RAID-Systeme

Hausarbeit

im Proseminar
Speicher- und Dateisysteme
Sommersemester 2012
FB MIN, Universität Hamburg

vorgelegt von

Kai Frederking
Matr.-Nr.: 6322593
Studiengang Informatik, B.Sc.

am 12. September 2012

Kurzfassung

Gegenstand der hier vorgestellten Arbeit ist ein einführender Überblick über die Geschichte, Bedeutung und Funktion von RAID-Systemen. Zielgruppe sind Studienanfänger. Fachliche Vorkenntnisse sind nicht erforderlich, allerdings wird Kenntnis von Computern auf dem "Normalbenutzer-Level" vorausgesetzt.

Der Begriff RAID (**R**edundant **A**rray of **I**nexpensive/Independent **D**isks) bezeichnet die Zusammenschaltung von Laufwerken zur Datenspeicherung zu Verbunden. Deren Zweck ist die Steigerung der Verfügbarkeit und/oder der Performanz gegenüber dem Betrieb der einzelnen Laufwerke.

Es gibt verschiedene Formen der Implementierung von RAID-Systemen, sowohl in Bezug auf die logische Funktion des Lufwerks-Verbundes (sogenannte RAID-Level), als auch bezüglich der technischen Umsetzung. Diese Implementierungen unterscheiden sich in der Priorisierung ihrer Ziele (Verfügbarkeit/Performanz) und im Preis des Gesamtsystems. Abhängig von fallspezifischen Anforderungen sind verschiedene Lösungen optimal.

RAID-Systeme bieten in Hinblick auf Ausfallsicherheit von Festplattensystemen ein exzellentes Preis-Leistungsverhältnis. Zur Leistungssteigerung sind sie meist gut geeignet.

Stichwörter: RAID, MTBF, Festplatten, Speichersysteme, Redundanz, Performanz, Ausfallsicherheit, Verfügbarkeit, Hausarbeit, Proseminar

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CPU	Central Processing Unit (Hauptprozessor)
DRAM	Dynamic Random Access Memory (verwendet als Hauptspeicher)
ECC	Error Correction Code (Code zur Fehlererkennung)
EDV	Elektronische Datenverarbeitung
HDD	Hard Disk Drive (Festplatte)
MTBF	Mean Time Between Failures (mittlere Zeit zwischen Ausfällen)
MTTF	Mean Time To Failure (mittlere Zeit bis zum Ausfall)
RAID	Redundant Array of Independent Disks (redundante Anordniung unabhängiger (Fest)platten) Ursprünglich: Redundant Array of Inexpensive Disks (redundante Anordniung billiger (Fest)platten)
SCSI	Small Computer System Interface (Schnittstelle zur Anbindung schneller Peripheriegeräte)
SLED	Single Large Expensive Disk (einzelne große teure Platte)
SSD	Solid State Drive (ein halbleiterbasierter Massenspeicher)
MB, GB, TB	Mega-, Giga-, Terabyte, hier synonym zu MiB, GiB, TiB genutzt, da die resultierenden Abweichungen bei den Betrachtungen keine Rolle spielen.

Vorwort

Die hier vorgelegte Hausarbeit erweitert und ergänzt die am 18.5.2012 im Rahmen des Proseminars Speicher- und Dateisysteme an der Universität Hamburg gehaltene Präsentation zum gleichen Thema. Die Präsentationsfolien sind unter diesem Link verfügbar:
http://wr.informatik.uni-hamburg.de/_media/teaching/sommersemester_2012/sds-12-frederking-raid_systeme-praesentation.pdf

1 Ziele

Ziel dieser Arbeit ist es, Studienanfängern im Bereich Informatik und anderen Interessierten einen grundlegenden Überblick über RAID-Systeme, deren Ursprung, Zweck und Nutzen zu geben. Sie versteht sich als Einführung in das Thema. Sowohl die technischen als auch die historischen Inhalte sind teils stark vereinfacht.

Spezielle Vorkenntnisse sind zum Verständnis des Dargestellten nicht erforderlich, allerdings ist es vorteilhaft, wenn sich der Leser in groben Zügen mit der Funktion von Computern und Massenspeichern auskennt. Mathematik der Sekundarstufe hilft beim Verständnis der Ausfallwahrscheinlichkeiten redundanter Systeme.

2 Geschichte

2.1 Historische Situation

In den 60er Jahren des letzten Jahrhunderts wurde der Computermarkt von Großrechnern dominiert. Die Zahl der Installationen war aus heutiger Sicht gering, die Preise für Rechner und Peripheriegeräte sehr hoch. In den 70er Jahren kamen sogenannte Minirechner auf, Maschinen geringerer Leistungsfähigkeit bei deutlich geringerem Preis. Sowohl Groß- als auch Minirechner nutzten oft die gleiche Peripherie, teils auch als Massenspeicher Plattensysteme.

Ein(!) Modul einer IBM 3380 Platteneinheit (1987).

Technische Daten:
Kapazität 1 GB, Zugriffszeit 20 ms, Transferrate 3MB/s, Preis ca. $60.000, Netzteil 6.600Watt

Quelle:
http://de.wikipedia.org/wiki/Datei:IBM3380DiskDriveModule.agr.jpg

Abbildung 1: "SLED" Plattenlaufwerk

Mit dem Aufkommen von PC Anfang der 80er Jahre, vorwiegend aufgrund der Markteinführung des IBM PC, wurden zunehmend billige und relativ leistungsfähige kleine Festplattenlaufwerke verfügbar.

"Kleine" Festplatte Maxtor XT-1140, 5.25", Ende der 80er (Größenvergleich: 2.5" Platte mit 6 GB)

Technische Daten:
Kapazität ca. 100MB, Zugriffszeit 35ms, Transferrate 1MB/s, Preis ca. $1.500, Netzteil 10 Watt

Quelle:
http://en.wikipedia.org/wiki/File:5.25_inch_MFM_hard_disk_drive.JPG

Abbildung 2: "Inexpensive Disk"

2.2 Das Problem

Die teuren Großrechner-Festplatten, rückblickend auch SLED genannt - Single Large Expensive Disks (einzelne große teure Platten) - waren zwar schneller und zuverlässiger als ihre kleinen Geschwister, allerdings nicht um so viel, wie sie teurer waren.

Grundsätzlich stellten Festplatten eines der größten Ausfall- und Datenverlust-Risiken in der EDV (elektronische Datenverarbeitung, damals gebräuchlicher als "IT") dar. Insbesondere der Headcrash, auch "spanabhebende Datenverarbeitung" genannt, das Aufsetzen eines Schreib-Lese-Kopfes auf der Oberfläche einer rotierenden Datenplatte mit gegenseitiger Zerstörung, war gefürchtet.

Head Crash bei moderner Platte

Abbildung 3: Head Crash

Außerdem begrenzte die Geschwindigkeit der Festplatten zunehmend die Leistungsfähigkeit der Gesamtsysteme. Die Transfer- oder Datenrate lag um mehr als eine Größenordnung unter dem, was damaliger Hauptspeicher (DRAM) zu leisten vermochte. Wichtiger noch: Die Zugriffszeit, also die Zeit die erforderlich ist, den Schreib-Lesekopf zu positionieren (Spurwechsel) und anschließend im Mittel eine halbe Plattenumdrehung zu warten, bis die gewünschten Daten unter dem Kopf angekommen sind (Latenz), verringerte sich kaum mit fortschreitender Entwicklung der Festplatten, während alle anderen Systemkomponenten exponentiell mit der Zeit schneller wurden. [Dies gilt übrigens heute noch und ist einer der Gründe für den Erfolg der SSD gegenüber modernen Festplatten.]

2.3 Der Lösungsansatz

Schon 1978 patentiert Ken Ouchi bei IBM ein "System for recovering data stored in failed memory unit.“, das dem im Folgenden vorgestellten RAID 5 sehr ähnelte. Er ist allerdings seiner Zeit - und insbesondere dem Markt - voraus und findet deswegen kaum Beachtung. Erst 1980 entwickelt Seagate die erste 5.25“ Festplatte, der IBM PC folgt 1981 und 1986 wird mit dem SCSI Interface eine geeignete Bus-Schnittstelle für kleine und mittlere Systeme standardisiert.

Vor diesem Hintergrund veröffentlichten David A. Patterson, Garth A. Gibson und Randy H. Katz, von der University of California, Berkeley 1987 ihren wegweisenden Vorschlag „A Case for Redundant Arrays of Inexpensive Disks“ = „Ein Argument für redundante Anordnungen billiger Platten“.

Hierin schlugen sie vor, kleine, billige Platten zu Verbunden zusammenzuschließen. Diese Verbunde können dann bezüglich ihrer Verfügbarkeit und Performanz wesentlich teureren einzelnen Laufwerken (SLED) deutlich überlegen sein.

Anmerkung: Das Akronym RAID wurde später aus Marketinggründen uminterpretiert. Aus "inexpensive" (billig) wurde "independent" (unabhängig), was inhaltlich vollständiger Unsinn ist, da die Platten im Verbund zwar "individuell" (also nicht notwendigerweise in einem Gehäuse), aber auch logisch komplett abhängig vom Rest des Verbundes sind.

Im folgenden schauen wir uns an, wie diese Vorteile zu Stande kommen. Zuerst werfen wir einen Blick auf den Sicherheitsgewinn.

2.3.1 Redundanz als Sicherheitsfaktor

Werden Daten nicht nur einmal, sondern auf zwei oder mehr Festplatten vorgehalten, so kann beim (rechtzeitig erkannten) Verlust eines einzelnen Datenträgers durch Defekt der Betrieb ohne Datenverlust fortgesetzt werden.

Zuerst werfen wir einen Blick auf ein Maß für Verfügbarkeit, bevor wir dieses nutzen, um die verschiedenen System-Varianten zu bewerten.

2.3.1.1 Ausfallrate und MTBF

Ausfallraten (NICHT: Fehlerraten) von technischen Geräten und Anlagen werden oft als Kehrwert ihrer MTBF angegeben, der „Mean Time Between Failures“ (mittlere Zeit zwischen Ausfällen). Dieses Zeitintervall ist nach IEC 60050 (191) definiert als:

"*Der Erwartungswert der Betriebsdauer zwischen zwei aufeinanderfolgenden Ausfällen."*

Kennen wir die MTBF eines Geräts, dann können wir die Wahrscheinlichkeit p(T) eines Ausfalls des Gerätes in einem Zeitintervall der Länge T nach Inbetriebnahme errechnen als:

$$p(T) = 1 - e^{-T / MTBF}$$

Bei Festplatten ist der erste meist auch der letzte Ausfall, daher spricht man hier auch von "Mean Time To Failure" (MTTF, mittlere Zeit bis zum Ausfall) oder man betrachtet den Austausch durch ein gleichartiges Laufwerk als Wiederaufnahme des Betriebs. Für die Rechnung spielt dies keine Rolle.
Die MTBF für Festplatten beträgt typischerweise um 10^6 Stunden, also in der Größenordnung 100 Jahre. Die Ausfallwahrscheinlichkeit innerhalb eines Jahres ist also näherungsweise $1 - e^{-1 / 100} \approx 1\%$

Zu beachten ist im Einzelfall, dass MTBF vom Hersteller des Geräts unter Laborbedingungen ermittelt werden und damit "best case" Szenarien repräsentieren. In der Praxis können die Ausfallraten teils erheblich höher ausfallen, abhängig von Betriebstemperatur, Last, mechanischen Einflüssen, Vibrationen und anderen Außeneinflüssen.

2.3.1.2 Datenspiegelung und Ausfallwahrscheinlichkeit

Um die Vorteile redundanter Datenhaltung zu sehen reicht es, uns den einfachsten Fall vor Augen zu führen: Die Datenspiegelung auf zwei gleichen Platten, wobei für alle Daten "Platte 2 = Platte 1" gilt.

Ein Datenverlust bei Plattenschaden tritt nur dann auf, wenn beide Platten gleichzeitig ausfallen, also wenn bei einem Ausfall einer der beiden Platten die zweite Platte vor Restaurierung des ersten Opfers ebenfalls stirbt.

Nennen wir das Intervall, in dem wir einen Defekt bemerken, die Platte austauschen und die Spiegelung mit dem verbliebenen Laufwerk wieder herstellen können Δt.

$p(\Delta t)$ sei die Wahrscheinlichkeit, dass in diesem Zeitintervall eine Platte ausfällt. Nehmen wir für folgendes Beispiel ein Δt von 10h und ein $p(\Delta t)$ von 1/50.000 an.

Läuft das Speichersystem 10.000h (etwas mehr als ein Jahr), dann fällt eine ungespiegelte Platte mit der Wahrscheinlichkeit
$1 - (1 - p)^{(10000 / \Delta t)} = 1 - (49999 / 50000)^{1000} \approx 0{,}02 =$ **2%**
aus,
ein gespiegeltes Paar, bei dem beide innerhalb von 10h ausfallen müssen, mit
$1 - (1 - p^2)^{(10000 / \Delta t)} \approx 0{,}0000004 =$ **0,00004 %**

Dies ist ein sensationeller Verfügbarkeitsgewinn, insbesondere wenn man dies auf einen Serverpark mit hunderten von Laufwerken extrapoliert.

In einem komplizierteren Fall, in dem wir die Daten nicht spiegeln, sonder statt auf vier auf fünf Platten verteilen (mehr hierzu im Folgenden), tritt Datenverlust schon beim

Ausfall von zwei der fünf Platten auf. Die Wahrscheinlichkeit hierfür ist $\binom{5}{2}$ = 10 mal höher als der Ausfall zweier von zwei Platten, also 0,0004% in obigem Beispiel. Immer noch dramatisch besser als die nicht-redundanten 2%.

2.3.1.3 Seiteneffekte der Datenspiegelung

Da für die redundante Datenhaltung zusätzliche Daten zu schreiben sind, erhöht sich natürlich der Zeitaufwand für den Transport der Daten. Da aber der Transport über den Datenbus erheblich schneller ist als das Schreiben der Daten auf die Platten selbst, da auf die Platten gleichzeitig und nebenläufig zugegriffen werden kann und da der Datentransfer üblicherweise über einen dedizierten Controller gesteuert wird, fällt dieser Overhead nicht ins Gewicht.

Eine größere Rolle für die Verfügbarkeit des Gesamtsystems spielt die Ausfallwahrscheinlichkeit der an der Datenspiegelung beteiligten Komponenten. Diese addiert sich zum (stark reduzierten) Ausfallrisiko des Laufwerksverbundes. Glücklicherweise sind die Ausfallraten von Halbleiterkomponenten um mehrere Größenordnungen geringer als die von (mechanischen) Festplatten, so dass sie den Sicherheitsgewinn nur leicht schmälern. Allerdings ist zu berücksichtigen, dass RAID Controller einen sogenannten potentiellen "Single Point of Failure" darstellen, mit Konsequenzen eines Ausfalls, die deutlich über die eines einzelnen Laufwerksversagens hinausgehen. Dies spielt eine Rolle bei der Wahl geeigneter Komponenten und auch bei Grenznutzenbetrachtungen bei mehrfach-redundanter Datenhaltung. Mehr dazu später.

2.3.2 RAID als Performanz-Erhöher

Wenn wir bei Massenspeichern über Leistungsfähigkeit sprechen, dann meinen wir vorwiegend drei Dinge:

Kapazität, Zugriffsgeschwindigkeit und Übertragungsgeschwindigkeit.

Möglichkeiten zur Partitionierung oder zur Zusammenfassung zu logischen Laufwerken seien hier unter "Kapazität" subsummiert, Mount-Zeiten unter "Zugriff".

Schalten wir mehrere kleinere Platten zusammen, so ergibt sich als Kapazität des Verbundes (bestenfalls) die Summe der einzelnen Kapazitäten abzüglich der Redundanz, also der nur für Datenkopien oder Prüfsummen reservierten Bereiche. Kapazitätsgewinn und größere logische Laufwerke sind zwar oft ein Nebeneffekt von RAID Systemen, allerdings meist nicht ihr Hauptzweck und auch mit anderen Mitteln zu erreichen.

Die Zugriffsgeschwindigkeit kann dann gesteigert werden, wenn Daten redundant auf unterschiedlichen Platten abgelegt werden und das System beim Lesen gleichzeitig auf

alle Kopien zugreift. Hierbei verringert sich zumindest die Latenz, da die zuerst verfügbaren Daten genutzt werden können.

Die Übertragungsgeschwindigkeit erhöht sich dann, wenn Daten die zu einem Zugriff gehören auf verschiedene Platten verteilt sind und gleichzeitig Übertragen werden können. Grenzen setzt hier die Busgeschwindigkeit, die jedoch üblicherweise weit über der Geschwindigkeit der einzelnen Platten liegt. Theoretisch lässt sich die Übertragung hinreichend großer Datenmengen durch Verteilung auf mehr Platten beliebig beschleunigen, wobei jedoch Zugriffszeiten und Buskapazitäten praktische Grenzen setzen.

3 Implementationen

3.1 Komponenten eines RAID-Systems

Ein RAID-System besteht aus den zu verbindenden Laufwerken, den zur Verbindung genutzten Komponenten und der Software, welche die Ansteuerung des Verbunds übernimmt.

Im Folgenden sei davon ausgegangen, dass es sich bei den Laufwerken um Festplatten handelt. Dies ist das weit überwiegende Einsatzgebiet für RAID.

Nicht näher eingehen will ich in diesem Rahmen auf die Anbindung an den Datenbus (SCSI, SATA, ...), auch wenn sie genau genommen Bestandteil des RAID (allerdings nicht für dieses spezifisch) ist.

3.1.1 Die Laufwerke

Die Laufwerke eines RAID-Verbundes sollten alle von gleicher Kapazität sein. Dies ist praktisch von Vorteil wenn es um das Vorhalten von Ersatz geht. Wichtiger aber ist: In den meisten Konfigurationen bestimmt die kleinste Platte die Gesamtkapazität des Verbundes. Eine 2TB Platte, gespiegelt mit einer 1TB HDD hat eine Nutzkapazität von 1TB + 1TB Redundanz, es wird also Platz bezahlt, aber verschenkt.

Idealerweise sollten alle Platten typgleich sein. Damit lassen sich Kompatibilität der Komponenten und zueinender passende Leistungsdaten wie zum Beispiel Zeitverhalten sicherstellen. Es senkt außerdem den logistischen Aufwand. Ein gewisses Risiko stellt allerdings die Möglichkeit von Serienfehlern dar, Herstellerseitigen Problemen oder Defekten in ganzen Produktions-Chargen. Dieses Risiko ist üblicherweise klein gegenüber den Kosten seiner Vermeidung.

Wichtig, aber heute auch üblich ist eine automatische Fehlererkennung der Laufwerke, wie zum Beispiel S.M.A.R.T. Ohne eine solche Fehlererkennung kann ein "sterbendes" Laufwerk zum einen den Verbund unbemerkt durch Datenfehler kompromittieren, zum anderen das oben erwähnte "Δt" so verlängern, dass der Ausfall zweier Platten wahrscheinlicher wird. Selbst bei automatischer Fehlererkennung ist zu beachen, dass diese nicht jeden Fehlerzustand zuverlässig erkennt.

3.1.2 Der (oder die) Controller

Zur Ansteuerung der angeschlossenen Laufwerke und zur logischen Implementation des RAID-Verbundes dienen RAID-Controller. Diese unterscheiden sich zum einen durch

die angebotene Funktionalität (hierzu mehr unter "RAID-Level" später) und durch die Art ihrer Realisierung: Vorwiegend als Hardware, als reine Software oder als Hybrid aus beidem.

3.1.2.1 Software RAID

Bei einem Software RAID werden, wie der Name schon sagt, alle Funktionen ohne spezielle Hardware - außer den zusätzlichen Platten - rein in Software realisiert. Praktisch alle gängigen, modernen Betriebssysteme bieten diese Möglichkeit bereits an.

Der Vorteil dieser Lösung ist der Preis, da er sich auf die Kosten der Laufwerke beschränkt, sowie das Fehlen ausfallträchtiger zusätzlicher Komponenten.

Der Nachteil liegt darin, dass die Steuerung des RAID die CPU belastet, da ihr diese Arbeit nicht von einem Hardware-Controller abgenommen wird. Auch die Auslastung des Systembusses steigt im Allgemeinen. Für Systeme mit geringer Auslastung dieser Komponenten (CPU, Bus) im Regelbetrieb kann dies durchaus akzeptabel sein.

3.1.2.2 Hardware RAID

Hier übernimmt ein dedizierte Hardwarecontroller pro Server, pro Platteneinheit, pro Array, oder sogar pro Platte die Aufgabe der Datenverteilung. Dieser hat oft eine eigene CPU und zusätzlichen Cache-Speicher und kann so die System-CPU entlasten. Der Durchsatz wird optimiert.

Die zusätzliche Hardware stellt selber ein Ausfallrisiko dar, allerdings ist dieses um mehrere Größenordnungen geringer als das eines Plattenausfalls. Es ist aber dieses Restrisiko, zusammen mit dem Ziel einen "Single Point of Failure" mit potentiell desaströsen Folgen zu vermeiden, dass zu den oben erwähnten Lösungen mit mehreren Controllern pro Installation führt

3.1.2.3 Hybrid: Host RAID

Hybrid-Lösungen, meist Host RAID genannt vereinigen die Nachteile der beiden anderen, sind aber billig und oft schon „on-board“ bei PC-Mainboards. Einige Chips übernehmen Teile der Steuerlogik, überlassen aber den Großteil der Arbeit ihren in Software auf der System-CPU laufenden Treibern.

Für professionelle Nutzung sind Host RAID ungeeignet, für nicht-professionelle zumindest fragwürdig. Die Zuverlässigkeit der verwendeten Komponenten und die Qualität der Treiber kann durchaus weit unter der (theoretischen) des Plattenverbundes liegen. Ein Ausfall kann schlimmstenfalls den gesamten Datenbestand unbrauchbar machen.

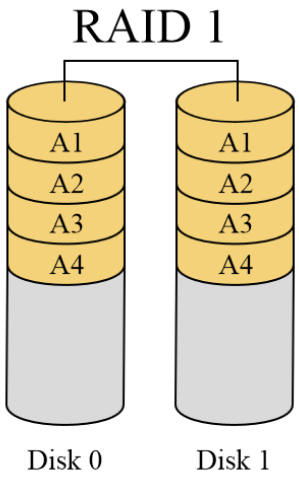
3.2 RAID Level

Die verschiedenen Varianten von RAID werden als RAID Level bezeichnet. Sie unterscheiden sich in ihrer vorwiegenden Zielsetzung - Verfügbarkeit, Performanz oder beides - und in der Art, wie diese verfolgt wird.

Ein gegebener RAID Controller unterstützt üblicherweise einen oder mehrere dieser Level, jedoch nicht alle. Welcher RAID Level in einem bestimmten Fall der geeignetste ist hängt von der spezifischen Aufgabenstellung und von Kosten-Nutzen-Betrachtungen ab.

Hier ein Überblick über die wichtigsten RAID Level und wie sie funktionieren.

3.2.1 RAID 1



Datenspiegelung oder auch Mirroring.

Alle Daten die auf Disk 0 geschrieben werden, werden auch auf Disk 1 gespeichert.

Quelle: http://de.wikipedia.org/wiki/RAID

Abbildung 4: RAID 1

RAID 1 hat den Vorteil voller Redundanz und dadurch hoher Sicherheit. Es ist einfach zu verstehen und auch einfach in der Implementation. Es ist auch möglich, auf mehr als eine Platte zu spiegeln, allerdings ist dies aus Kostengründen und aufgrund des abnehmenden Grenznutzens (schon bei drei Platten liegt die Ausfallwahrscheinlichkeit des Verbundes meist unter der des Restsystems) nur bei sicherheitskritischen Anwendungen eine Option.

Fällt eine Platte aus, so wird der Betrieb mit der verbleibenden fortgeführt, wobei die Spiegelung abgeschaltet wird. Die defekte Platte wird von einem Operator oder Admin per "hot swap" im Betrieb ausgetauscht, kann aber auch automatisch als "hot standby" sofort nach Ausfall zugeschaltet werden, wenn ein entsprechendes Reservelaufwerk sich schon im Verbund befindet. Dies ist allerdings bei RAID 5 (siehe unten) üblicher als bei RAID 1.

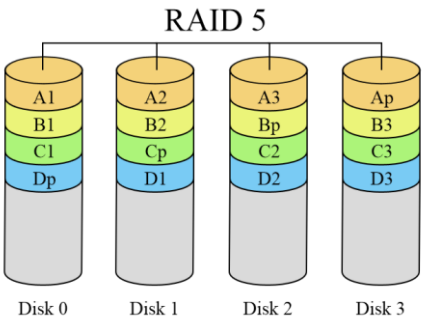
Nach Ersatz der ausgefallenen Platte beginnt der "Rebuild", üblicherweise ebenfalls ohne Unterbrechung des Rechenbetriebs. Die Daten der verbliebenen Platte werden jetzt auf das Ersatzlaufwerk gespielt, um die Konsistenz des Verbundes wieder herzustellen. Diese Phase ist aus verschiedenen Gründen kritisch. Zum einen dauert sie abhängig von Kapazität und Füllung der Platte mehrere Stunden. Ein Ausfall der überlebenden Platte würde jetzt zum Ausfall des Systems und wahrscheinlich zu Datenverlust führen. Rein statistisch ist die Wahrscheinlichkeit hierfür zwar gering, wenn jedoch der Ausfall der ersten Platte auf nicht-statistische Gründe zurückzuführen war - wie z.B. Ausfall der Kühlung, Probleme mit der gemeinsamen Stromversorgung, etc. - dann trägt die verbliebene Platte, die jetzt unter zusätzlicher Last Normalbetrieb + Rebuild läuft, ein erhöhtes Risiko. Außerdem kann es beim Kopieren großer Datenmengen - und genau das ist der Rebuild - zu zufälligen unkorrigierbaren Bitfehlern kommen, die die Integrität der Daten verletzen. Zwar ist die Wahrscheinlichkeit für einen solchen Fehler nur von der Größenordnung 10^{-15} pro kopiertem Bit, allerdings hat jedes Terabyte auch fast 10^{13} Bit, was das Risiko in den Prozentbereich hebt.

Ist der Wiederaufbau abgeschlossen wird wieder "Spiegelbetrieb" aufgenommen.

Da alle Daten in einem RAID 1 doppelt vorliegen, ist eine Leistungssteigerung möglich. Es können sowohl gleichzeitig von verschiedenen Platten Daten übertragen werden (wie bei RAID 0, siehe unten), was die Übertragungsrate erhöht, als auch konkurrent auf Daten zugegriffen werden, was die mittlere Latenz senkt.

Der Nachteil von RAID 1: Verdoppelter Preis für Speicherplatz, da zu jeder Platte mit Nutzdaten eine (mindestens gleich große) Platte für Spiegeldaten vorhanden sein muss.

3.2.2 RAID 5



Block-Level Striping mit verteilter Paritätsinformation.

A, B, C und D sind "Stripes" gleicher Größe (z.B. 64kB pro Laufwerk)

A1 ... D3 sind "Nutzdaten"

Ap ... Dp sind Paritätsinformationen, gebildet über die zugehörigen Nachbar-Stripes gleichen Buchstabens

Quelle: http://de.wikipedia.org/wiki/RAID

Abbildung 5: RAID 5

Die Idee hinter RAID 5 ist es, ohne eine Verdoppelung der Kosten für Laufwerke doch einen fast so hohen Sicherheitsgewinn wie bei RAID 1 zu erzielen. Hierzu werden die Daten nicht vollständig redundant gehalten, sondern lediglich so auf n Platten verteilt, dass sie jederzeit bei Ausfall einer Platte aus den verbliebenen n-1 rekonstruiert werden

können. Typisch sind hierbei Systeme mit n gleich drei oder fünf Platten (also gerade nicht den im Bild gezeigten vier).

Wie oben gezeigt erhöht sich zwar das Ausfallrisiko von zwei aus fünf Platten gegenüber einer RAID 1 Lösung um den Faktor Zehn, liegt allerdings noch 5000-mal unter dem einer einzelnen Platte. Und während der Plattenpreis bei Datenspiegelung sich verdoppelt, liegt er bei 4-aus-5 lediglich um 25% über der unsicheren Variante.

Realisiert wird diese Redundanz dadurch, dass die Platten in Stripes fester Größe aufgeteilt werden (siehe Abbildung 5) auf die die Daten verteilt werden. Hierbei ist eine Platte pro Stripe (allerdings nicht die gleiche Platte für jeden Stripe) für Paritätsinformationen reserviert. Diese Paritätsinformation wird per XOR aus den "Nutzdaten" gebildet. In der Abbildung ist also Ap = A1 xor A2 xor A3.

Um zu verhindern, dass bei Änderung der Daten auf einer Platte alle Platten gelesen werden müssen, um dann die Paritätsinformation zu aktualisieren, arbeiten manche RAID 5 Systeme so, dass sie erst die "alten" Daten lesen, aus diesen und den neu zu schreibenden eine inkrementelle Parität bilden (0=keine Änderung, 1=Bit hat sich geändert), dann erst schreiben und anschließend die Paritätsinformation mittels Inkrement anpassen. Logisch das Gleiche wie ein komplettes XOR, allerdings vermeidet es, auf alle Laufwerke zugreifen zu müssen.

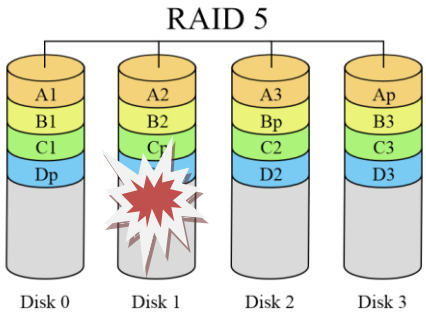


Abbildung 6: RAID 5 Recovery

0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	1

Wenn's bei RAID 5 kracht ...

Disk 1 sei ausgefallen und damit die Nutzdaten A2, B2, D1 verloren, ebenso die Paritätsinformationen Cp, die aus C1, C2 und C3 gebildet wurden.

Die Tabelle unten zeigt die zum Zeitpunkt des Unfalls auf den Platten vorhandenen Daten - stark vereinfacht, ein Bit repräsentiert einen ganzen Stripe.

Rote Daten sind jetzt verloren und nicht mehr lesbar, grüne Daten sind Paritätsinformationen, gebildet durch XOR der Nutzdaten der zugehörigen Stripes.

Wie leicht nachzuvollziehen ist, können alle verlorenen Daten durch einfaches "XORen" der verbliebenen Daten der gleichen Zeile (Stripes) wieder hergestellt werden, unabhängig davon, ob es Nutz- oder Paritätsinformationen sind:

A2: 0 xor 0 xor 0 = 0
B2: 0 xor 0 xor 1 = 1
Cp: 1 xor 0 xor 1 = 0
D1: 1 xor 1 xor 1 = 1

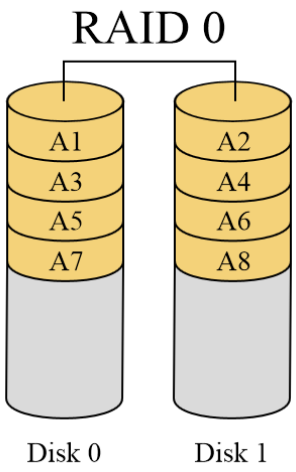
Das Verfahren bei Ausfall einer Platte entspricht weitgehend dem bei RAID 1 beschriebenen: Unterbrechen des "Spiegel-" bzw. "Paritäts-Modus", Ersetzen des defekten Laufwerks per Hot Swap oder aus Hot Spare, Rebuild.

Der Unterschied liegt darin, dass zur unterbrechungsfreien Fortsetzung des Betriebs nicht mehr alle Daten vorliegen. Daher müssen diese bei Bedarf "on the fly" aus den Daten der überlebenden Platten rekonstruiert werden. Auch ist der Wiederaufbau der ersetzten Platte nicht einfach durch Kopieren zu bewerkstelligen, sondern muss als Rekonstruktion über XORen der Stripes der anderen Platten erfolgen. Wie auch bei RAID 1 ist dies eine kritische Phase, insbesondere da die zusätzliche Last auf Bus und Platten bei RAID 5 in dieser Situation noch größer ist.

Ist die Wiederherstellung erfolgt wird der Normalbetrieb wieder aufgenommen.

Auch mit RAID 5 ist eine Leistungssteigerung möglich. Zwar nicht in Sachen Zugriffszeit, da die Redundanz dafür zu gering ist, allerdings in Bezug auf die Übertragungsrate, da die Daten schon verteilt (striped) vorliegen. Dies gilt allerdings nur für lesenden Zugriff - Schreibvorgänge, insbesondere viele kleine Schreibzugriffe, dauern länger als bei einzelnen Laufwerken, da bei jedem Schreibzugriff auch die Parität neu zu schreiben ist, wozu wiederum alle Platten gelesen werden müssen.

3.2.3 RAID 0



Striping ... wenn's schnell gehen muss

Eigentlich kein RAID, da nicht redundant:

Die Information ist verteilt, aber nicht doppelt vorhanden.

Fällt eine Platte aus, dann sind beide nicht mehr verfügbar, Daten sind möglicherweise komplett verloren.

Quelle: http://de.wikipedia.org/wiki/RAID

Abbildung 7: RAID 0

RAID 0 ist ein "uneigentliches" RAID, da es nicht redundant ist und sogar statistisch die Verfügbarkeit des Systems senkt. Dies liegt daran, dass das einzige Ziel dieser Konfiguration die Steigerung der Übertragungsrate ist.

Hierzu werden die logisch einem Laufwerk zugeordneten Daten auf zwei oder mehr physische Laufwerke verteilt. Daher können sowohl beim Lesen, als auch beim

Schreiben hinreichend großer Dateien mehrere Platten gleichzeitig genutzt werden. Die Zugriffsgeschwindigkeit verbessert sich allerdings nicht.

Der Preis hierfür ist eine gesteigerte Ausfallwahrscheinlichkeit, da jetzt der Defekt einer einzelnen Platte das Gesamtsystem lahmlegt. Fällt ein einzelnes Laufwerk in einem gegebenen Zeitintervall mit der Wahrscheinlichkeit p aus, dann ist die Wahrscheinlichkeit des Versagens für einen RAID 0 Verbund aus n Platten in der gleichen Zeit $1 - (1 - p)^n$.

Wirklich attraktiv wird diese Lösung erst in Verbindung mit einem "echten" RAID, welches zum Leistungsgewinn des RAID 0 auch noch ein Plus an Sicherheit bringt.

3.2.4 Seltener genutzt oder obsolet

Wie schon die Lücke in der Nummerierung der RAID Level vermuten lässt, gibt es auch noch weitere Varianten. Diese sind jedoch nicht annähernd so weit verbreitet wie die bereits genannten. Hier nur ein kurzer Überblick als Lückenfüller.

3.2.4.1 RAID 2: Bit-Level Striping mit Hamming-Code

Die Daten werden auf einen Hamming-Code umcodiert und dann bitweise auf mehrere Platten verteilt. Die länge des Codeworts in Bit bestimmt die (Mindest-)Anzahl der Platten. Inzwischen obsolet, früher im Großrechnerbereich verwendet.

3.2.4.2 RAID 3: Byte-Level Striping, separate Paritätsplatte

Ähnlich RAID 4 (siehe unten), allerdings enthalten die Stripes keine Datenblöcke, sondern sind lediglich ein Byte groß. Praktisch durch RAID 5 abgelöst.

3.2.4.3 RAID 4: Block-Level Striping, separate Paritätsplatte

Wie RAID 5, allerdings sind die Paritätsinformationen nicht auf alle Platten verteilt, sondern werden für alle "Nutzplatten" auf einer dedizierten "Paritätsplatte" gehalten. Nachteil: Die Paritätsplatte wird zum Flaschenhals des Systems, da sie an jedem schreibenden Zugriff beteiligt ist. Praktisch durch RAID 5 abgelöst.

3.2.4.4 RAID 6: Block-Level Striping mit doppelter Paritätsinformation

Ähnlich RAID 5, allerdings wird die Parität nicht nur auf einem, sondern auf zwei Laufwerken vorgehalten. Daher wird der Ausfall von bis zu zwei Laufwerken gleichzeitig ohne Datenverlust verkraftet. Erfordert im Vergleich mit RAID 5 eine weitere zusätzliche Festplatte und ist bezüglich Sicherheit - je nach Gesamtzahl der Platten - zwischen RAID 5 und RAID 1 angesiedelt, leistungsmäßig wegen der zusätzlichen (Paritäts-)Schreibvorgänge aber langsamer als beide. Findet noch Verwendung.

3.2.5 Kombinationen

Da ein RAID-Verbund als ein einzelnes logisches Laufwerk betrachtet werden kann, ist es auch möglich, einen Verbund von RAID-Verbunden herzustellen. Sinn macht dies dann, wenn dadurch unterschiedliche Stärken verschiedener RAID Level kombiniert werden, bzw. Schwächen ausgeglichen.

So ist zum Beispiel ein RAID 10 ein RAID 0 das aus mehreren RAID 1 besteht. Umgekehrt beim RAID 01. Auch wenn Umkehrungen nicht die gleichen Eigenschaften in Bezug auf Leistung und Sicherheit haben, so verfolgen doch beide das gleiche Ziel: Die Sicherheit eines RAID 1, gepaart mit der Geschwindigkeit eines RAID 0. Andere Varianten sind z.B RAIDs 05/50, 15/51. Einige dieser Kombinationen, wie z.B. RAID 10, sind tatsächlich von praktischer Bedeutung, allerdings würde ihre Diskussion den Rahmen einer Einführung sprengen.

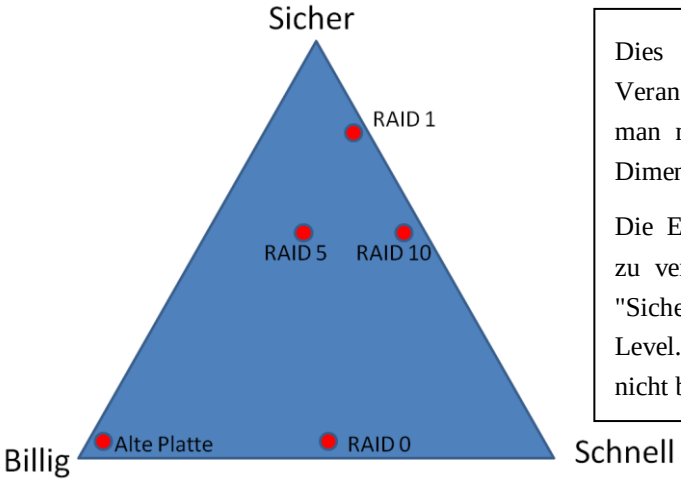
3.2.6 Uneigentliche RAID

Im Zusammenhang mit RAID wird man auch gelegentlich auf Begriffe wie NRAID (Non-RAID) oder JBOD (Just a Bunch Of Disks) treffen. Dieses sind zwar auch Festplattenverbunde, allerdings ohne die im Vorangegangenen geschilderte Funktionalität eines RAID, sei es in Bezug auf Verfügbarkeit oder Performanz. Meist ist das Ziel dieser "uneigentlichen" lediglich die Zusammenfassung zu größeren logischen Volumen.

3.3 Praktische Gesichtspunkte

Bei der Entscheidung über das "Ob" und das "Welches" im RAID-Bereich stellen sich im Allgemeinen vier Fragen. Die erste, "Wie viel TB Kapazität" sei hier einmal als durch das Systemumfeld beantwortet angenommen. Die anderen drei sind die nach Sicherheit, Geschwindigkeit und Preis. Diese befinden sich in einem Spannungsfeld und jede Verbesserung in eine der drei Richtungen geht auf Kosten eines oder beider verbleibender Parameter.

Wie so oft setzen fachlich gute Entscheidungen eine genaue Kenntnis der Anforderungen und des (betrieblichen) Umfelds voraus.



Dies Diagramm ist nur zur groben Veranschaulichung gedacht, insbesondere da man mit einem Dreieck nicht wirklich drei Dimensionen repräsentieren kann.

Die Eckpunkte sind als "Anziehungspunkte" zu verstehen - je näher ein roter Punkt an "Sicher" ist, desto sicherer ist sein RAID-Level. RAID 10 ist z.B. schnell, sicher, aber nicht billig.

Abbildung 8: Bermuda-Dreieck der Systementscheidung

Abschließend sei noch darauf hingewiesen wovor RAID nicht schützt. So leistungsfähig RAID auch als Mittel zum Schutz vor (vorwiegend) statistischen Ausfällen von Festplatten sein kann, so wenig hilft es gegen eine ganze Reihe von anderen Fehlerquellen.

Hierzu gehören natürlich logische Fehler wie Programmfehler, fehlerhafte Systemkonfigurationen und so weiter. Malware, Sabotage, Brand, Wasserschaden, ... um es kurz zu fassen: RAID ersetzt nie eine Datensicherung und erlaubt auch keine Verlängerung von Sicherungszyklen.

Selbst wenn es speziell um Festplatten-Integrität geht, gibt es eine Reihe von Ausfallgründen, die nicht statistisch, bzw. unabhängig sind und vor denen Hardware- und Daten-Redundanz daher nur begrenzt schützen können, so z.B. Stromausfall, Überhitzung und Fehlerkaskaden.

Controllerfehler wurden schon angesprochen und auch wenn sie bei guten Komponenten selten sind, so können ihre Folgen doch schwerwiegend sein.

Generell gilt, dass nicht jeder Fehler bei Auftreten erkannt wird und oft um so schwerwiegendere Folgen hat, je länger es bis zum Bemerken erster Symptome dauert. Unerkannte Fehler im RAID können gelegentlich erst beim „Rebuild“ zum Vorschein kommen und dann mit Datenverlust einhergehen.

Auch hier gilt wieder das Spannungsfeld zwischen Sicherheit, Leistung und Preis. So kann die Chance, einen Fehlerer rechtzeitig zu erkennen erhöht werden, indem Maßnahmen wie „read after write“ ergriffen werden, oder durch Deaktivierung von Disk-Caches (um sicherzustellen, dass was geschrieben ist sich auch tatsächlich auf der Platte befindet). Dies verlangsamt allerdings das System, oft nicht unerheblich. Durch zusätzliche Investitionen kann dies gelegentlich wieder kompensiert werden.

Und dann gibt es da noch die Fehler, die man sich erst durch die Sicherheitsmaßnahmen ins Haus holt. So z.B. das RAID 5 „Write Hole“: Fällt das System nach dem Schreiben von Nutzdaten, aber vor Aktualisierung der Parity-Informationen aus, so kann es zu Inkonsistenzen des Datenbestandes führen, die nach der Wiederinbetriebnahme nicht einmal festgestellt werden.

Solange man sich allerdings dieser Grenzen des Anwendungsbereichs bewusst ist, stellen RAID (> 0) eine exzellente Möglichkeit dar, die Verfügbarkeit großer Systeme sicherzustellen.

4 Zusammenfassung und Ausblick

Wie wir gesehen haben stellen RAID einen effektiven, effizienten und kostengünstigen Schutz vor Datenverlust durch Plattenausfälle und Plattenschäden dar. Sie sind eine Maßnahme zur Erhöhung der Verfügbarkeit, allerdings nicht zur Datensicherung.

Wenn man die Verteilung der Daten auf mehrere Platten für einen Geschwindigkeitsgewinn nutzt, dann entschärfen RAID den Flaschenhals zwischen Sekundärspeicher und CPU/Primärspeicher, gelegentlich allerdings im Austausch gegen Sicherheit.

Welche technische und logische Implementation im Einzelfall zu wählen ist hängt ab vom Anwendungsumfeld und den spezifischen Anforderungen.

In der Praxis wird heute wohl kein Serverpark mehr zu finden sein, in dem RAID nicht in der einen oder anderen Form zum Einsatz kommen.

Die Zukunft der RAID-Technologie ist eng verbunden mit der der (mechanischen) Festplatte. Diese gerät unter zunehmenden Druck durch SSD, auch wenn diese im Preis pro GB noch um mehr als eine Größenordnung höher liegen und momentan noch eine begrenzte Lebenserwartung haben.

Die Frage könnte daher sein: Was bringen RAID für SSD? Technisch lassen sich SSD genau wie Festplatten zu Verbunden zusammenschalten. Allerdings sind die Nutzenbetrachtungen auf SSD nur begrenzt anwendbar.

Zum einen sind die Hauptfehlerquellen bei SSD nicht statistisch, sondern "verschleißbedingt", zum anderen trägt redundante Datenhaltung gerade zu diesem Verschleiß bei. SSD Speicherzellen haben eine begrenzte Zahl von Schreibzyklen bevor sie versagen und das doppelte Schreiben von Daten verdoppelt auch diese Zyklen.

Des Weiteren verfügen SSDs bereits über eine Fehlerkorrektur in Form eines ECC (Error Correction Code).

Zur Steigerung der Übertragungsgeschwindigkeit sind RAID auch bei SSD geeignet, wobei aufgrund der hohen Datenrate der SSD hier jedoch der Datenbus zum Engpass werden kann. Zu beachten ist auch, dass SSD großer Kapazität prinzipiell schneller sind als kleine, das Zusammenschalten kleiner zu einem Leistungsverbund also wenig Sinn macht.

Trotz all dieser Einwände hängt noch viel von der zukünftigen Entwicklung dieses Speichermediums ab. Es ist durchaus denkbar, dass auch die SSD sich zukünftig in RAIDs wiederfindet.

Anhang A: Übersicht der RAID-Level

Nachfolgend sind die wichtigsten RAID Level mit ihren hauptsächlichen Eigenschaften und ihrer (von mir geschätzten) Verbreitung aufgelistet. Die Bewertung der Leistungsfähigkeit ist nur eine grobe Richtlinie, da diese im Einzelfall von Details der Konfiguration, wie z.B. der Zahl der Platten und der Art der verwendeten Komponenten abhängt.

Tabelle 1: RAID Level

Level	**Hauptziel**	**S**	**G**	**P**[1)]	**Verbreitung**
kein RAID	(Normalbetrieb)	-	-	++	groß
RAID 0	Geschwindigkeit	--	+	+(+)	groß
RAID 1	Verfügbarkeit	++	0	--	groß
RAID 2	Verfügbarkeit, Geschw.	+(+)	+	--(-)	praktisch obsolet
RAID 3	Verfügbarkeit	+	0/-	0	Ersetzt durch 5
RAID 4	Verfügbarkeit	+	0/-	0	Ersetzt durch 5
RAID 5	Verfügbarkeit	+	0	0	groß
RAID 6	Verfügbarkeit	+(+)	0/-	-	gering
RAID 01/10	Verfügbarkeit, Geschw.	+(+)	+	-(-)	moderat
RAID 05/50	Verfügbarkeit, Geschw.	0	+	0	moderat

1) **S**icherheit / **G**eschwindigkeit / **P**reis

Geschwindigkeit bezieht sich auf einen Mix aus Zugriffszeit und Übertragungsrate. Preis bezieht sich auf Platten plus Controller, niedriger Preis = "+", hoher = "-".

Quellenverzeichnis

David Patterson, Garth A.Gibson, Randy Katz „A Case for Redundant Arrays of Independent Disks (RAID)“, Computer Sciences Division UC Berkeley, 1987 http://www.eecs.berkeley.edu/Pubs/TechRpts/1987/CSD-87-391.pdf

Holger Uhlig http://www.heinlein-support.de/upload/slac08/Heinlein-RAID_Mathematik_fuer_Admins.pdf , Heinlein Professional Linux Support GmbH

IBM http://www-03.ibm.com/ibm/history/exhibits/storage/storage_3380c.html

Tobias Allinger http://referate.mezdata.de/sj2003/festplatte_tobias-allinger/ausarbeitung/geschichte.html

http://de.kioskea.net/contents/histoire/disque.php3

Wikipedia http://de.wikipedia.org/wiki/RAID
http://en.wikipedia.org/wiki/RAID
http://de.wikipedia.org/wiki/Festplattenlaufwerk
http://en.wikipedia.org/wiki/Hard_disk_drive
http://de.wikipedia.org/wiki/Mean_Time_Between_Failures

Als Dokumentenvorlage diente der zur nicht-kommerziellen Nutzung freie Entwurf von **Prof. Dr. Wolf-Fritz Riekert**, Hochschule der Medien Stuttgart: http://v.hdm-stuttgart.de/~riekert/theses/

(Datum der Zugriffe für alle Hyperlinks in diesem Dokument: 27. August 2012)