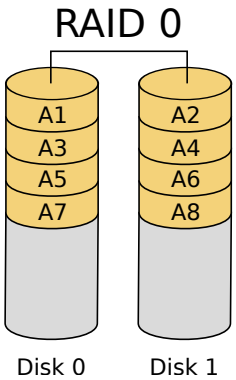


Abbildung: RAID 0 – Striping [4]

RAID 0…

- Erhöhung der Geschwindigkeit durch Striping
- Vorteile
  - Geschwindigkeit kann erhöht werden
  - Mehrere Festplatten können zusammengefasst werden
- Nachteile
  - Keinerlei Redundanz

Wiederherstellung

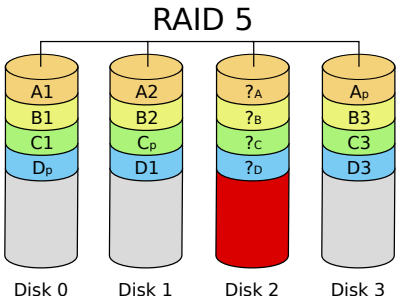


Abbildung: RAID 5 – Wiederherstellung [4]

- $?_A = A1 \oplus A2 \oplus A_p$, $?_B = B1 \oplus B2 \oplus B3, \ldots$

Wiederherstellung…

- Während der Wiederherstellung können weiter Anfragen bearbeitet werden
  - Hot spare: Festplatten sind angeschlossen und werden im Fehlerfall automatisch benutzt
  - Hot swap: Festplatte kann zur Laufzeit gewechselt werden
  - Cold swap: System muss abgeschaltet werden

RAID 6

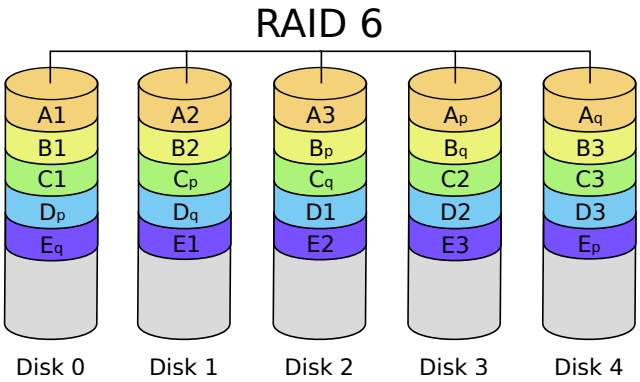


Abbildung: RAID 6 – Block-Striping mit verteilter doppelter Parität [4]

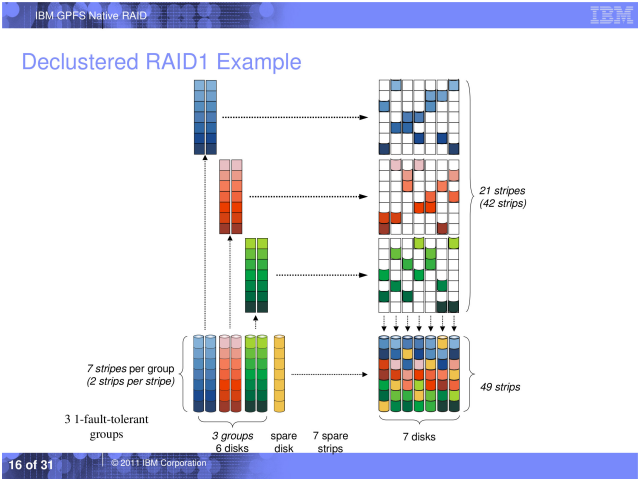
RAID 6…

- Erhöhung der Zuverlässigkeit durch Parität
- Vorteile
  - Ausfallsicherheit wird im Vergleich zu RAID 5 erhöht
- Nachteile
  - Zusätzliche Last durch Paritätsberechnung
  - XOR nicht mehr ausreichend

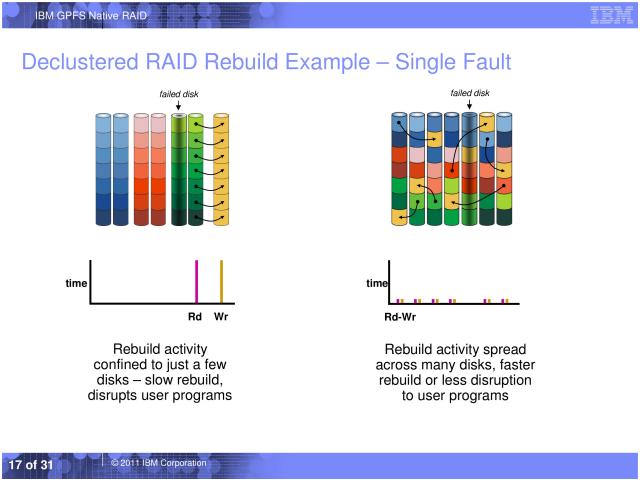
Probleme

- Ausfälle
  - Festplatten sind üblicherweise ähnlich alt
  - Selbe Baureihe
- Wiederherstellung
  - Lesefehler auf anderen Festplatten
  - **Dauer (30 min in 2004, 11 h in 2017)**
- Zuverlässigkeit
  - **Write Hole**

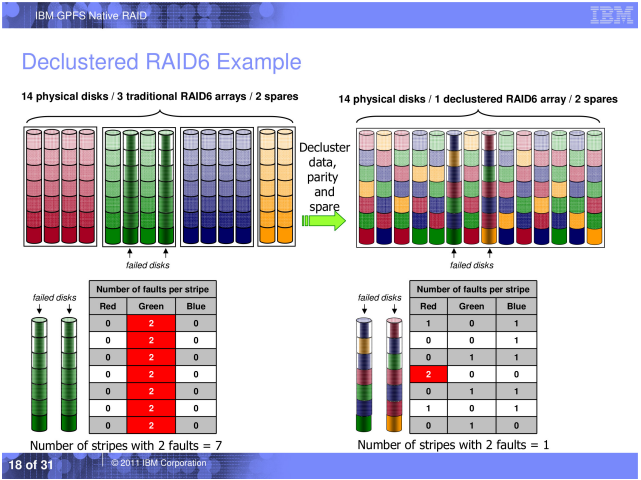
Probleme… [1]


IBM GPFS Native RAID
IBM
Declustered RAID1 Example
21 stripes
(42 strips)
7 stripes per group
(2 strips per stripe)
49 strips
3 1-fault-tolerant
groups
3 groups
6 disks
spare
disk
7 spare
strips
7 disks
16 of 31
© 2011 IBM Corporation

Probleme… [1]


IBM GPFS Native RAID
IBM
Declustered RAID Rebuild Example – Single Fault
failed disk
failed disk
time
Rd
Wr
time
Rd-Wr
Rebuild activity confined to just a few disks – slow rebuild, disrupts user programs
Rebuild activity spread across many disks, faster rebuild or less disruption to user programs
17 of 31
© 2011 IBM Corporation

Probleme… [1]


IBM GPFS Native RAID
IBM
Declustered RAID6 Example
14 physical disks / 3 traditional RAID6 arrays / 2 spares
14 physical disks / 1 declustered RAID6 array / 2 spares
Decluster data, parity and spare
failed disks
failed disks
failed disks
Number of faults per stripe
Red Green Blue
0 2 0
0 2 0
0 2 0
0 2 0
0 2 0
0 2 0
0 2 0
Number of stripes with 2 faults = 7
failed disks
Number of faults per stripe
Red Green Blue
1 0 1
0 0 1
0 1 1
2 0 0
0 1 1
1 0 1
0 1 0
Number of stripes with 2 faults = 1
18 of 31
© 2011 IBM Corporation

Probleme…

- Schreiben in RAID
  1. Alten Block lesen
  2. Alte Parität lesen
  3. Alten Block und Parität XOR-verknüpfen
  4. Neuen Block und Ergebnis XOR-verknüpfen (neue Parität)
  5. **Neuen Block schreiben**
  6. **Neue Parität schreiben**

Probleme…

- Write Hole kann bei mehreren RAID-Leveln auftreten
  - Am populärsten bei RAID 5
- Schreiben des neuen Blocks und der neuen Parität müsste atomar erfolgen
  - Daten und Parität können sonst inkonsistent sein
- Inkonsistenz fällt erst bei Wiederherstellung auf
- Mögliche Lösungsansätze
  - Unterbrechungsfreie Stromversorgung
  - Regelmäßiges Synchronisieren des Arrays

Leistungsbewertung

- Unterschiedliche Leistungskriterien
- Datendurchsatz
  - Große Datenmengen werden sequentiell gelesen oder geschrieben
  - Beispiele: Foto-/Videoverarbeitung, numerische Anwendungen
- Anfragendurchsatz
  - Kleine Datenmengen werden in vielen kleinen Einzelanfragen gelesen oder geschrieben
  - Beispiele: Datenbanken, Metadatenverwaltung

Festplatten und SSDs

- Datendurchsatz
  - Festplatten: 150–200 MB/s
  - SSDs: 500 MB/s
- Anfragendurchsatz
  - Festplatten
    - 75–100 IOPS (7.200 RPM)
    - 175–210 IOPS (15.000 RPM)
  - SSDs
    - 8.600 IOPS (alt)
    - 85.000–90.000 IOPS (aktuell)
- Zugriff auf Teilblöcke/-seiten kann Leistung erheblich reduzieren (Größe üblicherweise 4 KiB)

RAID

- Datendurchsatz
  - Alle Festplatten sollen an einer Anfrage beteiligt sein
  - Gesamtleistung addiert sich
  - Möglichst kleine Blockgrößen
- Anfragendurchsatz
  - Jede Festplatte soll eine Anfrage alleine abarbeiten können
  - Viele Anfragen durch Parallelität
  - Möglichst große Blockgrößen

RAID…

- RAID 0
  - Reines Striping
  - Hoher Daten- und Anfragendurchsatz
  - Anpassung der Blockgröße notwendig
- RAID 1
  - Reine Spiegelung
  - Höhere Leseleistung, geringere Schreibleistung
- RAID 2/3
  - Bit-/Byte-Striping
  - Hoher Datendurchsatz, geringer Anfragendurchsatz

Leistungsbewertung…

- RAID 4
  - Block-Striping
  - Hoher Datendurchsatz
  - Hoher Leseanfragendurchsatz, geringer Schreibanfragendurchsatz
- RAID 5/6
  - Block-Striping
  - Hoher Datendurchsatz, höherer Anfragendurchsatz

Leistungsbewertung…

- Schreiben in RAID
  1. **Alten Block lesen**
  2. **Alte Parität lesen**
  3. Alten Block und Parität XOR-verknüpfen
  4. Neuen Block und Ergebnis XOR-verknüpfen (neue Parität)
  5. **Neuen Block schreiben**
  6. **Neue Parität schreiben**

Leistungsbewertung…

- Beim Lesen nimmt eine Festplate teil
- Beim Schreiben nehmen mindestens zwei Festplatten teil
  - Read-Modify-Write
  - Schlecht für Datendurchsatz
- Verbesserung der Leistung durch Caching
  - Hardware-RAID-Controller haben üblicherweise große dedizierte Caches

Ausblick

- RAID wird auf mehreren Abstraktionsebenen eingesetzt
  - Festplatte, Dateisystem, verteiltes Dateisystem
- Software-RAID
  - Auch direkt ins Dateisystem integriert, z.B. btrfs/ZFS

Zusammenfassung

- Festplattenkapazität und -durchsatz entwickeln sich unterschiedlich schnell
- RAID erhöht Kapazität, Durchsatz und Zuverlässigkeit
- Leistungsbewertung schwierig
  - Daten- vs. Anfragendurchsatz
  - Lese- vs. Schreibvorgänge
- Ursprünglich RAID 1–5
  - Heutzutage auch RAID 0, 6 und Mischformen

Quellen

[1] Veera Deenadhayalan. General Parallel File System (GPFS) Native RAID.
https://www.usenix.org/legacy/events/lisa11/tech/slides/deenadhayalan.pdf.

[2] Wikipedia. Festplattenlaufwerk.
http://de.wikipedia.org/wiki/Festplattenlaufwerk.

[3] Wikipedia. Hard disk drive.
http://en.wikipedia.org/wiki/Hard_disk_drive.

[4] Wikipedia. Standard RAID levels.
http://en.wikipedia.org/wiki/Standard_RAID_levels.