

Speichergeräte und -verbünde

Hochleistungs-Ein-/Ausgabe



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Michael Kuhn

2019-04-02

Wissenschaftliches Rechnen

Fachbereich Informatik

Universität Hamburg

Speichergeräte und -verbünde

Orientierung

Speichergeräte

Speicherverbünde

Leistungsbewertung

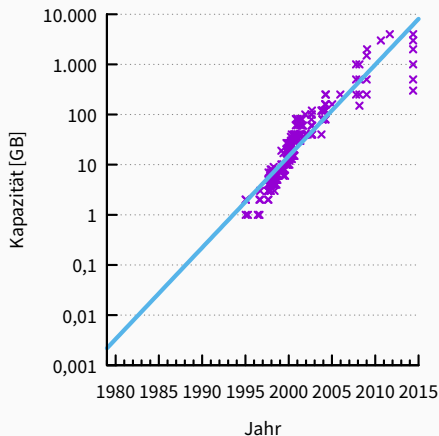
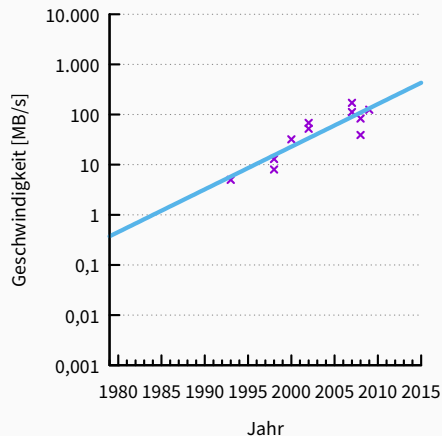
Ausblick und Zusammenfassung

Datenreduktion	Optimierungen	Leistungsanalyse	Anwendung
			Bibliotheken
			Paralleles verteiltes Dateisystem
			Dateisystem
			Speichergerät/-verbund

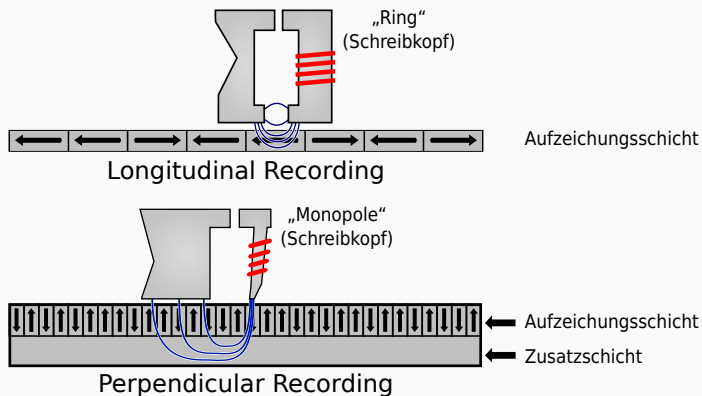
Abbildung 1: E/A-Schichten und orthogonale Themen

- 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

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

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

- 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



- Neu: "heat-assisted magnetic recording" und "shingled magnetic recording"

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

- 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

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

- 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

- 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

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

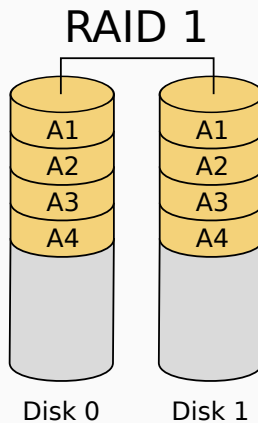


Abbildung 2: RAID 1 – Spiegelung [4]

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

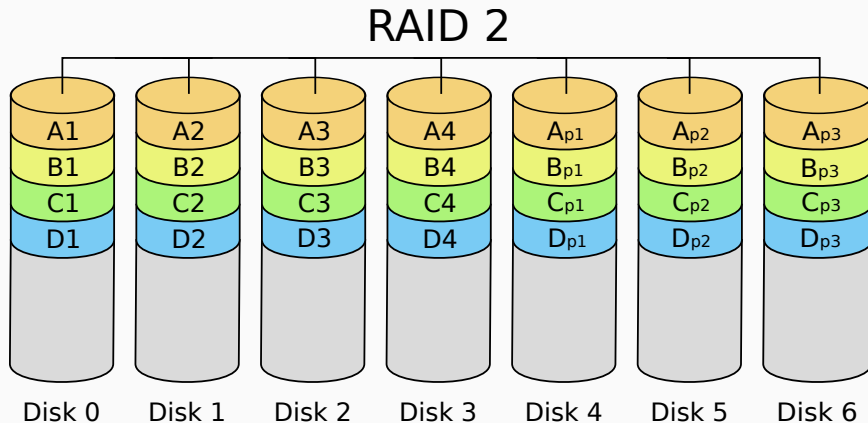


Abbildung 3: RAID 2 – Bit-Striping mit Hamming-Codes [4]

- 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

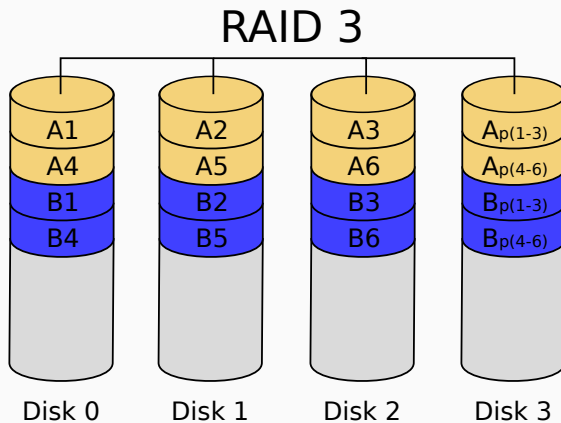


Abbildung 4: RAID 3 – Byte-Striping [4]

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

- 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
 - $A \oplus B \oplus C \oplus D \oplus E = P$

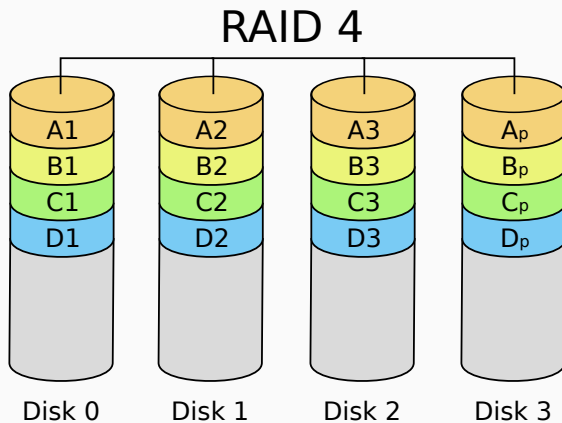


Abbildung 5: RAID 4 – Block-Striping [4]

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

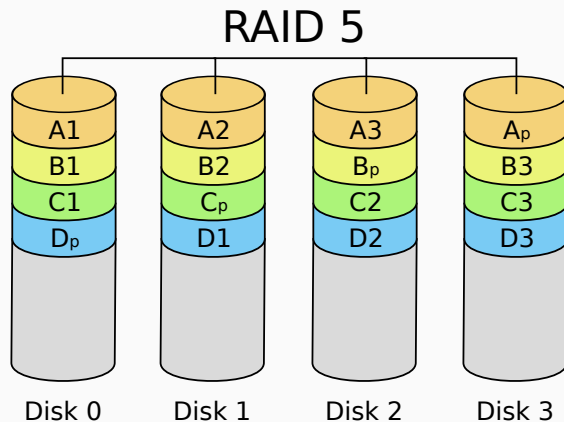


Abbildung 6: RAID 5 – Block-Striping mit verteilter Parität [4]

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

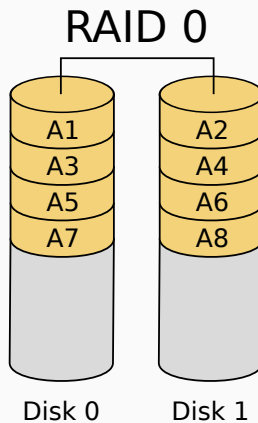


Abbildung 7: RAID 0 – Striping [4]

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

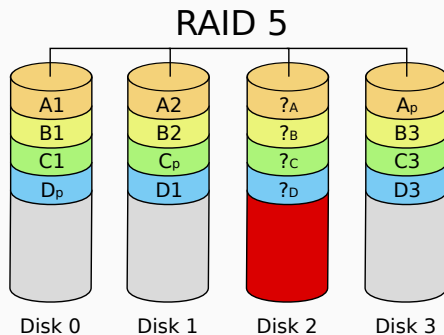


Abbildung 8: RAID 5 – Wiederherstellung [4]

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

- Während der Wiederherstellung können weiter Anfragen bearbeitet werden
 - Hot spare: Festplatte ist angeschlossen und wird im Fehlerfall automatisch benutzt
 - Hot swap: Festplatte kann zur Laufzeit gewechselt werden
 - Cold swap: Festplatte kann nur bei abgeschaltetem System gewechselt werden

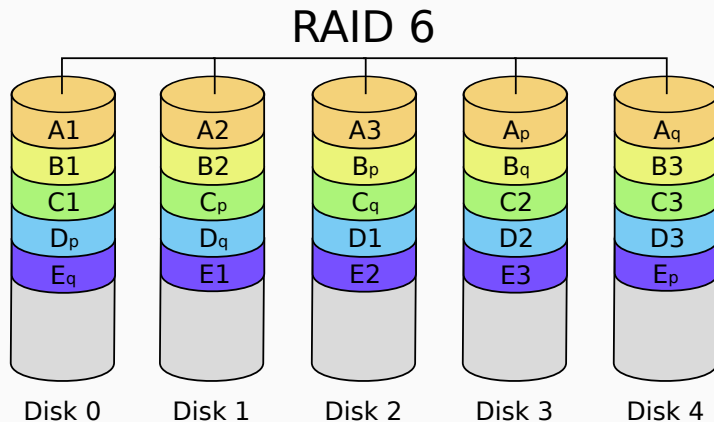
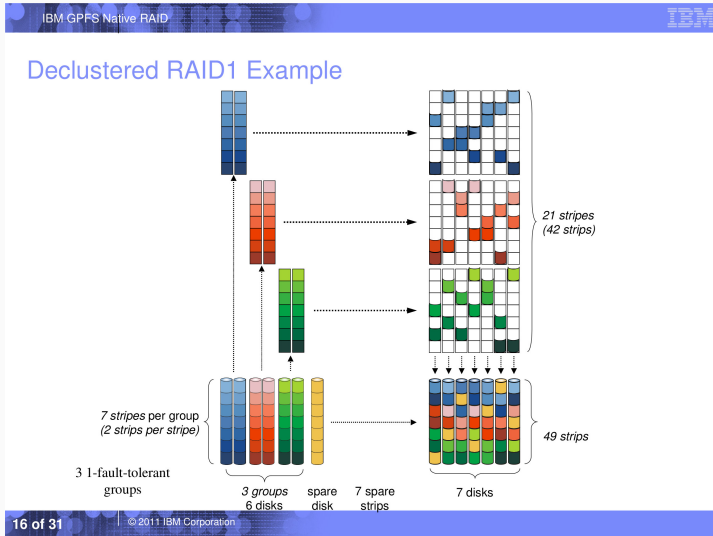


Abbildung 9: RAID 6 – Block-Striping mit verteilter doppelter Parität [4]

- Erhöhung der Zuverlässigkeit durch Parität
- Vorteile
 - Ausfallsicherheit wird im Vergleich zu RAID 5 erhöht
- Nachteile
 - Zusätzliche Last durch Paritätsberechnung
 - Unterschiedliche Implementierungen, teilweise mit erhöhtem Rechenaufwand

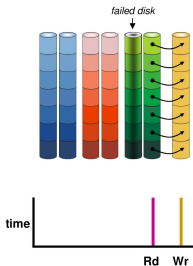
- Ausfälle
 - Festplatten sind üblicherweise ähnlich alt
 - Selbe Baureihe
- Wiederherstellung
 - Lesefehler auf anderen Festplatten
 - **Dauer (30 min in 2004, 13–14 h in 2019)**
- Zuverlässigkeit
 - **Write Hole**



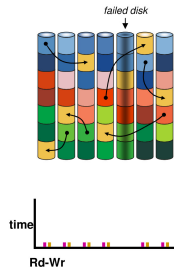
IBM GPFS Native RAID

IBM

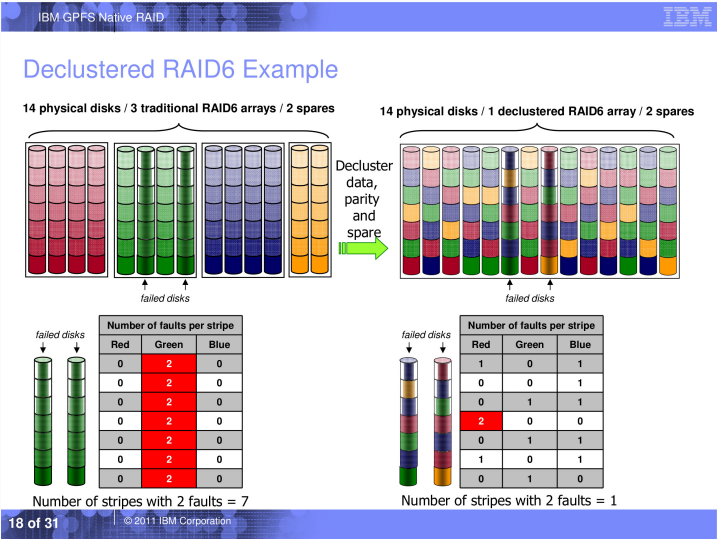
Declustered RAID Rebuild Example – Single Fault



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



- 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**

- 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
 - Es kann nicht bestimmt werden, welche Daten korrekt sind
- Mögliche Lösungsansätze
 - Unterbrechungsfreie Stromversorgung
 - Regelmäßiges Synchronisieren des Arrays

- 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

- Datendurchsatz
 - Festplatten: 150–250 MB/s
 - SSDs: 0,5–3,5 GB/s
- Anfragendurchsatz
 - Festplatten
 - 75–100 IOPS (7.200 RPM)
 - 175–210 IOPS (15.000 RPM)
 - SSDs
 - 90.000–600.000 IOPS
- Zugriff auf Teilblöcke/-seiten kann Leistung erheblich reduzieren
 - Größe üblicherweise 4 KiB

- 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 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

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

- 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**

- Beim Lesen nimmt eine Festplatte 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

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

- 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

- [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.