Neue RAID-Level im Überblick

Doppelt gesichert

Mit der Datenflut wachsen die Speichersysteme, die sie auffangen sollen. Mit deren Größe steigen aber auch die möglichen Folgen eines Ausfalls. Andererseits sind Lese- oder Schreibfehler und auch das komplette Versagen einer Disk nur eine Frage der Zeit. Einen Ausweg bieten Plattengruppen mit eingebauter Redundanz, wie sie die bekannten Parity-RAID-Verfahren erzeugen. Sie können den Ausfall einer einzelnen Disk kompensieren. Allerdings wird dabei selten bedacht, dass die klassischen RAID-Konfigurationen für Storage-Systeme heutiger Dimensionen keinen hinreichenden Schutz mehr gewähren. Stattdessen bieten sich neue Lösungen an.

Marcus Schuster

Die Behauptung, die hergebrachten und gut bekannten RAID-Level reichten heutzutage nicht mehr aus, klingt zunächst nach einer Übertreibung – doch ein Blick auf die Details macht klar, dass sie sich auf Tatsachen gründet.

Betrachtet man etwa ein herkömmliches RAID-Set, das *n* Festplatten durch RAID-Level 5 zusammenfasst, dann offenbaren sich die prinzipiellen Schwächen klassischer Parity-RAID-Level, sobald eine Festplatte ausgefallen ist (Degraded-Zustand) oder eine ausgefallene Festplatte durch eine neue ersetzt wird (Rebuild-Zustand). In diesen beiden Betriebszuständen sind die Daten dieses RAID-Sets in erhöhtem Maße gefährdet: Weder darf eine zweite Festplatte ausfallen, noch darf bei der Rekonstruktion der ausgefallenen Platte oder bei anderen I/O-Prozessen ein einziger Sektor auf den verbliebenen Festplatten schadhaft oder unlesbar sein.

Die Wahrscheinlichkeit für einen zweiten Plattenausfall hängt von der Gesamtzahl der Festplatten, ihrer durchschnittlichen Betriebsdauer bis zum Fehlerfall (MTTF, Mean Time to Failure) und der Zeitspanne ab, die die Wiederherstellung der ausgefallenen Platte benötigt. Die Wiederherstellungszeit (MTTR, Mean Time to Repair) beinhaltet den Austausch der Festplatte und die Rebuild-Dauer, die stark von der Einzelplattenkapazität und auch von der Auslastung des RAID-Systems abhängt.

Eine Berechnungsgrundlage für die MTTDL (Mean Time to Data Loss) liefert **(1)**. Bei *N* Festplatten, die einen einzigen RAID-5-Verbund bilden, wird allgemein angenommen:

$$MTTDL = \frac{MTTF(disk1) \cdot MTTF(disk2)}{N \cdot (N-1) \cdot MTTR(disk1)}$$

Für einen einzigen RAID-Verbund aus *N* Festplatten, der zwei beliebige Festplattenausfälle tolerieren kann, nimmt man an:

$$MTTDL = \frac{MTTF(disk1) \cdot MTTF(disk2) \cdot MTTF(disk3)}{N \cdot (N-1) \cdot (N-2) \cdot MTTR^2(disk)}$$

Einfache theoretische Modelle sehen alle Plattenausfälle als unabhängig voneinander an, und für die MTTF jeder Platte wird ein identischer Wert zugrunde gelegt. Dies führt üblicherweise zu sehr hohen MTTDL-Werten.

Für eine praxisgerechtere Beurteilung sollte man allerdings davon ausgehen, dass Plattenausfälle korreliert auftreten. Insbesondere fallen Festplatten zu Beginn ihrer Betriebszeit oder nach langer Einsatzzeit mit einer höheren Wahrscheinlichkeit aus. Bei identischen Festplatten aus der gleichen Charge ist deshalb mit einer Häufung der Plattenausfälle nach langer Einsatzzeit zu rechnen. Zudem sind die Festplatten eines RAID-Systems ähnlichen Umgebungsbedingungen ausgesetzt. So wirken sich Einflüsse, wie Erschütterungen, Spannungsschwankungen, hohe Umgebungstemperaturen und allgemeine Alterungsprozesse, auf alle Platten des Verbundes in ähnlicher Weise aus.

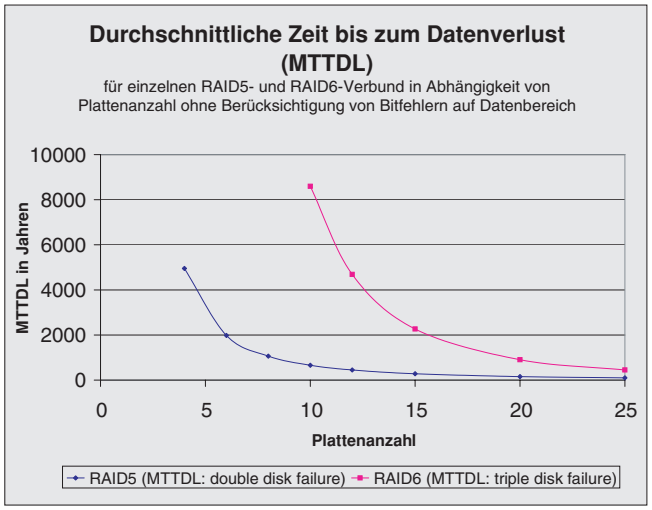


Abbildung 1: Als Berechnungsgrundlage für dieses Beispiel diente: MTTF(disk1)=500000h, MTTF(disk2)=50000h, MTTF(disk3)=5000h, MTTR=48h.

Diese Einflüsse lassen sich nur annäherungsweise abschätzen, indem man korrelierte Plattenausfälle durch von Ausfall zu Ausfall sinkende MTTF-Werte berücksichtigt. Für den Ausfall der ersten Platte lässt sich der MTTF-Wert des Datenblattwertes verwenden. Für jedes weitere Versagen sollte man dann beispielsweise einen um das Zehnfache geringeren MTTF-Wert berücksichtigen.

Ein Berechnungsbeispiel zeigt **Abbildung 1**. Ist es auch nahezu unmöglich, ein reales System mit diesen Schätzwerten exakt zu beschreiben, so erhält man doch zumindest einen Eindruck, welchen Nutzen ein RAID-Level bietet, der auch den Ausfall zweier Festplatten toleriert.

Eine weitere Gefährdung des Datenbestandes geht von Einzelbitfehlern oder defekten Plattensektoren aus. Während der Wiederherstellungszeit muss gewährleistet sein, dass auf den verbliebenen Festplatten alle Sektoren fehlerfrei zu bearbeiten sind. Das Risiko eines partiellen

Datenverlusts durch defekte Plattensektoren während eines Rebuild lässt sich ebenfalls abschätzen. Die Wahrscheinlichkeit ist abhängig von der Gesamtkapazität des RAID-Sets und der Bitfehlerwahrscheinlichkeit der verwendeten Festplatten. Bitfehlerwahrscheinlichkeiten geben die Festplattenhersteller in einem Größenordnungsbereich von $1:10^{14}$ bis $1:10^{15}$ an. Nimmt man eine unabhängige Wahrscheinlichkeit p für Bitfehler an, so errechnet sich eine Wahrscheinlichkeit b dafür, dass s Bit fehlerfrei zu lesen sind aus:

$$b = (1 - p)^s$$

Einen Eindruck vermittelt **Abbildung 2**. Es zeigt sich, dass eine Variation der Bitfehlerrate einer einzelnen Platte von nur einer Zehnerpotenz gravierende Auswirkungen bei großer Gesamtkapazität des RAID-Sets hat.

Heute lassen sich unter Verwendung klassischer Parity-RAID-Level problemlos RAID-Sets von 5 TByte und mehr konfigurieren. Bei einer Bitfehlerwahrscheinlichkeit von $1:10^{14}$ bleibt ein Risiko von etwa 30 Prozent, durch defekte Sektoren inkonsistente Dateisysteme oder fehlerhafte Daten zu erhalten. Da sind Zweifel angebracht, ob die angestrebte Verfügbarkeit auf diese Weise tatsächlich zu erreichen ist. In Kombination mit dem Risiko doppelter Plattenausfälle oder datenzerstörender Ausfälle weiterer Komponenten des RAID-Systems wird die Zuverlässigkeit noch weiter minimiert. Man sollte sich durch MTTDL-Angaben von beispielsweise 100 Jahren, die sich bei entsprechenden Abschätzungen ergeben, nicht täuschen lassen, denn es besteht dabei immer noch ein Restrisiko von fünf Prozent innerhalb einer Betriebsdauer von fünf Jahren einen Datenverlust zu erleiden (**Abbildung 3**). Das Restrisiko ergibt sich aus der MTTDL und der Betriebsdauer t zu $1\text{-}R(t)$, wobei $R(t)$ ein Maß für die Zuverlässigkeit (Reliability) ist und sich wie folgt aus der MTTDL berechnet:

$$R(t) = \exp\left(\frac{-t}{MTTDL}\right)$$

Als Ausweg bleibt zunächst der Einsatz alternativer klassischer RAID-Level, wie beispielsweise RAID 10, RAID 30, RAID 40, RAID 50 – auf Kosten der Speicherplatzeffizienz. Mehrfache Plattenausfälle sind hier aber nur in bestimmten Plattenkombinationen tolerierbar. Allein wegen der wachsenden Festplattenanzahl und der steigenden Kapazität einzelner Platten gestaltet sich die Suche nach einem Kompromiss zwischen Verfügbarkeit, Fehlertoleranz und Speicherplatzeffizienz immer schwieriger.

Eine nahe liegende Idee

Die prinzipiellen Schwächen der klassischen RAID-Level erkannte die Forschung schon früh. Bereits der RAID-Level 2, der durch Hamming-Codes zumindest Bitfehler korrigieren kann, stellt einen ersten Ansatz dar, um diese Schwierigkeiten zu umgehen. Weitere Konzepte folgten bereits in den 90er Jahren **(2)**. Wegen der damals geringeren technischen Möglichkeiten oder wegen des geringeren Problemdrucks erwuchs daraus allerdings noch keine Implementierung.

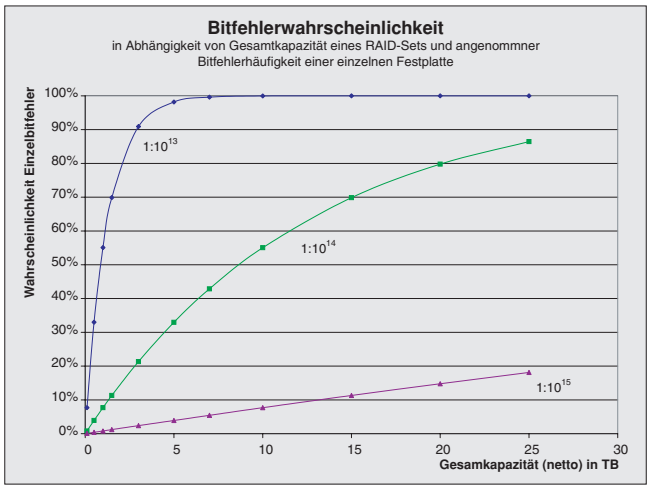


Abbildung 2: Bitfehlerwahrscheinlichkeiten für unterschiedliche Bitfehlerraten: Bereits relativ geringfügig höhere Werte wirken sich bei großen RAID-Sets dramatisch aus.

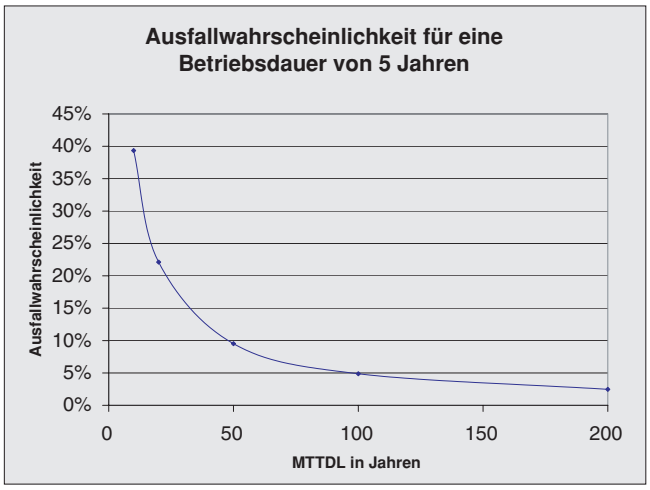


Abbildung 3: Der Zusammenhang zwischen MTTDL und Ausfallwahrscheinlichkeit: Ein fünfprozentiges Restrisiko für einen Datenverlust in den ersten fünf Jahren bleibt bestehen.

Festplatten-anzahl	RAID-01, RAID-10	RAID-3, RAID-4, RAID-5	RAID-6, RAID-DP etc.
3	-	66,70%	-
4	50%	75%	50%
5	-	80%	60%
8	50%	87,50%	75%
10	50%	90%	80%

Abbildung 4: Speichereffizienz (Verhältnis Netto- zu Bruttokapazität).

Inzwischen setzen sich neue RAID-Implementierungen zunehmend auf dem Markt durch. Ihre gemeinsame Grundidee ist nahe liegend: Eine zweite, unabhängige Prüfsumme bewirkt, dass der Ausfall eines beliebigen Plattenpaares zu verkraften ist. Damit verbunden ist auch ein besserer Schutz vor defekten Plattensektoren, weil zusätzliche Korrekturmöglichkeiten verfügbar sind. Man investiert als Kompromiss an die Speichereffizienz die Kapazität einer zusätzlichen Festplatte und gewinnt dadurch bessere Fehlertoleranz und Verfügbarkeit, denn nun darf ein beliebiges Paar Festplatten im gleichen Zeitfenster ausfallen.

Als weitergehende Verallgemeinerung dieses Konzepts bildet man ein Array von *n+m* Festplatten. Die Datenpakete liegen im Speicherbereich, der *n* Festplatten. Dann werden *m* unabhängige Prüfsummen gebildet, die sich auf die verbleibende Speicherkapazität der *m* Einzelfestplatten verteilen. Man spricht hier allgemein von RAID-n+m, was deutlich machen soll, dass *m* Festplatten gleichzeitig ausfallen können. RAID-n+m ist lediglich eine Bezeichnung für eine RAID-Level-Klasse, die über die Einzelheiten der verwendeten Algorithmen zunächst nichts aussagt. Konkrete Vertreter der RAID-n+2-Klasse sind beispielsweise RAID 6, RAID DP oder RAID 5DP.

Unterschiedlich umgesetzt

Verfügbare Implementierungen, wie RAID 6, RAID DP, RAID 5DP, RAID^n und andere, unterscheiden sich zunächst im Verfahren, eine zweite oder weitere unabhängige Prüfsummen zu berechnen. Diese Prüfsummen lassen sich einerseits durch zweifache XOR-Bildung gewinnen oder andererseits durch alternative fehlerkorrigierende Kodierungsverfahren. Ein weiteres Unterscheidungsmerkmal betrifft die Daten- und Prüfsummenpakete auf die Festplatten verteilen. Das kann analog zu RAID 4 mit dedizierten Prüfsummenplatten oder ähnlich RAID 5 mit gleichmäßig über alle Platten verteilten Prüfsummen erfolgen. (**Abbildung 5 a und b**)

Zweifaches XOR

Beispiele für dieses Verfahren sind EVENODD **(2)**, RDP (Row Diagonal Parity) **(3)**, RAID DP und RAID 5DP. RAID DP von Network Appliance ist eine spezielle Form allgemeiner RDP-Verfahren und steht im NetApp-Betriebssystem zur Verfügung. RAID 5DP (RAID 5 Double Parity) ist die Bezeichnung von HP für einen RAID-Level der VA7000-Serie.

Allgemein entspricht die erste Prüfsumme der üblichen XOR-Prüfsumme, die auch bei RAID 3, RAID 4 und RAID 5 verwendet wird. Sie wird horizontal über die einzelnen Datenpakete gebildet. Die zweite Prüfsumme wird ebenfalls

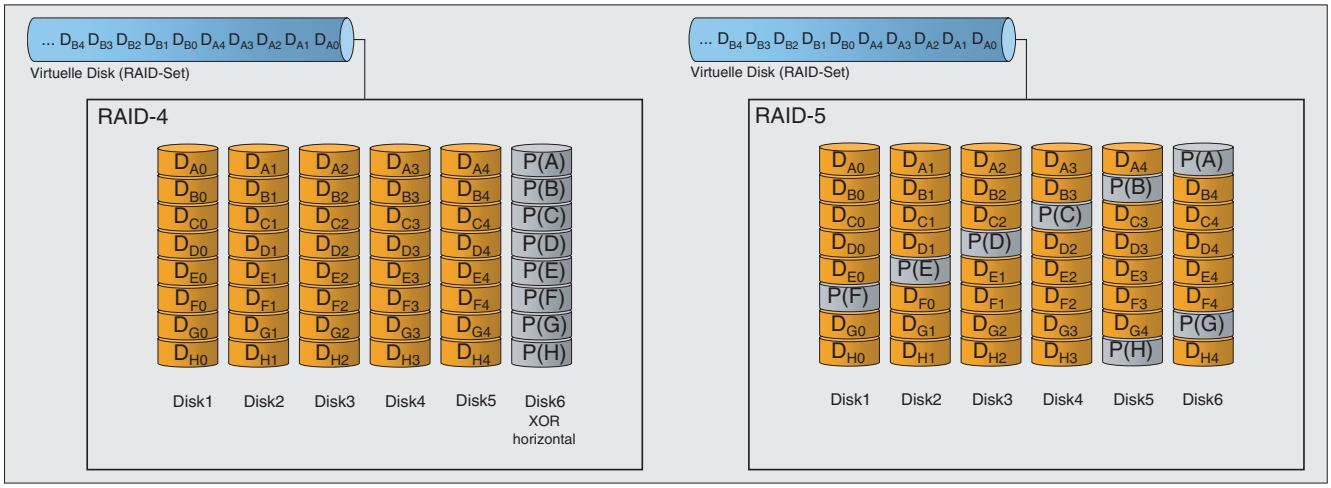
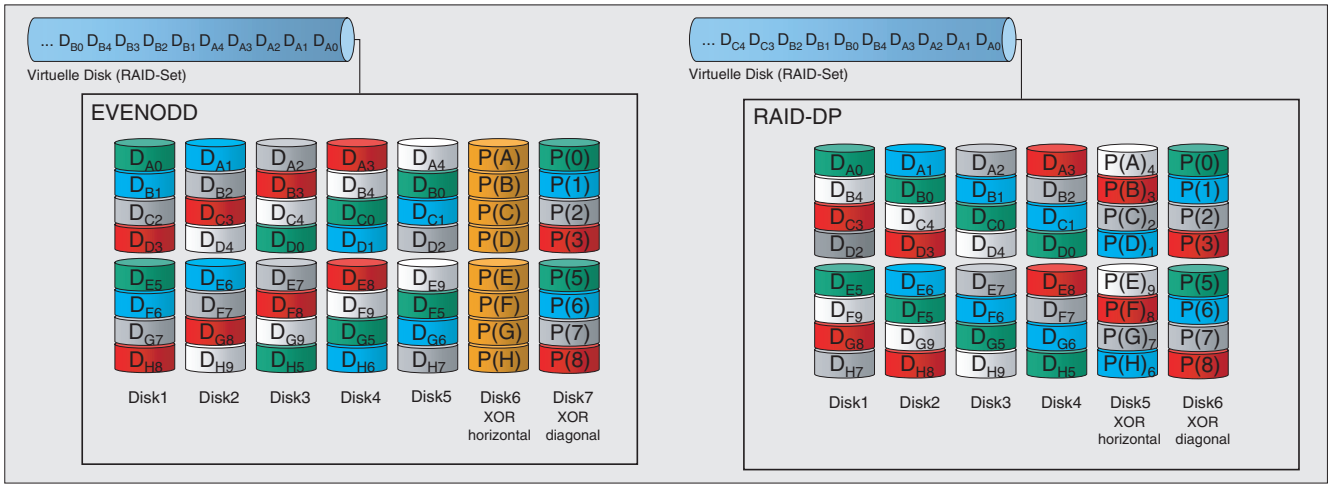


Abbildung 5a und b: RAID 4 und RAID 5 unterscheiden sich in der Speicherung der Prüfsummen.

durch einen XOR-Algorithmus errechnet. Die Verwendung diagonal liegender Datenpakete für die zweite XOR-Prüfsumme gewährleistet ihre Unabhängigkeit von der ersten.
Die Verteilung der Daten- und der Prüfsummenpakete kann wie bei RAID 4 mit dedizierten Daten- und Prüfsummenplatten erfolgen, was den Vorteil einer unkomplizierten Absicherung bestehender RAID-4-Sets durch Erweiterung um eine weitere Prüfsummenplatte bietet. Die bekannten Einschränkungen von RAID 4, die durch dedizierte Prüfsummenplatten entstehen, bleiben dabei jedoch erhalten.
Umgekehrt ist eine Rückkehr zu RAID 4 ebenfalls ohne weiteres durch Abkoppeln der zusätzlichen Prüfsummenplatte möglich. Die von RAID 5 bekannte gleichmäßige Verteilung von Daten- und Prüfsummenblöcken auf alle Platten lässt sich ebenfalls einsetzen, wobei darauf zu achten ist, dass die Umverteilung die Unabhängigkeit der Prüfsummen nicht zerstört.
Als Beispiele für zweifaches XOR im RAID-4-Stil dienen EVENODD und RAID DP. Beiden Verfahren gemeinsam ist die Tatsache, dass die Unabhängigkeit beider Prüfsummen nur bei Verwendung einer bestimmten Anzahl von Festplatten gegeben ist. Bei EVENODD muss die Anzahl der Datenplatten eine Primzahl sein. Bei RAID DP (**Abbildung 6 a und b**) muss die Anzahl der Datenplatten zuzüglich der horizontalen Prüfsummenplatte einer Primzahl entsprechen. Für den Fall einer beliebigen Festplattenanzahl wird der Algorithmus um virtuelle Festplatten, die keine Daten – also nur Nullen – enthalten, erweitert, bis das jeweilige Primzahlkriterium erreicht ist. Diese virtuellen Festplatten dienen nur als Platzhalter im Rechenverfahren, um die Unabhängigkeit beider Prüfsummen zu gewährleisten.
Als weitere Gemeinsamkeit fällt auf, dass man nicht alle diagonalen Prüfsummen benötigt. Jeweils eine Diagonale – in den Abbildungen weiße Platten – bleibt unbenutzt. Zudem lassen sich diese Verfahren auf herkömmlicher RAID-Hardware mit XOR-Engine ohne weiteres implementieren.
RAID DP bezieht im Unterschied zu EVENODD die Daten der horizontalen Prüfsummenplatte in die diagonale Prüfsummenbildung mit ein, was zu einer Verringerung der XOR-Operationen gegenüber EVENODD, insbesondere bei kleiner Festplattenanzahl, führt.

Alternative Prüfsummenbildung

Das langfristige Ziel, allgemeine RAID-n+m-Level verfügbar zu machen, ist mit mehrfachen XOR-Algorithmen nicht zu erreichen. Sie sind auf den simultanen Ausfall zweier Platten beschränkt. Nur allgemeinere, fehlerkorrigierende Kodierungsverfahren führen hier zum Ziel. Kandidaten sind beispielsweise Reed-Solomon-Codes, Vandermonde-based Reed-Solomon-Codes, Bose-Chaudhuri-Hocquenghem-Codes, die von RAID 2 bekannten Hamming-Codes, Gallager- und Tornado-Codes. In Zukunft werden zunächst RAID-n+2-Implementierungen verfügbar sein, deren Kodierungsverfahren aber häufig auf den allgemeinen RAID-n+m-Fall erweiterbar sind. Die zugrunde liegende Mathematik ist durchaus nicht trivial. Die Berechnungen sind nicht so leicht nachvollziehbar wie die XOR-Prüfsummenbildung. Zudem sind die



Abbildungen 6 a und b: EVENODD und RAID DP.

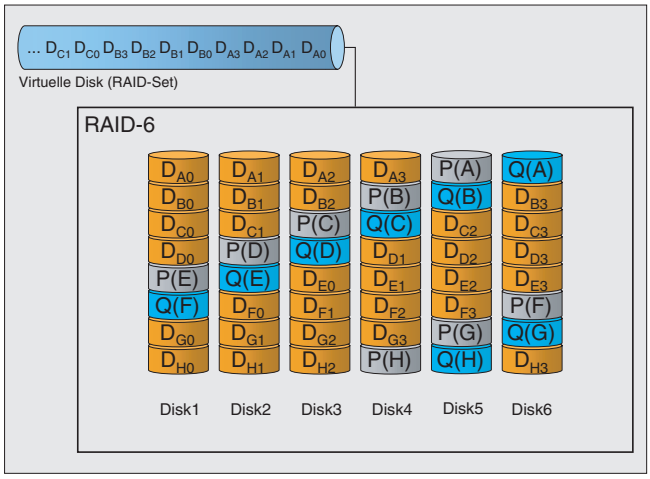


Abbildung 7: Das Prinzip von RAID 6: Die unabhängigen Prüfsummen P und Q werden nach einem ähnlichen Prinzip wie bei RAID 5 gespeichert.

Verfahren rechenintensiv. Im Unterschied zu den zweifachen XOR-Verfahren arbeiten diese Kodierungsverfahren im Allgemeinen innerhalb eines Stripe, wodurch während Schreiboperationen weniger Daten zu bearbeiten sind. Fraglich bleibt, ob daraus – wegen des höheren Rechenaufwands – ein Performancevorteil erwächst.

RAID 6

Aus den Datenpaketen berechnet man mittels Reed-Solomon-Code zwei Prüfsummen P und Q, Syndrome genannt. Reed-Solomon-Codes sind fehlerkorrigierende Kodierungsverfahren aus den 60er Jahren. In einer allgemeineren Sichtweise als bei CRC-Verfahren, betrachtet man statt Bits Zeichen aus mehreren Bits und konstruiert einen endlichen algebraischen Zahlkörper (Galois-Körper), auf dem eine Addition und Multiplikation definiert sind. Diese Operationen benutzt man zur Berechnung der Prüfsummen. Reed-Solomon-Verfahren sind Blockkodierungsverfahren, die jede vorgegebene Anzahl von Fehlern in einem Block korrigieren können. Sie finden unter anderem auch bei der Datenspeicherung auf CDs, DVDs und bei DVB-Technologien Verwendung.

Für den Spezialfall zweier Syndrome P und Q operiert man auf dem Zahlkörper $GF(2^8)$, was die Zahl der Platten auf 256 beschränkt. Die Berechnung von P reduziert sich hierbei auf einfaches XOR. Lediglich die Berechnung von Q ist aufwändiger **(4)**. Als Software-Lösung ist RAID 6 (**Abbildung 7**)seit Kernel 2.6.2 Bestandteil des md-Treibers von Linux. Viele Hersteller bieten hardwarebeschleunigte RAID-6-Implementierungen an.

RAID^n

RAID^n bezeichnet eine Familie alternativer RAID-Level der Tandberg-Tochter Inostor. Der patentierte Algorithmus verwendet laut Firmenangaben keine Reed-Solomon-Codes und ist hauptsächlich in Hardware von Tandberg und Inostor integriert. Es gibt auch eine Softwarelösung in Form von Linux-Kernel-Modulen .

RAID X

RAID-X ist ein scheinbar noch in Arbeit befindlicher Level von ECC Technologies, der RAID 3 erweitert.

Fazit

Wachsende Speichervolumen machen alternative RAID-Lösungen jenseits der klassischen RAID-Level nötig. Externe Hardwarelösungen, PCI-RAID-Controller oder Software mit zeitgemäßen RAID-Implementierungen sind marktreif und verfügbar. Bereits die Vielzahl der Ansätze und der uneinheitliche Sprachgebrauch zeigen jedoch, dass eine Standardisierung noch nicht in Sicht ist. Deshalb dürfte die weitere Entwicklung in diesem Bereich in den nächsten Jahren spannend zu verfolgen sein. *(jcb)* ■■■

Infos

(1) P.M. Chen, E.K. Lee, G.A. Gibson, R.H. Katz und D.A. Patterson: „RAID: High-Performance, Reliable Secondary Storage", in ACM Computing Surveys, Vol 26, No. 2, June 1994, S. 145-185.

(2) M. Blaum, J. Brady, J. Bruck und J. Menon: „EVENODD: An efficient scheme for tolerating double disk failures in RAID architectures", in Proceedings of the Annual International Symposium on Computer Architecture, S. 245-254, 1994.

(3) Peter Corbett, Bob English, Atul Goel, Tomislav Grcanac, Steven Kleiman, James Leong und Sunitha Sankar: „Row-Diagonal Parity for Double Disk Failure Correction", in Proceedings of the Third USENIX Conference on File and Storage Technologies März/April 2004.

(4) H. Peter Anvin: „The mathematics of RAID-6": (http://www.kernel.org/pub/linux/kernel/people/hpa/)

Der Autor

Marcus Schuster arbeitet als System-Engineer bei der transtec AG, Tübingen. Dort gehört es zu seinen Aufgaben, die für eine Vermarktung vorgesehenen, externen Hardware-RAID-Systeme zu prüfen, sie auf Fehler zu untersuchen und diese zu bereinigen.