ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 1996

Seria: INFORMATYKA z. 30 Nr kol.1315

Jarosław FLAK

# MACIERZE DYSKOWE - WZROST NIEZAWODNOŚCI SERWERÓW PLIKÓW

Streszczenie. Redundancyjne macierze dyskowe są obecnie najlepszym rozwiązaniem zabezpieczającym urządzenia dyskowe przed skutkami awarii, a także zwiększającym szybkość realizacji operacji dyskowych. W artykule przedstawiono cechy i fundamentalne koncepcje rozwiązań macierzy dyskowych, analizę niezawodności macierzy, problemy związane z występowaniem awarii oraz projektowaniem macierzy.

## DISK ARRAYS - IMPROVEMENT OF RELIABILITY OF FILE SERVERS

Summary. Redundant disk arrays are nowadays the best solution to overcome disk failures and to improve disk performance. This article describes features and fundamental ideas of disk arrays, contains the analisys of their reliability, failure occurrence problems, and topics covering failures and the design of disk arrays.

## MATRICES DES DISQUES - AMELIORATIONS ET CERTITUDES DES SERVEURS DES FICHIERS

Resume. Les matrices de disques sont actuellement parmis les meilleurs resolutions protegeant les disques contre les differentes pannes, permettant aussi d'augmenter la vitesse de réalisation des operations des disques. Dans cet article seront presentés les differentes charecteristiques, les conceptions de resolutions des matrices de disques ainsi q'une analise de la certitude de ces matrices.
Il y seront présentées aussi les differentes resolutions dans le cas d'une panne ainsi que les méthodes prises pour étudier un projet des matrices de disques.

# 1. Wprowadzenie

Jedną z ważniejszych usług, którą musi zapewniać sieć komputerowa, jest udostępnianie danych, których żąda jej użytkownik. W sieciach z serwerem do świadczenia usługi dostępu do danych zgromadzonych na dyskach sieciowych służy wyspecjalizowana maszyna - serwer plików. Konieczne jest minimalizowanie czasu obsługi żądań odwołujących się do dysków, a więc podnoszenie efektywności serwera plików. Jego wydajność ma duże znaczenie dla szybkości obsługi użytkownika, jednak dedykowane serwery plików dają sobie z reguły radę ze sprawną obsługą żądań. Ważniejszym może problemem jest ochrona danych zgromadzonych na dyskach serwera przed ich utratą w razie awarii. Jeśli danych takich jest dużo i są często modyfikowane, tworzenie kopii zapasowych może zajmować dużo czasu i być nieopłacalne. Wygodnym rozwiązaniem jest wtedy zainstalowanie systemu, który pozwoli na przetrwanie awarii dysku, zapewni ciągłość pracy i odzysk danych ze zniszczonego dysku. Przy okazji poprawi jeszcze w czasie normalnej pracy szybkość operacji dyskowych. Cechy takie posiada macierz dyskowa.

W latach osiemdziesiątych prędkość przesyłu danych dyskowych w systemach superkomputerowych okazała się już niewystarczająca. Powstał wówczas pomysł połączenia kilku niezależnych dysków twardych systemu komputerowego w jedną macierz dyskową, która umożliwi zrównoleglenie dostępu do dysków i zwiększy w ten sposób wydajność systemu. Macierz taką nazwano redundancyjną macierzą tanich dysków (ang. Redundant Array of Inexpensive Disks) - RAID.

Obecnie zainteresowanie macierzami RAID wyraźnie wzrasta. Dzieje się tak z kilku powodów. Jednym z nich jest gwałtowny rozwój układów elektroniki, co wpływa na zwiększanie mocy obliczeniowej procesorów i pojemności pamięci. Szybkość dostępu do danych na dysku jest wyraźnie mniejsza, a wpływa ona w bardzo dużym stopniu na wydajność całego systemu. Osiągi procesorów RISC zwiększane są aż o ok. 50 % rocznie, a czas dostępu do dysków (zależny od mechaniki dysku) zwiększa się tylko o ok. 10 %, a szybkość przesyłu danych (zależna od mechaniki dysku i upakowania danych) o ok. 20 % rocznie (dane z roku 1994). Przewiduje się, że przepaść pomiędzy stale zwiększaną wydajnością procesorów a wydajnością dysków magnetycznych będzie się jeszcze powiększała. Szuka się więc rozwiązań nie w mechanicznym polepszaniu parametrów, bo już wiadomo, że rozwój tej dziedziny nie nadąża za rozwojem procesorów, ale w logicznym łączeniu dysków i zrównoleglaniu operacji dyskowych.

Drugą przyczyną wzrostu zainteresowania macierzami RAID jest rozwój i tworzenie nowych aplikacji (video, hypertext, multimedia, CAD), operujących na dużych strukturach danych i korzystających intensywnie z pamięci masowych. Również w sieciach komputerowych, do których jest przyłączonych ok. 60 % wszystkich komputerów, rośnie zapotrzebowanie na szybkie, duże dyski - ich pojemność zwiększana jest o ok. 40 % rocznie.

Macierze dyskowe są obecnie produktami handlowymi. Ich nabywcami są banki, giełdy, systemy rejestracji klientów. Są to instytucje wymagające od systemów nieprzerwanej pracy i niezawodności składowania danych. Jest to w tej chwili najbardziej ceniona zaleta macierzy, powodująca też wzrost ich popularności. Jednak mimo tego, że jest to produkt handlowy, w systemach macierzy dyskowych pozostało jeszcze wiele problemów do rozwiązania. Wciąż trwają prace nad udoskonaleniem tych systemów, pełnym wykorzystaniem ich możliwości i optymalnym zastosowaniem w praktyce.

## 2. Cechy macierzy dyskowych

W dziedzinie rozwoju dysków magnetycznych nie przewiduje się raczej radykalnej poprawy ich parametrów mechanicznych i szybkiego wzrostu ich prędkości. Rozwiązaniem logicznym problemu zwiększenia wydajności pamięci masowych może być macierz dyskowa. Łączy ona wiele niezależnych dysków fizycznych w jeden wielki dysk o dużej wydajności i rozdziela dane na kilka dysków fizycznych, pozwalając na równoległy, szybszy do nich dostęp.

Jednak wzrost liczby dysków powoduje wzrost niezawodności systemu. Duże macierze są bardziej podatne na awarie niż pojedyncze dyski. Na przykład w macierzy 20-dyskowej wystąpienie awarii jest 20- krotnie bardziej prawdopodobne. O ile średni czas do wystąpienia awarii dla jednego dysku wynosi 200 000 godzin, czyli ok. 23 lata, to dla dwudziestu dysków wynosi on 10 000 godzin, tj. ok. 1 roku. Dla zwiększenia niezawodności systemu wprowadzono więc redundancję danych w postaci kodów korekcyjnych. Umożliwia to odtworzenie danych z uszkodzonego dysku i pozwala na uniknięcie utraty danych nawet przez dłuższy czas niż dla pojedynczego dysku. Jednak powoduje to konieczność aktualizacji redundancyjnych informacji po każdym zapisie, co zmniejsza z kolei szybkość systemu. Ogólnie mówiąc, macierze o małej redundancji są szybsze, o dużej redundancji - bezpieczniejsze.

Podsumowując, można powiedzieć, że redundancyjna macierz dyskowa posiada następujące cechy:

— jest zestawem wielu dysków fizycznych, widzianych przez system operacyjny jako jeden duży dysk logiczny;
— dane są rozrzucone na wszystkie dyski fizyczne zestawu umożliwiając równoległy odczyt lub zapis;
— zapisywana jest redundancyjna informacja służąca do odtworzenia danych w razie awarii dysku.

Zaletami macierzy dyskowych są:

— szybkość dostępu do danych na dysku;
— bezpieczeństwo danych;
— ciągłość pracy, czyli możliwość wymiany uszkodzonego dysku i rekonstrukcji danych bez przerywania pracy systemu.

Obecnie można zaobserwować zapotrzebowanie na pamięci masowe gwarantujące przede wszystkim bezpieczeństwo danych. Udoskonala się więc systemy dysków macierzowych w tym kierunku, starając się jednak nie stracić przy tym na ich szybkości. Zostało opracowanych szereg projektów wykorzystujących rozdzielanie i redundancję danych. Ich kombinacje tworzą szereg opcji dla użytkowników i projektantów, zapewniając subtelne korzyści między szybkością, niezawodnością i kosztami. Są one trudne do oszacowania bez znajomości alternatywnych rozwiązań. Poniżej zostaną opisane podstawowe organizacje macierzy dyskowych, ich zalety i wady, a także zostaną porównane ich szybkości, koszty i niezawodność.

## 2.1. Rozwiązania stosowane w macierzach dyskowych

Redundancyjne macierze dyskowe wykorzystują dwie podstawowe koncepcje w celu zapewnienia wysokiej wydajności i bezpieczeństwa:

— rozdzielania danych na kilka dysków fizycznych zestawu, aby polepszyć szybkość;
— tworzenia i zapisywania redundancyjnej informacji, aby polepszyć niezawodność.

Między wydajnością a niezawodnością macierzy istnieje pewna zależność. Jeśli zwiększymy redundancję, macierz będzie bardziej niezawodna, ale mniej wydajna. Wpływa na to również sposób rozmieszczenia na dyskach danych i informacji nadmiarowej.

### 2.1.1. Paskowanie danych

Rozdzielanie informacji na kilka dysków polega na podziale jej na bloki (paski), które są zapisywane jednocześnie na oddzielnych dyskach. Mechanizm ten nazywany jest paskowaniem (ang. striping).

Operacje dostępu do dysku dokonujące odczytu lub zapisu danych będziemy nazywać operacjami wejścia/wyjścia, w skrócie: I/O (ang. Input/Output).

Żądania odczytu danych mieszczących się w jednym pasku angażują tylko ten dysk, na którym ów pasek się znajduje. Umożliwia to równoległy odczyt lub zapis danych, co zwiększa szybkość obsługi operacji I/O i przesyłu danych. Zapewnia też równomierne obciążenie wszystkich dysków, co zapobiega sytuacji, w której niewielka liczba dysków jest obciążona, podczas gdy inne są bezczynne.

Równoległość obsługi może być realizowana na dwa sposoby:

— Wiele niezależnych operacji I/O dla małych porcji danych może być obsłużonych równolegle przez oddzielne dyski; zwiększa to szybkość obsługi operacji I/O (skraca czas oczekiwania zgłoszenia w kolejce).

— Pojedyncze żądanie dostępu do danych rozproszonych na wielu dyskach może być obsłużone jednocześnie przez te dyski (wskazane dla dużych porcji danych obsługiwanych przez jedną procedurę I/O); zwiększa to szybkość przesyłu danych.

Dla równomiernej pracy dysków i różnych rodzajów żądań I/O duże znaczenie ma długość paska. Większość macierzy dyskowych może być klasyfikowana na podstawie dwóch kryteriów:

— granulacji przeplotu danych (długości bloku przy paskowaniu);

— metod obliczania redundancyjnej informacji i sposobu jej rozmieszczenia na dyskach macierzy.

Ze względu na rozmiar bloku przeplot danych może być określony jako drobnoziarnisty i gruboziarnisty.

Macierze o granulacji drobnoziarnistej przeplatają dane w relatywnie małych jednostkach. Każde żądanie I/O, niezależnie od rozmiaru, dotyczy wszystkich dysków macierzy. Umożliwia to bardzo dużą prędkość przesyłu danych dla wszystkich żądań I/O, ale posiada też wady:

— tylko jedno zgłoszenie może być obsługiwane w danym momencie;

— wszystkie dyski zużywają czas na pozycjonowanie głowicy dla każdego żądania.

Macierze o granulacji gruboziarnistej stosują przeplot danych w relatywnie dużych jednostkach. Żądanie I/O małej ilości danych dotyczy niewielu dysków, a żądanie dużej ilości danych może dotyczyć wszystkich dysków. Pozwala to na równoległą obsługę wielu małych zgłoszeń i dużą prędkość przesyłu danych dla dużych zgłoszeń.

### 2.1.2. Redundancja informacji

Wszystkie dyski fizyczne macierzy tworzą razem jeden duży dysk logiczny. Im więcej dysków w zestawie, tym lepsza wydajność wynikająca z paskowania. Niestety, jak wspomniano wyżej, większa liczba dysków obniża niezawodność macierzy. Aby przetrwać awarię dysku i kontynuować pracę nie tracąc danych, konieczna jest redundancja informacji.

Redundancja w macierzach dyskowych polega na wytworzeniu nadmiarowej informacji o pewnych blokach danych (np. sumy modulo 2) i następnie zapisaniu jej tak, aby podczas awarii jednego dysku możliwe było odtworzenie zapisanych tam danych na podstawie informacji redundancyjnej i danych z pozostałych dysków.

Dla przykładu załóżmy, że na czterech dyskach będą przechowywane dane, a na piątym - bity parzystości (suma modulo 2) dla odpowiednich bitów danych z czterech dysków. W razie awarii któregoś dysku można odtworzyć dane, które były na nim przechowywane, obliczając sumę modulo 2 dla odpowiednich bitów z wszystkich pozostałych dysków (również z dysku z informacją redundancyjną). Suma ta będzie poszukiwaną wartością, która była w tym miejscu na uszkodzonym dysku.

Stosowanie redundancji wymaga rozpatrzenia dwóch problemów:

- Wyboru metody obliczania redundancyjnej informacji. Większość dzisiejszych matryc stosuje parzystość (sumę modulo 2), chociaż niektóre korzystają z kodów Reed-Solomona lub Hamminga.
- Sposobu rozmieszczenia redundancyjnej informacji na dyskach macierzy. Istnieje ścisły związek między szybkością macierzy a rozmieszczeniem redundancyjnej informacji. Chociaż jest wiele sposobów jej rozmieszczenia, dzielimy je na dwie grupy:
  - koncentracja na małej ilości dysków;
  - równomierne rozmieszczenie na wszystkich dyskach; jest to bardziej pożądane, gdyż unika się nierównomiernego obciążenia dysków.

Chociaż podstawy paskowania i redundancji danych są koncepcyjnie proste, jest wiele możliwości zastosowania tych rozwiązań, w zależności od celów zastosowania macierzy.

## 3. Organizacje macierzy dyskowych

Istnieje szereg organizacji (poziomów) macierzy RAID, z których każdy dotyczy różnej szybkości przetwarzania i stopnia zabezpieczenia danych. Zdefiniowano siedem

(0 - 6) poziomów macierzy RAID. Zostaną one omówione niżej i będą stanowiły podstawę późniejszej oceny wydajności, kosztów i niezawodności różnych konfiguracji macierzy. W celu uniknięcia ewentualnych niezgodności numeracji poziomów w nawiasach zostaną także podane ich angielskie określenia. Poszczególne poziomy zostały przedstawione na rys. 1. Ciemniejsze pola dysków oznaczają miejsce przechowywania informacji redundancyjnej.

**RAID 0 (Nonredundant)**

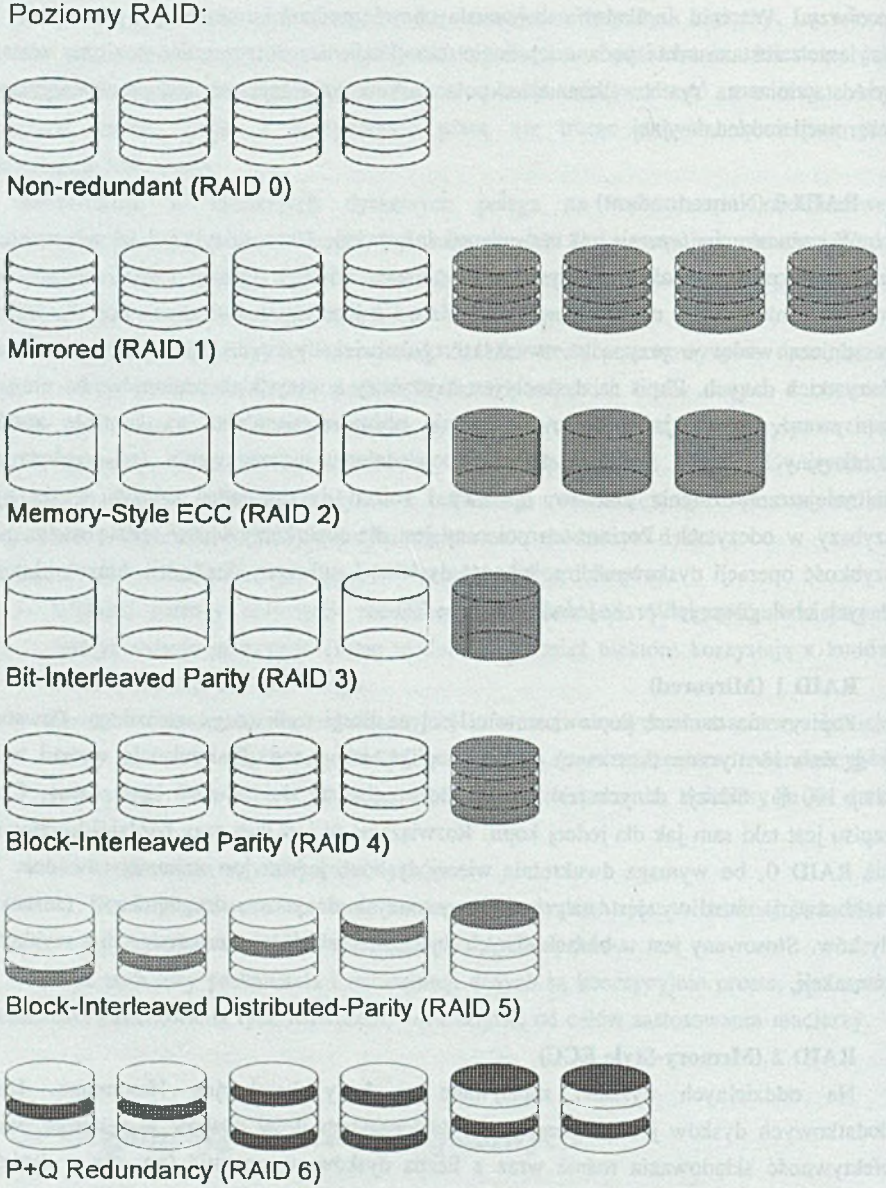
W rozwiązaniu tym nie ma redundancji informacji. Dane są tylko dzielone na bloki (paski) i rozrzucane na różne dyski zestawu (ang. striping). Odczyt i zapis odbywa się równocześnie z wielu dysków. Jest to najtańsze i najprostsze rozwiązanie. Posiada jednak zasadniczą wadę: w przypadku awarii któregokolwiek dysku następuje całkowita utrata wszystkich danych. Zapis na dyskach jest najszybszy z wszystkich poziomów, bo nie jest zapisywana informacja redundancyjna i nie poświęca się czasu na kontrolę kodów korekcyjnych. Tylko RAID 1 dzięki odpowiedniemu uszeregowaniu (ze względu na najmniejsze opóźnienia przesuwu głowicy i rotacji dysku) żądań odczytu może być szybszy w odczytach. Poziom ten polecany jest dla superkomputerów (gdzie ważna jest szybkość operacji dyskowych i pojemność dysków) i wskazany dla żądań dużych bloków danych obsługiwanych przez jedną procedurę I/O.

**RAID 1 (Mirrored)**

Zapisywane dane są kopiowane w całości na drugi dysk (ang. mirroring). Powstają więc dwa identyczne (lustrzane) dyski - duplikat informacji. Redundancja wynosi więc tutaj 100 %. Odczyt danych jest szybki, może odbywać się z dwóch kopii naraz. Czas zapisu jest taki sam jak dla jednej kopii. Rozwiązanie to jest dwa razy bardziej kosztowne niż RAID 0, bo wymaga dwukrotnie więcej dysków, jednak jest najmniej zawodne. W razie awarii możliwy jest natychmiastowy odzysk danych z drugiej kopii (zestawu) dysków. Stosowany jest w bazach danych wymagających dużej niezawodności i szybkości transakcji.

**RAID 2 (Memory-Style ECC)**

Na oddzielnych dyskach zapisywane są kody korekcyjne Hamminga. Ilość dodatkowych dysków jest proporcjonalna do logarytmu ilości dysków w systemie, więc efektywność składowania rośnie wraz z liczbą dysków w systemie (np. dla 4 dysków potrzeba 3 dodatkowe, dla 11 potrzeba 4). Dane są zapisywane w ten sposób, że na kolejnych dyskach zapisuje się kolejny bit lub bajt informacji (ang. bit/byte-interleaving). Odczyt i zapis następują równolegle ze wszystkich dysków (synchronicznie).




Rys. 1. Poziomy 0 - 6 macierzy RAID
Fig. 1. RAID levels 0 through 6

W RAID 2 w danej chwili może być obsługiwana tylko jedna operacja I/O - jest to poważne ograniczenie efektywności, występujące także w poziomie 3. W razie awarii do identyfikacji uszkodzonego dysku trzeba odczytać wszystkie dyski z kodami, ale do odtworzenia informacji wystarczy odczytać tylko jeden odpowiedni dysk z kodami. Jest to rozwiązanie polecane dla dużych komputerów typu host. Dla rozwiązań sieciowych uznaje się, że jest za drogie.

**RAID 3 (Bit-Interleaved Parity)**

Redundancyjna informacja (parzystość) zapisywana jest na jednym dysku. Dane są dzielone na bity lub bajty i zapisywane na kolejnych dyskach. Odczyt i zapis następują równolegle ze wszystkich dysków. Ponieważ kontroler dysku może łatwo zidentyfikować, który dysk uległ awarii, w RAID 3 jest tylko jeden dysk z informacją parzystości. Wszystkie dyski działają synchronicznie, ale możliwa jest tylko jedna operacja I/O w danej chwili. Rozwiązanie to nie nadaje się więc do przetwarzania transakcyjnego. Jest proste do implementacji. W razie awarii następuje łatwy odzysk informacji z dysku z parzystością. Zapis i odczyt jest szybki tylko dla dużych bloków danych. Poziom ten nadaje się dla stacji obróbki plików graficznych ze względu na szybkość.

**RAID 4 (Block-Interleaved Parity)**

Redundancyjna informacja zapisywana jest na jednym dysku. Dane są dzielone na bloki (paski) i zapisywane na wielu dyskach. Odczyt małych porcji informacji możliwy jest równocześnie z wielu dysków. Zapis trzeba realizować jak dla pojedynczego dysku ze względu na to, że jest tylko jeden dysk korekcyjny, na którym trzeba przy zapisie uaktualniać parzystość. Odczyt jest szybki, można realizować równolegle wiele żądań odczytu małych porcji informacji.

Wadą poziomu 4 (a także 5) jest zmniejszona szybkość zapisu niepełnych pasków. Jeśli zapis bloków następuje na wszystkich dyskach (nazwijmy to dużym odczytem), to oblicza się i zapisuje parzystość tylko dla nowych danych (nowych pasków na wszystkich dyskach). Jeśli zapis bloków dokonywany jest na małej liczbie dysków (mały odczyt), to parzystość obliczana jest jako różnica parzystości starych i nowych danych i zapisywana jest nowa parzystość. Konieczne jest dokonanie aż czterech operacji I/O: odczytu dawnych danych, zapisu nowych danych, odczytu starej parzystości i zapisu nowej parzystości na to samo miejsce (procedura read-modify-write). Rozwiązanie RAID 4 jest praktycznie nie stosowane - zastępowane jest o wiele lepszym RAID 5.

**RAID 5 (Block-Interleaved Distributed-Parity)**

Redundancyjna informacja (parzystość) zapisywana jest na wszystkich dyskach. Dane są dzielone na paski i zapisywane na wielu dyskach. Odczyt i zapis następują równolegle z wielu dysków. Możliwa jest równoczesna obsługa wielu żądań I/O małych porcji informacji.

Poziom RAID 5 jest najczęściej stosowany. Posiada chyba najmniej wad i wiele zalet - jest optymalny pod względem wydajności, bezpieczeństwa, ceny i możliwości konfiguracji. Nie ma już tutaj przeciążonego dysku z parzystością - "wąskiego gardła" dla operacji zapisów. Wszystkie dyski są równomiernie wykorzystane. Dla dużych zapisów, dużych odczytów i małych odczytów uzyskuje się tu najlepsze wyniki ze wszystkich poziomów. Dla małych zapisów konieczne jest stosowanie procedury read-modify-write, która jest główną wadą tego rozwiązania, ale ciągle trwają badania nad poprawieniem szybkości "małych" zapisów. RAID 5 jest rozwiązaniem 20 - 50 % tańszym niż RAID 1, przy porównywalnym poziomie bezpieczeństwa. Polecany jest dla przetwarzania transakcyjnego i sieci lokalnych Novell i Unix.

**RAID 6 (P+Q Redundancy)**

Rozwiązanie podobne do RAID 5, ale redundancyjne informacje dla każdego z dysków są kopiowane jeszcze na różne dyski. Stopień niezawodności jest tu większy - dopuszczalna jest awaria dwóch dysków jednocześnie. Jednak prędkość zapisu małych porcji informacji została pogorszona (6 operacji we/wy przy małych zapisach).

Wraz z rozwojem technologii systemów RAID pojawiają się wciąż coraz to nowsze i doskonalsze pomysły, klasyfikowane przez producentów jako nowe poziomy (np. RAID 10, jako próba kombinacji RAID 1 i RAID 0). Jednak w większości opierają się one na fundamentalnych rozwiązaniach omówionych powyżej.

## 4. Porównanie wydajności i kosztów

Porównanie poziomów RAID jest sprawą dosyć trudną. Trzeba wziąć pod uwagę trzy parametry: wydajność, koszty i niezawodność. Każdy z poziomów określa raczej konfigurację i sposób użycia niż implementację macierzy. Sprawia to dodatkowe kłopoty, bo np. RAID 5 w pewnych implementacjach zachowuje się podobnie jak inne poziomy. Występuje tu zależność doboru konfiguracji macierzy od celów, jakim ma służyć. Wybór

jednostki porównawczej także zależny jest od podejścia. Przyjęto uniwersalną jednostkę porównawczą:

$$k = \frac{liczba\ \ operacji\ \ I/O}{koszt\ \ macierzy} \qquad (1)$$

Tabela 1 zawiera porównanie poszczególnych poziomów RAID w odniesieniu do RAID 0, ze względu na przyjętą jednostkę porównawczą k. Zakłada się jednakowy koszt systemów, proporcjonalny do ilości dysków w macierzy. Grupa parzystości jest to zestaw dysków, dla których jest obliczana i zapisywana redundancyjna informacja. G oznacza liczbę dysków w grupie parzystości, mały zapis/odczyt - operacja na jednej jednostce paska (jednym dysku), duży zapis/odczyt - operacja na całym pasku (wszystkich dyskach w grupie parzystości).

Tabela 1

Porównanie poziomów macierzy ze względu na przyjęty wskaźnik k

| poziom RAID | odczyt mały | zapis mały | odczyt duży | zapis duży | efektywność składowania |
|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 / 2 | 1 | 1 / 2 | 1 / 2 |
| 3 | 1 / G | 1 / G | (G-1) / G | (G-1) / G | (G-1) / G |
| 5 | 1 | max (1/G, 1/4) | 1 | (G-1) / G | (G-1) / G |
| 6 | 1 | max (1/G, 1/6) | 1 | (G-2) / G | (G-2) / G |

Na podstawie wcześniejszego opisu i analizy tabeli można stwierdzić, że poziomy 1 - 4 są podklasami poziomu 5. Przykładowo system poziomu 1 jest równoważny systemowi poziomu 5 przy wielkości grupy parzystości równej 2, a poziom 3 jest podklasą poziomu 5 z ograniczoną wielkością paska danych. Wartość wskaźnika k dla poziomu 3 jest zawsze mniejsza lub równa wartości dla poziomu 5.

Klasyfikację macierzy poziomów 1 - 5 można więc sprowadzić do zagadnienia konfiguracji macierzy poziomu 5. Podobnie niezawodność wszystkich poziomów 1 - 4 rozpatrywać jako niezawodność różnych konfiguracji RAID 5.

## 5. Niezawodność

Niezawodność jest cechą macierzy dyskowych, która chyba w największym stopniu przyczyniła się do ich popularności. Dzięki przechowywaniu redundancyjnej informacji w macierzach złożonych z wielu dysków uzyskuje się większy stopień bezpieczeństwa przechowywania danych niż dla pojedynczego dysku. Awaria jednego dysku nie powoduje awarii całej macierzy, a jedynie włącza program odtwarzania danych znajdujących się na dysku, który uległ awarii. Odczytując parzystość i dane z wszystkich pozostałych dysków w grupie parzystości można odtworzyć utracone dane.

Zakładając, że awarie dysków są niezależne od siebie, możemy oszacować średni czas, po którym nastąpi awaria dwóch dysków jednocześnie i nie będzie już możliwości odzyskania danych. Czas ten (średni czas do utraty danych) oznaczymy przez MTTDL (ang. Mean Time To Data Loss). Patterson [3] podaje następujący wzór do wyliczenia tego czasu dla poziomu RAID 5:

$$MTTDL = \frac{MTTF^2(disk)}{N \times (G - 1) \times MTTR\ (disk)}, \tag{2}$$

gdzie MTTF (ang. Mean Time To Failure) oznacza średni czas do awarii pojedynczego dysku, N - liczbę wszystkich dysków w macierzy, G - rozmiar grupy parzystości, MTTR (ang. Mean Time To Repair) - średni czas usunięcia awarii pojedynczego dysku.

Jeśli założymy MTTF = 200000 godzin, N = 100, G = 16, MTTR = 1 godzina, to uzyskamy czas ok. 3000 lat.

Dla macierzy dyskowej z dwoma redundancyjnymi dyskami w grupie parzystości (poziom RAID 6) MTTDL wynosi:

$$MTTDL = \frac{MTTF^3(disk)}{N \times (G - 1) \times (G - 2) \times MTTR^2(disk)} \tag{3}$$

Konfiguracja ta daje jeszcze większą niezawodność. Dla poprzednich założeń otrzymujemy astronomiczną wartość 43 miliony lat. Dla mniejszej liczby dysków okresy te są jeszcze dłuższe. Wyliczenia te jednak, jakkolwiek prawdziwe, dają tylko wyidealizowany obraz niezawodności macierzy. Przy realistycznym spojrzeniu trzeba wziąć pod uwagę jeszcze inne czynniki, które mają wpływ na obniżenie niezawodności systemu. Są to awarie systemu, niekorygowalne błędy bitów lub skorelowane awarie dysków, drastycznie obniżające niezawodność.

**Awarie systemu i błędy parzystości**

Awarią systemu będziemy nazywać takie zdarzenia, jak: zanik napięcia zasilania, awaria sprzętu, błąd operatora systemu, błąd oprogramowania powodujący przerwanie operacji I/O macierzy. Mogą one występować częściej niż uszkodzenia samych dysków. Takie awarie przerywają operacje zapisu powodując stany, gdy parzystość nie jest jeszcze zaktualizowana, a dane są już zapisane lub odwrotnie. W obu przypadkach parzystość jest niepoprawna i nieprzydatna w przypadku awarii dysku. Nie istnieją rozwiązania pozwalające wykluczyć możliwość takiej awarii, chociaż można obniżyć prawdopodobieństwo jej wystąpienia, np. przez duplikowanie sprzętu i zapewnienie niezależnych źródeł zasilania. Aby uniknąć utraty parzystości przy awariach systemu informacje o parzystości powinny zostać zarejestrowane w pamięci nieulotnej przed rozpoczęciem każdej operacji zapisu. Po zakończonym zapisie dane te są już niepotrzebne. Sprzętowe implementacje RAID mogą efektywnie realizować taką rejestrację, implementacje programowe obniżają jednak wydajność macierzy.

**Niekorygowalne błędy bitów**

Są to niewykrywalne przekłamania występujące przy odczycie lub zapisie informacji z dysku. Powody występowania tego zjawiska nie są do końca jasne. Nie wiadomo, czy błąd powstał podczas zapisu danych czy tylko przy aktualnym odczycie. Obecnie dyski odznaczają się częstością wystepowania niekorygowalnych błędów bitów, wynoszącą ok. 1 na $10^{14}$ odczytów bitów. Jest to częstość, z jaką błędy wykrywane są przy odczycie podczas normalnej pracy dysku. W przypadku uszkodzenia dysku jego zawartość musi być odtworzona przez odczytanie danych z nieuszkodzonych dysków macierzy. Przykładowo, rekonstrukcja dysku w 100 GB macierzy wymaga odczytu ok. $2 \times 10^8$ milionów sektorów. Częstość występowania błędów, wynosząca 1 na $10^{14}$ bitów, powoduje, że jeden na $2.4 \times 10^{10}$ sektorów nie zostanie poprawnie odczytany. Jeśli założymy, że przekłamania bitów czytanych sektorów są niezależne od siebie, to prawdopodobieństwo poprawnego odczytu wszystkich sektorów 100 GB macierzy wynosi:

$$\frac{2 \times 10^8 \times (1 - \frac{1}{2.4 \times 10^{10}})}{2.4 \times 10^{10}} = 99.2\ \% \qquad (4)$$

Oznacza to, że średnio 0.8 % awarii dysków o takich parametrach będzie powodowało stratę danych na skutek niekorygowalnego błędu bitu. Jest to groźne zjawisko, znacznie obniżające niezawodność dla pewnych konfiguracji.

**Skorelowane awarie dysków**

Najprostszy model obliczania niezawodności zakładał, że uszkodzenia dysków są niezależne od siebie. W rzeczywistości występowanie czynników środowiskowych i produkcyjnych powoduje, że awarie dysków są skorelowane ze sobą. Mogą to być przypadki pożaru, zaniku lub wahania napięcia, awaria współdzielonego sprzętu dla kilku dysków, błędy seryjne. Awarie są także częstsze w pewnych okresach używania, np. całkiem nowego lub starego sprzętu. Skorelowane awarie dysków znacznie obniżają niezawodność macierzy, powodując niebezpieczeństwo następnej awarii, zanim zostanie odtworzona zawartość poprzednio uszkodzonego dysku.

## 5.1. Realistyczne spojrzenie na niezawodność

Biorąc pod uwagę omówione czynniki, możemy oszacować bardziej praktycznie, w jakim stopniu macierz dyskowa zabezpiecza przed utratą danych. Analiza tych zagadnień pokazuje, że całkowita utrata informacji może nastąpić na skutek jednego z trzech zdarzeń:

— podwójnej awarii dysków;
— awarii systemu po awarii dysku;
— awarii dysku po niekorygowalnym błędzie bitu podczas odtwarzania.

Jak wspomniano już wyżej, w rozwiązaniach sprzętowych przed awarią systemu (szczególnie po awarii dysku) możemy się częściowo zabezpieczyć stosując dodatkowe źródła zasilania i nieulotną pamięć, aby nie stracić parzystości. Aby skonstruować prosty model skorelowanych awarii założymy, że każda następna awaria dysku jest 10 razy bardziej prawdopodobna niż poprzednia, dopóki uszkodzony dysk nie zostanie zrekonstruowany. Do obliczeń niezawodności macierzy RAID 5 i RAID 6 przyjęto dalej następujące parametry:

— liczba dysków = 100
— pojemność dysku = 5 GB
— wielkość sektora = 512 bajtów
— niekor. błąd bitu = 1 na $10^{14}$
— grupa parzystości = 16 dysków
— MTTF (dysk) = 200000 godzin
— MTTF (dysk2) = 20000 godzin
— MTTF (dysk3) = 2000 godzin
— MTTR (dysk) = 1 godzina
— MTTF (sys) = 1 miesiąc.

Prawdopodobieństwo p (dysk) poprawnego odczytu wszystkich sektorów na dysku, obliczone na podstawie rozmiaru dysku, sektora i częstości występowania błędu bitu, wynosi 99.96 %. Na podstawie [3] obliczono MTTDL dla każdego z przypadków prowadzących do utraty danych. Wyniki przedstawia tabela 2. Dla poziomu 5 wszystkie trzy zdarzenia mają zbliżony wpływ na niezawodność. Najbardziej niebezpieczne jest wystąpienie awarii dysku po niekorygowalnym błędzie bitu. Prawdopodobieństwo P (10 lat) utraty danych w ciągu okresu 10 lat dla tego zdarzenia wynosi aż 26 %. Przy zastosowaniu implementacji sprzętowej uzyskuje się lepszy sumaryczny wynik, można bowiem nie brać pod uwagę sytuacji, w której po uszkodzeniu dysku nastąpi awaria systemu. Jednak najlepszy rezultat wynosi tylko 28 lat. P (10 lat) dla implementacji sprzętowej i programowej jest zbliżone i wynosi średnio 34 %.

Tabela 2

Obliczenia niezawodności dla RAID 5

| zdarzenie | MTTDL | MTTDL |
|---|---|---|
| podwójna awaria dysków | $\frac{MTTF\ (dysk) \times MTTF\ (dysk2)}{N \times (G - 1) \times MTTR\ (dysk)}$ | 304 lata |
| awaria systemu i awaria dysku | $\frac{MTTF\ (sys) \times MTTF\ (dysk)}{N \times MTTR\ (dysk)}$ | 164 lata |
| błąd bitu i awaria dysku | $\frac{MTTF\ (dysk)}{N \times (1 - p\ (dysk)^{G-1})}$ | 38 lat |
| sprzętowa | suma harmoniczna powyższych wielkości | 28 lat |
| programowa | suma harmoniczna bez przypadku awarii systemu i awarii dysku | 34 lata |

W tabeli 3 zawarto obliczenia dla poziomu 6. Tutaj zdecydowanie obniża niezawodność awaria systemu po awarii dysku. Zaleca się więc zastosowanie dla tego poziomu dodatkowego źródła zasilania. Dla implementacji sprzętowej jest to doskonała konfiguracja, w której tracimy dane średnio po 24000 lat. Jest to bardzo niezawodny system, odporny na skorelowane awarie i niekorygowalny błąd bitu, nawet przy

uszkodzeniu dwóch dysków jednocześnie. Rozwiązanie to wymaga jednak jednego dysku więcej w grupie parzystości niż w RAID 5.

Tabela 3

Obliczenia niezawodności dla RAID 6

| zdarzenie | MTTDL | MTTDL |
|---|---|---|
| potrójna awaria dysków | $\frac{MTTF(dysk) \times MTTF(dysk2) \times MTTF(dysk3)}{N \times (G-1) \times (G-2) \times MTTR^2(dysk)}$ | 43488 lat |
| awaria systemu i awaria dysku | $\frac{MTTF(sys) \times MTTF(dysk)}{N \times MTTR(dysk)}$ | 164 lata |
| błąd bitu i podwójna awaria dysków | $\frac{MTTF(dysk) \times MTTF(dysk2)}{N \times (G-1) \times (1-p(dysk)^{G-2}) \times MTTR(dysk)}$ | 54501 lat |
| sprzętowa | suma harmoniczna powyższych wielkości | 162 lata |
| programowa | suma harmoniczna bez przypadku awarii systemu i awarii dysku | 24188 lat |

Przeprowadzone obliczenia niezawodności dają pewne pojęcie o aktualnych parametrach macierzy dyskowych. Jednakże są to tylko szacunkowe obliczenia, przybliżające rzeczywiste warunki. Mogą one być wskaźnikiem przy porównywaniu innych organizacji macierzy o różnych konfiguracjach. Przedstawiono tu różne typy awarii, które macierz może przetrwać i takie, które powodują utratę danych lub ograniczają niezawodność systemu. Jednak są jeszcze inne czynniki, trudniejsze do zaklasyfikowania i wyliczenia, np. niezawodność części oprogramowania czy różnych konfiguracji sprzętu. Projektując i implementując macierze dyskowe stosuje się różne, bardziej szczegółowe rozwiązania, które wpływają na polepszenie parametrów niezawodnościowych macierzy, między innymi: rozgrupowaną parzystość (rozmieszczoną jak najoptymalniej na dyskach), dyski zapasowe (umożliwiające automatyczne przełączenie się na nowy dysk po awarii), ortogonalny układ dysków w macierzy (umożliwiający pracę w razie uszkodzenia kontrolera dysków i całego łańcucha dysków). Gromadzi się i przechowuje różnorodne znaczniki i informacje o stanie systemu, implementuje się różne - mniej lub bardziej efektywne - algorytmy wychodzenia ze stanu

awarii ([2] i [4]), zapobiega się przesyłaniu nie odświeżonych danych (bez uaktualnionej parzystości). Zagadnienia te mają ścisły związek z poprawianiem także szybkości działania macierzy. Aby zwiększyć wydajność, stosuje się buforowanie i caching danych, zapis parzystości w najszybciej dostępnym miejscu na dysku (pływająca parzystość) lub opóźnianie zapisu parzystości (rejestracja parzystości). Oczywiście powoduje to z kolei spadek niezawodności. Trzeba więc pogodzić ze sobą lub zdecydować się na jeden z dwóch celów: wydajnie czy niezawodnie. Zależy to już od zastosowań i wymagań stawianych konkretnej macierzy dyskowej.

### 5.2. Implementacja systemów macierzy dyskowych

Macierz dyskową można zaimplementować (zainstalować) na kilka sposobów. Istnieją odpowiednie drivery i urządzenia (implementacje dla różnych platform sieciowych (SCO Unix, Novell NetWare, Microsoft Windows NT, IBM OS/2, Macintosh)).

Generalnie sposoby implementacji można podzielić na trzy grupy:

- SW: programowe (ang. software). Główne funkcje sterujące i kalkulacyjne zawarte są w firmowym oprogramowaniu. Typ ten wykorzystuje standardowe sterowniki SCSI.
- HW: sprzętowe (ang. hardware). Funkcje sterujące wykonywane są przez specjalizowane, wielokanałowe (kanały SCSI) sterowniki macierzowe DAC (ang. Disk Array Controlers), przystosowane do współpracy z magistralą EISA.
- FW: sprzętowo-programowe (ang. full-ware). Zasadniczo są implementacją sprzętową, lecz posiadają własny "system operacyjny". Są niezależne od platformy sprzętowej i systemowej komputera.

W zasadzie, z nielicznymi wyjątkami, nie istnieją samodzielne programy obsługi macierzy dyskowych. Producenci dostarczają z reguły kompletne rozwiązania w postaci zestawu dyskowego i programu obsługi, przystosowanego do konkretnego, sieciowego systemu operacyjnego.

## 6. Podsumowanie

Macierze dyskowe mogą być skonfigurowane w rozmaity sposób. W celu poprawienia niezawodności i dostępności danych proponuje się dodatkowo różne ulepszenia: powielanie napędów (zapasowe dyski), kontrolerów, dodawanie buforów dyskowych, dodatkowe źródła zasilania itd. Dosyć trudno jest porównać poziomy RAID ze względu

na różne parametry: wydajność, niezawodność lub koszt. Chociaż najefektywniejszym rozwiązaniem wydaje się RAID 5 lub 6, stały spadek cen dysków twardych może spowodować większą popularność podsystemów RAID 0, 1 lub 10. Widać jednak, że obecnie liczą się tylko trzy rozwiązania: RAID 0, 1 i 5, z przewagą tego ostatniego, który jest coraz bardziej udoskonalany. Jego rozszerzenie - RAID 6, jak wykazano, doskonale nadaje się do zastosowań, w których wymagana jest wysoka niezawodność systemu komputerowego, chociaż dobór optymalnej konfiguracji, ze względu na różnorodność rozwiązań, jest sprawą złożoną.

## LITERATURA

[1] Gibson G. A., Patterson D. A.: Designing Disk Arrays for High Data Reliability. Journal of Parallel and Distributed Computing, January 1993, Vol. 17, pp. 4-27.

[2] Hellerstein L., Gibson G. A., Karp R. M., Katz R. H., Patterson D. A.: Coding Techniques for Handling Failures in Large Disk Arrays. Algorithmica, June 1994, Vol. 12, No. 3-4, pp. 182-208.

[3] Chen P. M., Lee E. K., Gibson G. A., Katz R. H., Patterson D. A.: RAID: High-Performance, Reliable Secondary Storage. ACM Computing Surveys, June 1994, Vol. 26, No. 2, pp. 145-185.

[4] The conference paper: Backward Error Recovery in Redundant Disk Arrays. Proc. of the 1994 Computer Measurement Group (CMG) Conference, Vol. 1, pp. 63-74.

Recenzent: Dr hab inż. Tadeusz Czachórski

Wpłynęło do Redakcji 18 grudnia 1995 r.

## Abstract

The article presents the features that make the redundant disk arrays more and more popular, describes the existing concepts (striping and redundancy) and their influence upon the parameters of the array. The standard levels of RAID systems (0 through 6) are described, their performances and reliabilities as well as costs (Table 1) are compared. It

is stated that RAID levels 1 through 4 are subclasses of RAID level 5, which proves to be an universal solution.

Reliability calculations have been carried out, assuming that disk and system failures occur independently. The real factors that may cause the failure were taken into account (Table 2 and 3). The failures that may cause the loss of data were analysed. It has been proven that some specific kinds of failure, like correlated disk failures and uncorrectable bit errors, affect the reliability especially badly, while particular RAID levels also differ in their failure resistance. RAID level 6 with doubled parity distributed on disks is the most reliable one. The mean time to failure with data loss amounts to 24 thousand years for hardware implementation, while software implementation is much less failure resistant - its time is 162 years only. The article also mentions design, implementation and failure recovery issues.